Bölüm 6

Çok Modlu Yapay Zekâda Görsel Dil Modelleri: Mimari Temeller ve Sektörel Uygulamalar

İrem Cakcak[1]

Burhan Duman[2]

Özet

Yapay zekâ alanındaki son gelişmeler, metin ve görsel veriyi ortak bir temsil uzayında birleştirebilen Görsel-Dil Modellerinin (VLM) farklı disiplinlerde etkin biçimde kullanılmasını mümkün kılmıştır. Geleneksel tek-modlu görsel tanıma sistemlerinin ötesine geçen bu yenilikçi modeller; bağlamsal yorumlama, çok modlu muhakeme ve görev odaklı üretim süreçlerini destekleyen bütüncül yaklaşımlar sunmaktadır. Bu çalışma, Büyük Dil Modelleri (LLM) ve VLM tabanlı yaklaşımların temel mimari bileşenlerini detaylı bir şekilde ele alırken, uygulama boyutunda ağırlıklı olarak görsel dil modellerinin sektörel entegrasyonuna odaklanmayı amaçlamaktadır. Çalışmada öncelikle, her iki model ailesinin temelini oluşturan Transformer mimarisi; öz-dikkat (self-attention) mekanizmaları, tokenizasyon ve konumsal kodlama gibi süreçler bağlamında teknik olarak incelenerek LLM'lerin çalışma prensipleri açıklanmıştır. Devamında ise odak noktası tamamen çok modlu yapılara kaydırılarak, VLM ve eylem boyutuyla genişletilmiş Vision-Language-Action (VLA) mimarilerinin spesifik kullanım alanları detaylandırılmıştır. Bu kapsamda; otonom sürüş sistemlerinde uçtan uca planlama, robotik sistemlerde mekânsal temellendirme, sağlık alanında yapılandırılmış bilgi çıkarımı, tarımda uzaktan algılama ve alan-özgü tanı sistemleri ile insan odaklı görsel analiz görevlerindeki güncel yaklaşımlar incelenmiştir. Yapılan incelemeler, VLM tabanlı sistemlerin pasif bilgi işleyiciler olmaktan çıkarak, yüksek seviyeli semantik rehberlik sağlayan karar destek bileşenlerine evrildiğini göstermektedir. Bununla birlikte, modellerin gerçek dünya sistemlerine aktarımında karşılaşılan hesaplama maliyetleri, gerçek zamanlılık kısıtları ve halüsinasyon riskleri gibi mühendislik darboğazları tartışılmış; gelecekteki eğilimin parametre-verimli ve alan-özgü hibrit mimarilere yöneleceği sonucuna varılmıştır.

1 Yüksek Lisans Öğrencisi, Isparta Uygulamalı Bilimler Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı , iremcakcak1012@gmail.com, 0009-0004-8929-3950

2 Dr. Öğr. Üyesi, Isparta Uygulamalı Bilimler Üniversitesi Teknoloji Fakültesi Bilgisayar Mühendisliği, burhanduman@isparta.edu.tr, 0000-0001-5614-1556

https://doi.org/10.58830/ozgur.pub1230.c4967

1. Giriş

Yapay zekâ sistemleri uzun yıllar boyunca belirli görevler için tasarlanan, çoğunlukla tek bir veri türüyle çalışan ve dar kapsamlı çıktılar üreten yapılar olarak geliştirilmiştir. Metin işleyen doğal dil sistemleri ile görüntüye odaklanan bilgisayarlı görü modelleri büyük ölçüde birbirinden bağımsız ilerlemiştir. Ancak son dönemde büyük ölçekli öğrenme modellerinin yaygınlaşması, hesaplama altyapısındaki ilerlemeler ve özellikle Transformer tabanlı mimarilerin olgunlaşması, bu ayrışmış yapıyı önemli ölçüde dönüştürmüştür. Bilgi işleme ve karar verme süreçleri artık daha bütüncül, daha bağlama duyarlı ve daha esnek sistemler üzerinden ele alınmaktadır.

Bu dönüşümün merkezinde yer alan Büyük Dil Modelleri (Large Language Models – LLM), yalnızca metni analiz eden sistemler olmaktan çıkarak; doğal dilin istatistiksel örüntülerini ve anlamsal yapısını öğrenebilen, çok adımlı muhakeme kurabilen, yönerge izleyebilen ve farklı görev türlerine uyarlanabilen genel amaçlı bileşenlere dönüşmüştür. LLM'ler, geniş ölçekli ön-eğitim süreçleri sayesinde yalnızca cevap üretmekle kalmamakta; bağlamı takip edebilmekte, tutarlı metinler oluşturabilmekte ve karmaşık problem yapılarını dil üzerinden temsil edebilmektedir.

Bununla birlikte, gerçek dünyadaki bilginin yalnızca metinden ibaret olmadığı açıktır. İnsan karar verme süreçleri çoğu zaman görsel, mekânsal ve dilsel bilgilerin birlikte değerlendirilmesine dayanır. Bu gereksinim, metin, görüntü ve video gibi farklı veri kaynaklarının ortak bir temsil uzayında birleştirilmesini zorunlu hâle getirmiştir. Bu noktada öne çıkan Görsel–Dil Modelleri (Vision–Language Models – VLM), görsel içerik ile dilsel ifadeler arasında semantik bir bağ kurarak sahne anlama, nesne–ilişki çıkarımı ve görsel bağlama dayalı akıl yürütme gibi görevlerde geleneksel tek-modlu yaklaşımların sınırlarını genişletmektedir.

Geleneksel bilgisayarlı görü sistemleri çoğunlukla "görselde ne var?" sorusuna sınıflandırma ya da nesne tespiti gibi çıktılar üretirken, VLM tabanlı yaklaşımlar bu içeriği dilsel olarak gerekçelendirebilmekte, görsel öğeler arasındaki ilişkileri açıklayabilmekte ve kullanıcı niyetine göre görev odaklı yanıtlar oluşturabilmektedir. Böylece kullanıcı etkileşimi de değişmektedir. Önceden tanımlı komut setleri ve etiketleme temelli arayüzler yerine; doğal dilde soru sorma, açıklama talep etme, görev tanımlama ve adım adım yönlendirme gibi daha esnek iletişim biçimleri ön plana çıkmaktadır. Bu yönüyle görsel–dil modelleri, yalnızca algısal doğruluğu artıran sistemler değil; insan–makine etkileşimini yeniden yapılandıran çok modlu karar destek bileşenleri olarak değerlendirilebilir.

Ancak çok modlu sistemlerin pratikte güvenilir biçimde kullanılabilmesi, yalnızca akıcı metin üretimi ya da yüksek doğruluklu görsel temsil öğrenimi ile sınırlı değildir. Gerçek dünya uygulamalarında model çıktısının görsel kanıta dayanması, bağlama uygun terminoloji kullanması ve tutarlı bir gerekçelendirme sunması kritik önem taşır. Özellikle otonom sürüş, sağlık, endüstriyel denetim ve tarım gibi karar maliyetinin yüksek olduğu alanlarda; halüsinasyon (görüntüde bulunmayan içerik üretimi), yanlış temellendirme, alan dışı genelleme zayıflığı, yüksek hesaplama maliyeti ve gerçek zamanlılık kısıtları temel mühendislik zorlukları olarak ortaya çıkmaktadır. Bu nedenle değerlendirme yalnızca model doğruluğu üzerinden değil; gecikme bütçesi, donanım sınırlamaları, veri dağılımı kayması ve sistem entegrasyon gereksinimleri üzerinden yapılmalıdır.

Son yıllarda literatürde gözlenen eğilim, görsel–dil modellerinin eylem boyutuyla bütünleşmesine yöneliktir. Görsel algı ve dilsel muhakemenin robotik ya da otonom sistemler gibi fiziksel dünyayla etkileşen yapılara aktarılması; planlama, görev ayrıştırma ve çok adımlı karar süreçlerinin yönetilmesini gerektirmektedir. Vision-Language-Action (VLA) mimarileri bu doğrultuda geliştirilmiş olup, dil tabanlı komutları görsel bağlamla ilişkilendirerek doğrudan eylem temsillerine dönüştürebilen bütüncül bir sistem yaklaşımı sunmaktadır. Bu evrim, sistemlerin denetimlenebilirlik, güvenlik ve açıklanabilirlik gereksinimlerini daha kritik bir noktaya taşımaktadır; çünkü üretilen çıktılar artık yalnızca metinsel öneriler değil, fiziksel sonuçlar doğurabilecek kararlar hâline gelmektedir.

Bu çalışma, söz konusu teknolojik ve mimari dönüşümü kavramsal temelleri ve uygulama alanları çerçevesinde sistematik biçimde ele almayı amaçlamaktadır. İlk eksende, LLM'lerin dayandığı Transformer mimarisi; öz-dikkat mekanizmaları, tokenizasyon ve konumsal kodlama süreçleri üzerinden incelenerek dil modellerinin teknik altyapısı açıklanmaktadır. İkinci eksende ise odak çok modlu yapılara kaydırılarak; VLM'lerin mimari türleri, eylem boyutuyla genişletilmiş VLA yaklaşımları, temel görev sınıfları ve otonom sürüş, sağlık, tarım ve robotik gibi alanlardaki uygulama örnekleri detaylandırılmaktadır. Bölümün temel katkısı, LLM ve VLM ekosistemini bütüncül bir çerçevede değerlendirmek; gerçek dünya entegrasyonunda karşılaşılan sınırlılıkları ortaya koymak ve gelecekteki eğilimin parametre-verimli, alan-özgü hibrit mimarilere doğru evrildiğini tartışmaktır.

2. Büyük Dil Modelleri (LLM)

Büyük dil modelleri geniş ölçekli metin veri kümeleri üzerinde eğitilen ve doğal dil işleme görevlerini yerine getirebilen derin öğrenme tabanlı modellerdir. Bu modeller; metin anlama, özetleme, çeviri, dil modelleme ve içerik üretimi

gibi görevlerde kullanılmaktadır. Büyük dil modellerin büyük çoğunluğu, dilsel bağlamı ve uzun menzilli bağımlılıkları etkili biçimde modelleyebilen Transformer mimarisine dayanmaktadır.

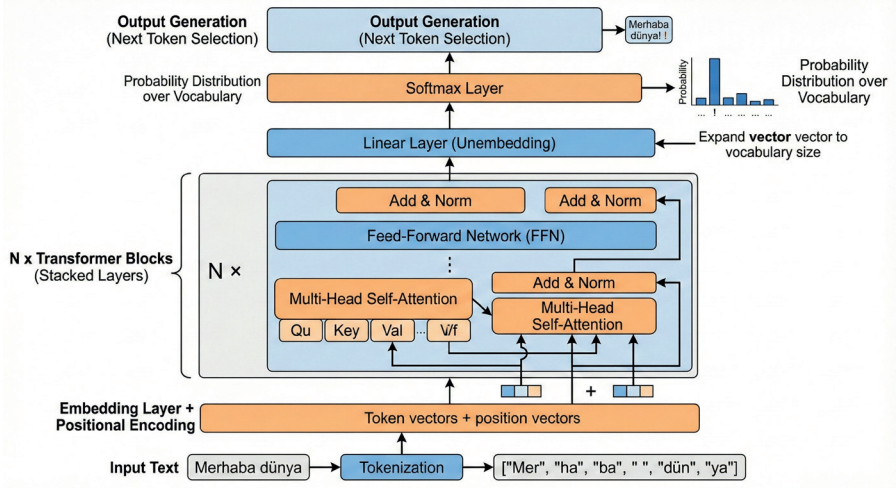
"Büyük" ifadesi yalnızca kavramsal bir nitelendirme değildir; modelin eğitildiği veri miktarının hacmine ve içerdiği parametre sayısının büyüklüğüne doğrudan işaret eder. Literatürde tanımlanan ölçeklenme yasaları (scaling laws), veri miktarı ve parametre sayısı arttıkça model performansının belirli bir doğrultuda ve öngörülebilir biçimde iyileştiğini göstermektedir.

2.1. Büyük Dil Modellerin Temel Mimarisi

Büyük dil modellerin temel mimarisi, Transformer tabanlı derin öğrenme yaklaşımlarına dayanmaktadır. Transformer mimarisi, önceki yinelemeli ağ yapılarının (RNN, LSTM) aksine, sıralı hesaplama zorunluluğunu ortadan kaldırarak dikkat (self-attention) mekanizması üzerinden paralel işlemeye olanak tanımaktadır. Bu özellik, modelin daha büyük veri kümeleri ve parametre sayılarıyla verimli biçimde eğitilmesini mümkün kılmıştır.

Günümüzde yaygın olarak kullanılan büyük dil modelleri, Transformer mimarisinin encoder, decoder veya encoder–decoder temelli farklı yapılandırmaları üzerine inşa edilmektedir.

2.1.1. Büyük Dil Modellerin Temel Yapısal Bileşenleri



Şekil 1: Transformer Tabanlı LLM'in Genel Çalışma Akışı

Yüksek soyutlama düzeyinde incelendiğinde, Transformer tabanlı bir büyük dil modeli (LLM) üç ana bileşenden oluşmaktadır (Şekil 1). İlk aşamada yer alan sözcükleyici (tokenizer), girdi metnini modelin işleyebileceği daha küçük alt birimlere (tokenlara) ayırır ve her birini sözlükte tanımlı benzersiz bir sayısal kimlikle temsil eder. Bu dijitalleştirme işlemi, metinsel verinin sayısal formata dönüştürülerek modelin dil üzerinde matematiksel işlemler yapabilmesine olanak tanır. Sayısallaştırılan bu girdiler, daha sonra modelin temel hesaplama merkezini oluşturan Transformer blokları yığınına aktarılır. Ardışık olarak düzenlenmiş onlarca katmandan oluşabilen bu bloklar, girdiyi derin bağlamsal temsillere dönüştürür. Modelin "katman derinliği" olarak da bilinen bu mimari büyüklüğü, karmaşık dilsel bağımlılıkları öğrenme kapasitesini doğrudan etkileyen en kritik tasarım parametrelerinden biridir. Sürecin son aşamasında ise dil modelleme kafası (language modeling head) devreye girer. Bu bileşen, Transformer bloklarından elde edilen bağlamsal temsilleri kullanarak modelin sözlüğündeki her bir token için bir olasılık dağılımı üretir ve metin üretim sürecinin bir sonraki adımını belirler.

2.1.2. Transformer Bloğunun İç Yapısı

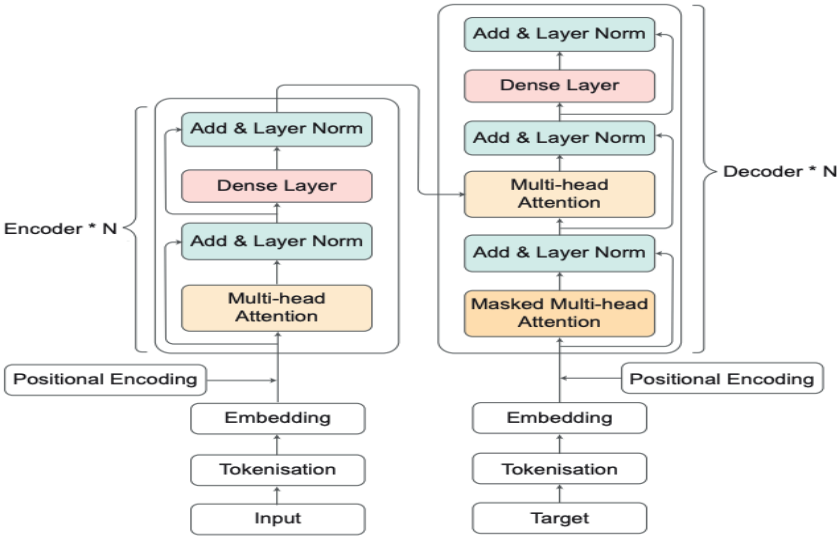
Transformer tabanlı bir LLM'de, her bir transformer bloğunun iç yapısı ve temel bileşenleri Şekil 2'de gösterilmektedir. Bir transformer bloğu, self-attention (öz-dikkat) mekanizması ve Feed-Forward Network (ileri beslemeli sinir ağı) bileşenlerinden oluşur.

2.1.2.1. Öz-Dikkat Mekanizması (Self-Attention) Katmanı

Self-attention mekanizması, bir tokenın temsili oluşturulurken dizideki diğer tokenlarla olan ilişkilerinin dikkate alınmasını sağlar. Bu sayede model, yerel bağlamın ötesinde uzun menzilli bağımlılıkları modelleyebilir ve bağlama duyarlı temsiller üretebilir. Öz-dikkat yapısı, özellikle sıralı verilerde bağlamsal ilişkilerin öğrenilmesinde kritik bir rol oynamaktadır.

2.1.2.2. İleri Beslemeli Sinir Ağı (Feed-Forward Network)

Self-attention katmanından elde edilen çıktılar, her token için bağımsız olarak uygulanan iki katmanlı bir Feed-Forward Network' e aktarılır. Bu bileşen, doğrusal olmayan dönüşümler aracılığıyla temsillerin zenginleştirilmesini sağlar. Transformer bloğu ayrıca artık (residual) bağlantılar ve katman normlama (layer normalization) bileşenlerini içerir. Artık bağlantılar derin ağlarda gradyan akışını desteklerken, katman normlama eğitim sürecinin kararlılığını artırmaya yardımcı olur.



Şekil 2: Transformer kodlayıcı–çözücü (encoder–decoder) mimarisi.

Kaynak: Vaswani vd. (2017).

2.1.3. Transformer Mimarisi Türleri

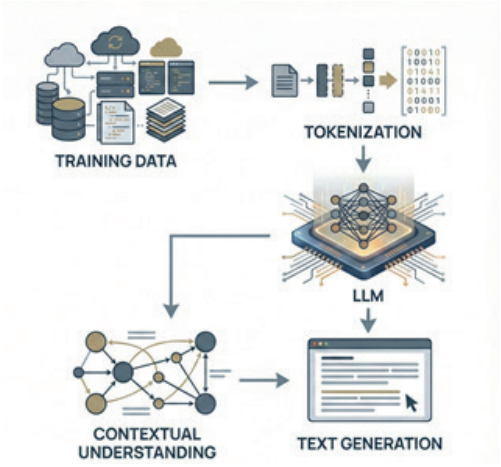
Transformer mimarileri, kullanım amaçlarına ve üstlendikleri doğal dil işleme görevlerinin niteliğine bağlı olarak üç temel varyasyonda tasarlanmaktadır. Kodlayıcı odaklı (encoder-only) modeller, metin üretiminden ziyade girdiyi bağlamsal olarak anlamaya odaklanır. BERT gibi mimariler bu gruba dahildir ve özellikle metin sınıflandırma, duygu analizi, öznitelik çıkarımı ve cümle temsili üretimi gibi analitik görevlerde yaygın biçimde kullanılmaktadır.

Buna karşılık, GPT serisi ve Phi gibi modellerin temsil ettiği kod çözücü odaklı (decoder-only) yapılar üretim merkezli bir yaklaşım benimser. Bu modeller oto-regresif (autoregressive) bir prensiple çalışarak her adımda bir sonraki tokenı tahmin eder ve metni ardışık biçimde üretir. Bu özellikleri sayesinde metin üretimi, soru yanıtlama, metin tamamlama ve diyalog sistemleri gibi görevlerde yüksek performans göstermektedir.

Kodlayıcı-kod çözücü (encoder–decoder) mimarileri ise hem girdiyi temsil eden bir kodlayıcı hem de bu temsile dayalı çıktı üreten bir kod çözücü bileşenini bir arada barındırır. T5 ve BART gibi modeller bu yapıyı kullanmaktadır. Bu mimari, özellikle bir girdi dizisinin farklı bir çıktı dizisine dönüştürülmesini gerektiren makine çevirisi, metin özetleme ve çeşitli sequence-to-sequence görevlerinde tercih edilmektedir.

2.2. Büyük Dil Modellerinin İşleyişi

Şekil 3'te gösterildiği üzere büyük dil modelleri metinsel dokümanları işlemek için çok aşamalı bir hesaplama süreci kullanır. Bu süreç; metnin model tarafından işlenebilir bir forma dönüştürülmesini, bağlamsal temsillerin oluşturulmasını ve kullanıcı girdisine uygun bir çıktının üretilmesini kapsar.



Şekil 3 : LLM çalışma prensibi

2.2.1. Tokenizasyon (Sözcükleme)

Tokenizasyon, büyük dil modelleri için temel bir ön işleme adımıdır ve metinsel girdilerin sayısal temsillere dönüştürülmesini sağlar. Bu süreçte, girdi metni model tarafından işlenebilecek tokenlara ayrıştırılır. Tokenlar; kelime, sub-word (alt-kelime) veya karakter düzeyinde tanımlanabilir ve bu temsil biçimi, modelin dil verisi üzerinde işlem yapabilmesi için zorunlu bir ara katman oluşturur.

Tokenizasyon sürecinde alt-kelime tabanlı yaklaşımlar, dildeki çeşitliliği daha dengeli biçimde ele alarak nadir kelimelerin ve biçimbilimsel varyasyonların modele dahil edilmesine olanak tanır. Bu nedenle, büyük ve heterojen metin kümeleriyle eğitilen modern büyük dil modelleride alt-kelime temsilleri yaygın olarak tercih edilmektedir.

Modern tokenizasyon yöntemleri, bilinmeyen kelime (Out-of-Vocabulary, OOV) problemini önemli ölçüde azaltarak modelin farklı dil yapıları ve yazım biçimleriyle karşılaşması durumunda daha kararlı sonuçlar üretmesine katkı sağlar. Özellikle boşluklardan bağımsız çalışan yöntemler, sosyal medya verileri

veya gürültülü metinler gibi düzensiz veri kaynaklarında daha esnek bir yapı sunmaktadır.

Yaygın olarak kullanılan tokenizasyon yöntemleri ve temel özellikleri Tablo 1'de özetlenmektedir.

Tablo 1:Tokenizasyon Yöntemleri

Yöntem	Temel Mekanizma	Seçim Kriteri	Kullanım Alanı / Özellikler
Byte-Pair Encoding (BPE)	Metindeki en sık görülen ardışık sembol çiftlerini birleştirerek sözlüğü genişletir.	Frekansı en yüksek çiftlerin seçimi.	Dengeli sözlük boyutu sağlar; alt-kelime temsilleriyle bilinmeyen kelime sorununu azaltır.
WordPiece	Alt birimleri, eğitim verisinin olasılığını en çok artıracak biçimde birleştirir.	Maksimum likelihood artışı sağlayan çiftlerin seçimi.	Daha kontrollü sözlük büyütme sunar; bağlam duyarlılığı daha yüksektir.
Unigram / SentencePiece	Geniş bir başlangıç token kümesi oluşturur; düşük olasılıklı tokenları iteratif şekilde eler.	Minimum kayıp veren token kümesinin belirlenmesi.	Boşluk bağımsız çalışır; gürültülü veya çok dilli veri üzerinde daha stabil performans gösterir.

2.2.2. Gömme (Embedding)

Şekil 4'te gösterildiği üzere, tokenizasyonun ardından gelen gömme (embedding) aşaması, ayrık token kimliklerini model tarafından işlenebilecek sürekli ve düşük boyutlu vektör temsillerine dönüştürür. Tokenlar yalnızca ayrık kimlikler (ID) olarak temsil edildiklerinde semantik bilgi taşımaz; gömme katmanı bu ayrık temsilleri ortak bir anlam uzayında konumlandırarak modele uygun bir giriş temsili oluşturur. Böylece kelimeler arasındaki anlamsal benzerlikler ve bağlamsal ilişkiler vektör uzayında ifade edilebilir hâle gelir ve Transformer tabanlı modellerin öz-dikkat mekanizması aracılığıyla bağlama duyarlı çıktılar üretmesinin temeli atılmış olur.

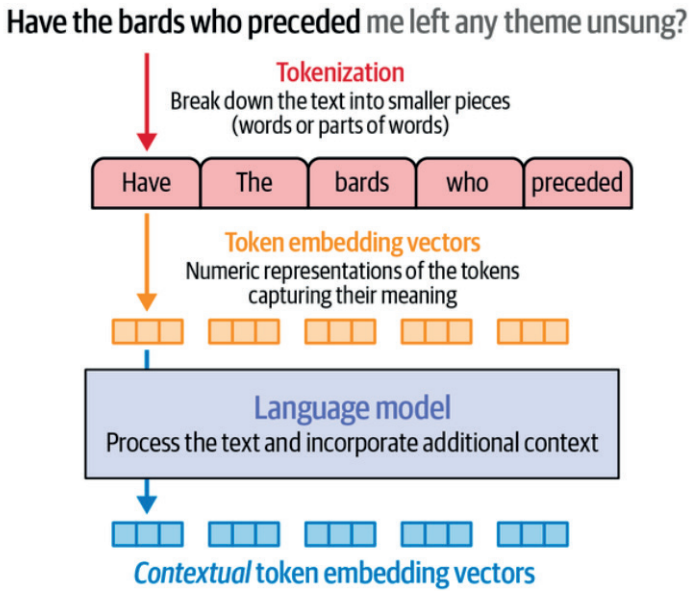
Matematiksel olarak bir tokenın gömme vektörü

$$e_t = W_e[t]$$

şeklinde tanımlanır. Burada $W_e \in \mathbb{R}^{|V| \times d}$ öğrenilen embedding matrisini, d modelin gizli boyutunu (model dimension), $e_t \in \mathbb{R}^d$ ise ilgili tokenın sürekli vektör temsilini ifade eder.

Embedding yaklaşımları tarihsel gelişim sürecinde statik ve dinamik olmak üzere ikiye ayrılmaktadır. Word2Vec ve GloVe gibi statik yöntemlerde her kelime bağlamdan bağımsız tek bir sabit vektörle temsil edilir; bu durum özellikle çok anlamlı kelimelerde sınırlılıklar doğurur. Transformer tabanlı modern modellerde ise token temsilleri bağlama duyarlı biçimde üretilir. BERT ve GPT gibi yapılarda bir tokenın temsili, bulunduğu bağlam ve çevresindeki diğer tokenlar dikkate alınarak dinamik biçimde güncellenir. Bu bağlamsal temsiller büyük ölçekli ön-eğitim (pre-training) süreçlerinde öğrenilmekte ve görev odaklı ince ayar (fine-tuning) aşamalarında optimize edilmektedir.

Ancak embedding vektörleri tek başına dizisel sıra bilgisini içermez. Bu nedenle Transformer mimarisinde gömme temsilleri konumsal kodlama (positional encoding) ile birlikte kullanılarak nihai giriş vektörü oluşturulur.



Şekil 4. Jetonların sözcükleme (tokenizasyon) sürecinden geçirilerek gömme (embedding) vektörlerine dönüştürülmesi ve bağlama duyarlı temsillerin elde edilmesi. Kaynak: Jay Alammar (2024)

2.2.3. Konumsal Kodlama (Positional Encoding)

Transformer tabanlı büyük dil modelleri metin dizilerinin sıralı yapısının doğru biçimde modellenmesi, dilin sözdizimsel ve anlamsal özelliklerinin yakalanabilmesi açısından önemli bir gerekliliktir. Ancak Transformer mimarisi tokenları paralel olarak işlediğinden, bu sıralı yapı bilgisini doğrudan temsil

etmez. Yalnızca gömme (embedding) temsillerine dayalı bir modelleme yaklaşımı, kelimelerin diziliminden kaynaklanan anlam farklılıklarını ayırt etmekte yetersiz kalabilir. Bu sınırlılığı gidermek amacıyla geliştirilen konumsal kodlama (positional encoding) yöntemleri, her bir tokenın dizideki konum bilgisini matematiksel olarak modele entegre etmektedir.

Konumsal bilgi genellikle başlangıçtaki gömme vektörüne eklenen bir konum vektörü aracılığıyla modele dahil edilir:

$$z_i = e_i + p_i$$

Burada $\mathbf{e}_i \in \mathbb{R}^d$, dizinin i. pozisyonundaki jetona ait embedding vektörünü; $\mathbf{p}_i \in \mathbb{R}^d$, aynı pozisyon için tanımlanan mutlak konumsal gömme vektörünü; $\mathbf{z}_i \in \mathbb{R}^d$ ise konumsal bilgi ile zenginleştirilmiş nihai giriş temsilini ifade etmektedir.

Literatürde konumsal bilginin modele entegrasyonu iki temel yaklaşım altında incelenmektedir:

- *Mutlak Konumsal Gömme (Absolute Positional Embeddings – APE):* Her pozisyon için sabit (sinüzoidal) veya veriden öğrenilebilir özgün bir vektör tanımlar. Ancak bu yöntemler genellikle önceden belirlenmiş bir maksimum dizi uzunluğuyla sınırlı kalır ve tokenlar arasındaki göreceli mesafeyi doğrudan modelleyemez.
- *Göreceli Konumsal Gömme (Relative Positional Embeddings – RPE):* Konumsal bilgiyi doğrudan dikkat (attention) mekanizmasına entegre ederek token çiftleri arasındaki mesafeyi baz alır. Uzun bağlamlı (long-context) modern dil modellerinde esneklik ve ölçeklenebilirlik sağlayan bu yaklaşımın öne çıkan iki türü bulunmaktadır:
    - *RoPE (Rotary Position Embeddings)*: Sorgu (query) ve anahtar (key) vektörlerine sinüzoidal rotasyon uygulayarak mesafe bilgisini hesaplamaya dâhil eder. Bu yapı, eğitimde görülmeyen uzun dizilerde bile modele güçlü bir genelleme yeteneği kazandırır.
    - *ALiBi (Attention with Linear Biases):* Konum bilgisini vektörlere eklemek yerine, dikkat skorlarına doğrusal bir sapma (ceza) olarak uygular. Tokenlar arası mesafe arttıkça dikkat ağırlığı sistematik olarak düşürülür, bu da özellikle uzun metinlerde oldukça kararlı bir çıkarım (ekstrapolasyon) davranışı sunar.

2.2.4. Dönüştürücü (Transformer) Blokları ve Öz-Dikkat (Self-Attention) Mekanizması

Transformer mimarisinin temel yapı taşı, birden fazla kez yinelenen dönüştürücü (transformer) bloklardır. Şekil 5'te gösterildiği üzere her bir blok; öz-dikkat (self-attention) mekanizması, iki katmanlı ileri beslemeli sinir ağı (feed-forward network), artık bağlantılar (residual connections) ve katman normlama (layer normalization) bileşenlerinden oluşur. Giriş temsilleri, embedding ve konumsal kodlama bileşenlerinin birleştirilmesiyle elde edilir ve blok içerisinde sırasıyla işlenir.

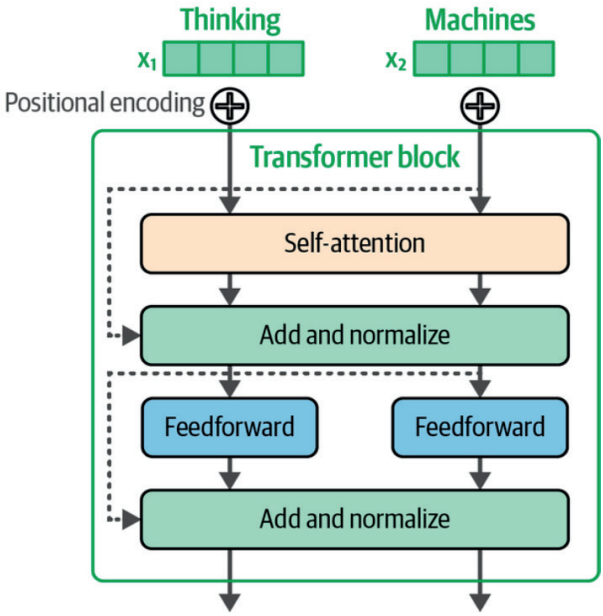


Figure 3-29. A Transformer block from the original Transformer paper.

Şekil 5: Transformer bloğunun genel yapısı

Blok yapısının ilk aşaması öz-dikkat mekanizmasıdır. Self-attention, her bir tokenın temsilini oluştururken dizideki diğer tüm tokenlarla olan ilişkisini dikkate alır. Bu sayede model yalnızca yerel komşuluk bilgisine değil, tüm dizi boyunca oluşan uzun menzilli bağımlılıklara erişebilir. Şekil 6'da şematik olarak gösterildiği üzere, bu mekanizma sorgu (Query, Q), anahtar (Key, K) ve değer (Value, V) temsilleri üzerinden çalışır.

Matematiksel olarak, giriş matrisi $X \in \mathbb{R}^{n \times d}$ olmak üzere, öz-dikkat için gerekli doğrusal projeksiyonlar şu şekilde tanımlanır:

$$Q = XW_Q, K = XW_K, V = XW_V$$

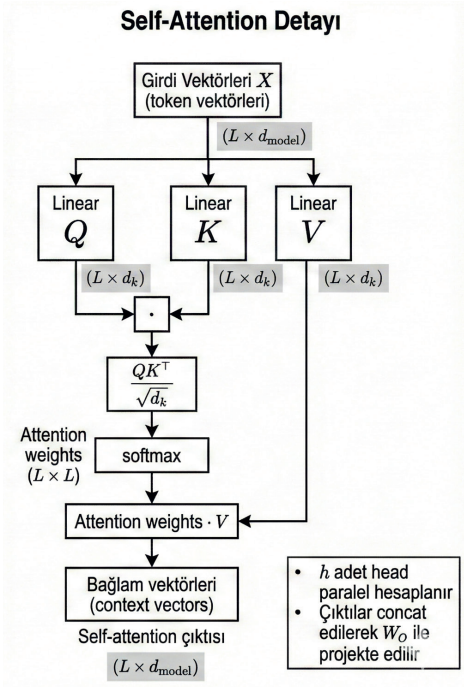
Burada W_Q, W_K, W_V öğrenilebilir ağırlık matrisleridir. Dikkat skorları, sorgu ve anahtar matrisleri arasındaki benzerlik üzerinden hesaplanır ve ölçeklenmiş noktasal çarpım (scaled dot-product) formunda ifade edilir:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Bu ifadede $QK^\top \in \mathbb{R}^{n \times n}$ matrisi, tüm token çiftleri arasındaki benzerlik skorlarını içerir. $\sqrt{d_k}$ ile ölçekleme işlemi, yüksek boyutlu temsillerde sayısal kararlılığı artırmak amacıyla uygulanır. Softmax işlemi her token için dikkat ağırlıklarını olasılık dağılımına dönüştürür ve önemli görülen tokenlara daha yüksek ağırlık verilmesini sağlar. Elde edilen dikkat ağırlıkları değer (V) matrisi ile çarpılarak her token için bağlama duyarlı, zenginleştirilmiş yeni bir temsil üretilir.

Öz-dikkat katmanının çıktısı doğrudan iletilmez; giriş temsili ile toplanarak artık bağlantı oluşturulur ve ardından katman normlama uygulanır. Bu "add & normalize" adımı, derin mimarilerde gradyan akışını stabilize ederek eğitimin daha kararlı ilerlemesine katkıda bulunur. Sonraki aşamada her tokena bağımsız olarak uygulanan iki katmanlı ileri beslemeli sinir ağı, doğrusal olmayan dönüşümler aracılığıyla temsilleri daha soyut özelliklere dönüştürür. Feed-forward çıktısına da tekrar artık bağlantı ve normlama uygulanarak blok tamamlanır.

Bu iki aşamalı yapı (self-attention + feed-forward), model derinliği boyunca yinelenerek daha üst düzey bağlamsal temsillerin öğrenilmesini sağlar. Şekil 5'te görüldüğü gibi, blok içerisindeki bu düzenli yapı Transformer mimarisinin hem paralel hesaplama verimliliğini hem de yüksek temsil kapasitesini mümkün kılan temel tasarım unsurudur.



Şekil 6: Self-attention mekanizması şematik gösterimi

2.2.5. İleri Beslemeli Ağ (Feed-Forward Network, FFN)

İleri beslemeli ağ ya da MLP (Multi-Layer Perceptron) alt bloğu, her transformer bloğunda çok başlı öz-dikkat katmanını izleyen temel bileşenlerden biridir ve modelin hesaplama kapasitesinin büyük kısmını üzerinde toplar. Öz-dikkat katmanı tokenlar arasındaki bağlamsal ilişkileri çözümlerken, FFN bu zenginleştirilmiş temsilleri her bir token için bağımsız olarak (position-wise) dönüştürür. Modelin ezberleme (memorization) ve genelleme (interpolasyon) yetenekleri ağırlıklı olarak bu katmanda şekillenir ve öğrenilen dilsel bilgilerin önemli bir bölümü burada depolanır.

Yapısal olarak FFN, iki doğrusal dönüşüm ve aralarındaki doğrusal olmayan bir aktivasyon fonksiyonundan oluşur. Öz-dikkat çıktısı olan $\in \mathbb{R}^{d_{model}}$, vektörü, ilk doğrusal katman (W_1) ile boyutu genellikle $4 \times d_{model}$ olacak şekilde genişletilir. Ardından (klasik yapıda ReLU, modern dil modellerinde ise çoğunlukla GELU veya SwiGLU gibi) bir aktivasyon fonksiyonundan geçirilen veri, ikinci doğrusal katman (W_2) ile tekrar d_{model} boyutuna sıkıştırılarak projeksiyon tamamlanır. Bu genişletme–sıkıştırma yapısı, modelin daha yüksek boyutlu bir temsil uzayında karmaşık, doğrusal olmayan örüntüler öğrenmesine

ve sonrasında bunları tekrar self-attention ile uyumlu boyuta projekte etmesine imkân tanır.

Eğitim sürecinde gradyan akışını ve yakınsamayı iyileştirmek için FFN bloğu; artık bağlantılar (residual connections) ve normalizasyon katmanlarıyla birlikte kullanılır. Modern tasarımlarda genellikle blok girişinde uygulanan ön-normalizasyon (LayerNorm veya RMSNorm) tercih edilmektedir. Bazı gelişmiş ve devasa ölçekli mimarilerde ise FFN, her token için yalnızca ilgili "uzman" ağların aktifleştiği Mixture-of-Experts (MoE) yapılarıyla genişletilmektedir. Bu bütünleşik yapı sayesinde modelin ifade gücü, yüksek kapasiteli ve seyrek (sparse) bir hesaplama modülüyle maksimize edilir.

2.2.6. Çıktı Katmanı ve Token Üretimi (Logits, Softmax, Sampling)

Büyük dil modellerinde metin üretimi, son Transformer bloğundan elde edilen gizli temsillerin sözlük uzayına projeksiyonu ile gerçekleştirilir. Her zaman adımında elde edilen $h_t \in \mathbb{R}^{d_{\text{model}}}$ boyutlu gizli vektör, doğrusal bir çıktı katmanı aracılığıyla model sözlüğündeki her bir token için bir uygunluk skoru (logit) üretir:

$$z_t = W_{\text{out}} h_t + b_{\text{out}}$$

Burada $W_{\text{out}} \in \mathbb{R}^{|V| \times d_{\text{model}}}$ ağırlık matrisi, $b_{\text{out}} \in \mathbb{R}^{|V|}$ sapma vektörü ve $|V|$ sözlük boyutudur. Elde edilen z_t vektörü ham skorları temsil eder ve henüz bir olasılık dağılımı değildir.

Bu skorlar Softmax fonksiyonu ile normalize edilerek olasılık dağılımına dönüştürülür:

$$p_t(i) = \frac{\exp\left(z_t(i)\right)}{\sum_{j=1}^{|\mathcal{V}|} \exp\left(z_t(j)\right)}$$

Bu işlem sonucunda her aday token için 0 ile 1 arasında bir olasılık değeri elde edilir ve toplamları 1 olacak şekilde normalize edilir. Eğitim aşamasında bu dağılım, gerçek hedef token ile karşılaştırılarak çapraz entropi (cross-entropy) kaybı hesaplanır ve model parametreleri geriye yayılım ile güncellenir.

Çıkarım (inference) sürecinde ise model, hesaplanan olasılık dağılımından bir sonraki tokenı seçerek metni adım adım üretir. Bu seçim süreci deterministik (greedy) veya olasılık temelli örnekleme stratejileri (top-k, top-p gibi) ile gerçekleştirilebilir. Ayrıca sıcaklık (temperature) parametresi kullanılarak dağılımın keskinliği ayarlanabilir; düşük sıcaklık değerleri daha tutarlı ve

öngörülebilir çıktılar üretirken, yüksek değerler daha çeşitli ve yaratıcı metinlere olanak tanır.

Metin üretimi oto-regresif bir mekanizma ile ilerler. Model her adımda bir sonraki tokenın koşullu olasılığını $P(x_t \mid x_{<t})$ biçiminde hesaplar ve seçilen token giriş dizisine eklenir. Bu süreç, belirlenen bir sonlandırma koşuluna ulaşılana kadar tekrarlanır. Böylece büyük dil modelleri, bağlama duyarlı ve tutarlı metinleri ardışık token üretimi yoluyla oluşturur.

3. Görsel–Dil Modelleri (Vision–Language Models, VLM)

Görsel dil modelleri, görüntü ve metin bilgisini ortak bir temsil çerçevesinde birleştirerek görsel içerik hakkında dil tabanlı çıkarım yapabilen çok modlu (multimodal) modellerdir. Bu modeller, yalnızca "görselde ne var?" sorusunu yanıtlayan geleneksel görsel tanıma yaklaşımlarının ötesine geçerek; nesneleri ve aralarındaki ilişkileri doğal dilde tanımlama, bağlamsal yorumlama ve soruya ya da komuta göre gerekçelendirilmiş yanıt üretme gibi daha üst düzey etkileşimleri mümkün kılar. Böylece kullanıcı, görsel içerikle etiketleme veya arama odaklı bir süreç yerine, soru sorma, açıklama isteme ya da yönlendirme verme gibi daha doğal bir iletişim biçimiyle etkileşime girebilir.

Görsel dil modellerinin pratik değeri, görsel ipuçlarını metinsel bağlamla birleştirerek tek bir görüntü üzerinden farklı türde görevleri destekleyebilmesinden kaynaklanır. Görüntü açıklama (image captioning), görsel soru–cevaplama (visual question answering, VQA) ve çapraz-modalite erişim (örneğin metinden görsel bulma ya da görselden metin bulma) gibi görevler, aynı model ailesi altında ele alınabilmekte; model, kullanıcının niyetine göre görseldeki ilgili kanıtları seçerek tutarlı bir dil çıktısına dönüştürebilmektedir. Bu bağlamda VLM yetenekleri literatürde belirli görev sınıfları altında incelenmektedir.

Görsel dil modellerinin başarısı büyük ölçüde web ölçekli görüntü–metin çiftleri üzerinde gerçekleştirilen ön-eğitim (pre-training) süreci sayesinde, her iki modaliteden zengin ve genellenebilir temsiller öğrenebilmesine dayanır. Bu sayede modeller, etiketli veriye sıkı biçimde bağımlı kalmadan yeni görüntüler üzerinde sıfır-atış (zero-shot) çıkarım yapabilir ve daha önce karşılaşmadıkları kavramları uygun metinsel ifadelerle ilişkilendirebilir. Bununla birlikte, görsel dil modellere yönelik ilgi hızla artmasına rağmen alan-özgül senaryolarda güvenilir kullanım için önemli güçlükler devam etmektedir. Üretilen çıktının görsel kanıta dayalı biçimde temellendirilmesi (grounding) her zaman yeterince açık olmayabilir; teknik veya klinik bağlamlarda gerekli olan terim doğruluğu ve ayrıntı düzeyi korunamayabilir; model eğitimde görmediği alanlara taşındığında genelleme performansı düşebilir.

Bu nedenle görsel dil modellerinin değerlendirilmesi, yalnızca genel amaçlı ölçütlerle sınırlı tutulmamalı; hedef uygulamanın gerektirdiği doğruluk, tutarlılık ve açıklanabilirlik kriterleri doğrultusunda alan verisi üzerinde test edilmesi ve gerektiğinde uyarlanması gerekmektedir. Bu çerçevede görsel dil modellerinin yetenekleri, literatürde yaygın biçimde kullanılan görev sınıfları üzerinden ele alınmaktadır. Görev temelli bu ayrım, görsel ve metinsel temsillerin etkileşim biçimini ve model çıktısının niteliğini görünür kılarak karşılaştırmalı değerlendirmeyi kolaylaştırır. Bu nedenle aşağıda, en sık kullanılan temel görev alanları kısaca özetlenmektedir.

3.1. Görsel Dil Modellerinin Temel Görev Alanları

Görsel dil modellerinin (VLM) görsel ve dilsel temsilleri ortak bir semantik uzayda birleştirme yeteneği, literatürde çeşitli temel görev alanları üzerinden değerlendirilmektedir. Bu görevler, modellerin çok modlu anlama, eşleştirme, muhakeme ve üretim kapasitesini ölçmeyi amaçlar.

Görsel Soru–Cevaplama (VQA), bir görüntü ve bu görüntüye ilişkin doğal dildeki sorunun birlikte değerlendirilerek uygun yanıtın üretilmesini hedefler. Bu görev; nesne tanıma, sayma ve uzamsal ilişkileri çözümleme gibi alt becerileri bir araya getirerek modelin çok modlu kavrama yeteneğini ortaya koyar.

Görsel Betimleme (Image Captioning), görüntüdeki sahnenin doğal dilde ifade edilmesidir. Modelin yalnızca görsel içeriği tanıması değil, aynı zamanda bunu akıcı ve bağlama uygun bir metne dönüştürmesi beklenir. Bu nedenle üretim (generation) odaklı VLM yeteneklerini yansıtır.

Görüntü–Metin Eşleştirme ve Getirme (Image–Text Retrieval), metin ve görüntü çiftlerinin ortak temsil uzayında hizalanmasına dayanır. Amaç, bir metne en uygun görüntüyü ya da bir görüntüye en uygun metni benzerlik skorları üzerinden belirlemektir. Bu yaklaşım, arama ve bilgi erişim sistemlerinde yaygın olarak kullanılmaktadır.

Görsel Muhakeme (Visual Reasoning), görüntüdeki öğeler arasındaki ilişkileri (neden–sonuç, karşılaştırma, olay akışı vb.) modelleyerek mantıksal çıkarım yapılmasını kapsar. Bu görev, basit tanımanın ötesine geçerek çok-adımlı akıl yürütme kapasitesini değerlendirir.

Görsel Referanslama (Visual Grounding), metindeki belirli bir ifadenin görüntüdeki ilgili nesne veya bölgeyle eşleştirilmesini hedefler. Bu yetenek, model çıktılarının görsel kanıta dayandırılmasını sağladığı için açıklanabilirlik açısından önemlidir.

Belge Zekâsı ve Düzen Anlama (Document Intelligence & OCR) ise yapılandırılmış dokümanlarda yalnızca metnin tanınmasını değil, aynı zamanda

sayfa düzeni ve görsel bileşenler arasındaki yapısal ilişkilerin yorumlanmasını içerir. Bu alan, kurumsal bilgi çıkarımı ve otomasyon uygulamalarında kritik bir rol oynamaktadır.

3.2. Görsel Dil Modellerinin Temel Yapısı

VLM'in mimari yapısı, görsel ve metinsel bilgiyi birlikte işleyebilmek amacıyla tasarlanmış üç temel bileşen etrafında şekillenir (Şekil 7). Bu bileşenler, farklı mimari düzenlemelere sahip modellerde değişen biçimlerde konumlandırılsa da, görsel dil modellerinin çalışma prensibinin temelini oluşturmaktadır.

3.2.1. Görüntü Kodlayıcı (Image Encoder)

Görüntü kodlayıcı, ham piksel değerlerinden oluşan görüntü veya video girdilerini, anlamsal açıdan zengin ve yoğun vektör temsillerine dönüştürmekle sorumludur. Bu bileşenin temel amacı, görsel içeriği dil bileşeninin işleyebileceği bir temsil uzayına taşımaktır. Bu süreçte yalnızca düşük seviyeli görsel özellikler (renk, kenar, doku) değil; aynı zamanda nesneler, sahne yapısı ve uzamsal ilişkiler de soyutlanmış biçimde kodlanır.

Literatürde erken dönem VLM çalışmalarında sıklıkla evrişimli sinir ağları (CNN) kullanılmıştır. Ancak güncel yaklaşımlarda, küresel bağlamı daha etkin biçimde modelleyebilmesi nedeniyle Vision Transformer (ViT) tabanlı kodlayıcılar yaygınlaşmıştır. ViT mimarilerinde görüntü, yamalara ayrılarak görsel belirteçler (visual tokens) hâline getirilir ve bu belirteçler dikkat mekanizmaları aracılığıyla tüm görüntü genelinde etkileşime girer. Bu yapı, özellikle dil bileşeniyle kurulacak etkileşim açısından esnek ve ayrıntılı bir temsil sunar.

Görüntü kodlayıcının çıktısı, mimariye bağlı olarak tek bir global görüntü vektörü (global embedding) veya görüntünün yamalarına karşılık gelen bir görsel token dizisi (patch-level tokens) biçiminde elde edilebilir. Token-temelli temsil, görüntünün farklı bölgelerine ait bilgiyi ayrı ayrı taşıdığı için nesne/bölge düzeyinde daha ayrıntılı temsil sağlarken; global temsil daha kompakt olup özellikle eşleştirme ve arama (retrieval) gibi görevlerde verimlidir.

Bu bileşende omurga olarak ResNet gibi CNN tabanlı yapılar veya ViT/DeiT/Swin Transformer gibi Transformer tabanlı kodlayıcılar kullanılmakta; ayrıca pek çok güncel VLM'de CLIP-ViT görsel kodlayıcısı yaygın bir tercih olarak öne çıkmaktadır.

3.2.2. Dil Bileşeni (Language Encoder / Decoder)

Dil bileşeni, metinsel girdilerin işlenmesinden ve gerekli durumlarda metin çıktısının üretilmesinden sorumludur. Bu bileşen, mimariye bağlı olarak yalnızca

bir dil kodlayıcıdan (encoder), yalnızca bir dil kod çözücüden (decoder) veya her ikisinin birlikte kullanıldığı bir yapıdan oluşabilir.

Dil kodlayıcılar, verilen metni anlamsal ve bağlamsal bir temsile dönüştürerek özellikle anlama ve eşleştirme temelli görevlerde kullanılır. Dil kod çözücüler ise görsel ve metinsel bağlamı birlikte değerlendirerek ardışık kelime üretimi gerçekleştirir ve görüntü açıklama ya da görsel soru–cevaplama gibi üretim odaklı görevlerin merkezinde yer alır. Güncel görsel dil modellerde bu bileşen çoğunlukla büyük dil modelleri müzerine inşa edilmekte; Transformer tabanlı mimariler, dilsel akıl yürütme ve tutarlılık açısından temel yapı taşı olarak kullanılmaktadır.

Encoder-only dil bileşenleri daha çok görüntü–metin eşleştirme/retrieval senaryolarında kullanılırken, decoder-only yapılar ve encoder–decoder modeller özellikle captioning ve VQA gibi üretim odaklı görevlerde öne çıkar. Güncel sistemlerde büyük dil modelleri çoğunlukla instruction-tuning ile çok modlu yönerge takibi (multimodal instruction following) yapacak şekilde uyarlanarak daha doğal ve görev yönelimli çıktılar üretir.

Bu kapsamda, görüntü–metin eşleştirme/retrieval odaklı yapılara CLIP, üretim ve yönerge takip eden multimodal çıktılara BLIP-2 / InstructBLIP ve LLM tabanlı görsel-diyalog yaklaşımlarına LLaVA örnek olarak verilebilir.

3.2.3. Modaliteler Arası Birleştirme (Cross-Modal Fusion)

Modaliteler arası birleştirme mekanizması, görüntü ve metin temsillerinin aynı bağlam içerisinde ilişkilendirilmesini sağlayarak görsel dil modellerinin ayırt edici yeteneklerini ortaya çıkaran bileşendir. Bu mekanizmanın temel işlevi, görsel ve dilsel bilgiler arasında hizalama kurmak ve iki modalite arasındaki karşılıklı bağımlılıkları öğrenmektir.

Bu süreç literatürde genellikle erken birleşim (early fusion), geç birleşim (late fusion) ve ara düzey / çapraz dikkat tabanlı birleşim (cross-attention fusion) olmak üzere üç ana yaklaşım altında incelenmektedir.

Geç birleşim (Late Fusion) yaklaşımında, görsel ve metinsel girdiler ayrı kodlayıcılarda işlenir ve yalnızca son aşamada benzerlik ölçümleri üzerinden ilişkilendirilir. İki akışlı (dual-encoder) mimariler bu yaklaşımın tipik örneklerini oluşturur. Bu yöntem hesaplama açısından verimli ve ölçeklenebilir olmakla birlikte, modaliteler arasında ayrıntılı etkileşim kurulmasına sınırlı imkân tanır.

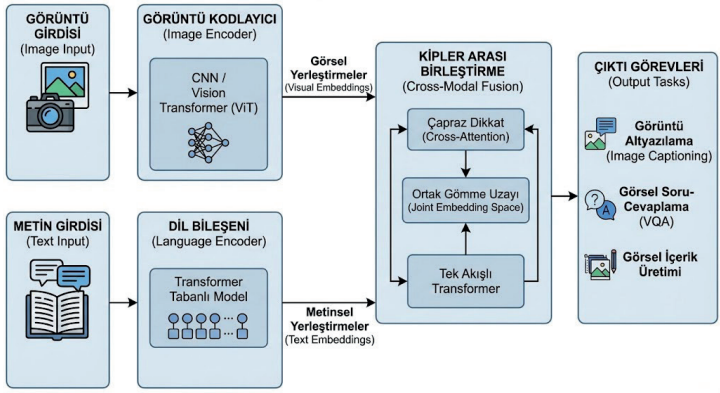
Erken birleşim (Early Fusion) yaklaşımında ise görsel ve metinsel temsiller, modelin giriş seviyesinde veya ilk katmanlarında birleştirilerek birlikte işlenir. Bu sayede her iki modalite baştan itibaren ortak bir temsil uzayında değerlendirilir.

Ancak bu yaklaşım, model karmaşıklığını ve hesaplama maliyetini artırabilmesi nedeniyle pratikte daha sınırlı kullanıma sahiptir.

Ara düzey birleşim ve çapraz dikkat (Cross-Attention Fusion), güncel VLM mimarilerinde en yaygın tercih edilen yaklaşımdır. Bu yöntemde görsel ve metinsel temsiller ayrı olarak elde edilir, ancak Transformer tabanlı çapraz dikkat katmanları aracılığıyla derinlemesine etkileşime sokulur. Tek akışlı (single-stream) ve iki akışlı (two-stream) mimariler, bu yaklaşımı farklı biçimlerde uygulayarak görsel ve dilsel bilgiler arasında çift yönlü bilgi akışı sağlar.

Çapraz dikkat mekanizmalarında, bir modalitenin temsilleri (ör. metin token'ları) diğer modaliteden (ör. görsel token'lar) bilgi "çekerek" hizalama kurar; böylece belirli kelimeler ile görüntünün ilgili bölgeleri arasında daha ince taneli ilişkiler öğrenilebilir.

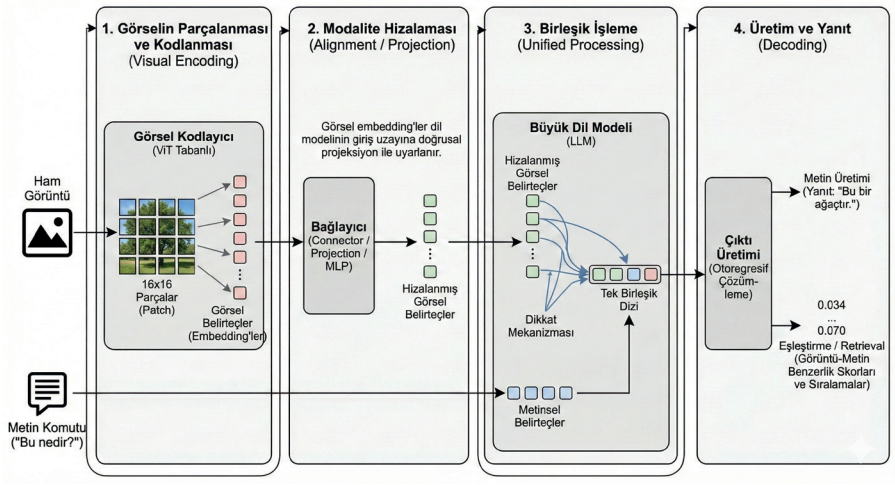
LLM tabanlı güncel görsel dil modellerde ise görsel kodlayıcı çıktısının dil modeline aktarılabilmesi için çoğunlukla projeksiyon/adapter benzeri bir köprü katmanı kullanılır; bu sayede görsel temsiller dil modelinin işleyebileceği biçime dönüştürülerek çapraz-modal etkileşim güçlendirilir.



Şekil 7: Görsel Dil Modelinin (VLM) genel mimarisi ve bileşenleri

3.3. Görsel–Dil Modellerinin İşleyişi

Görsel dil modelleri (VLM), görsel ve metinsel girdileri ortak bir anlamsal uzayda temsil ederek çok modlu çıkarım yapmayı hedefler. Bu modellerin çalışma mantığı, farklı modalitelerden gelen bilgilerin sayısal temsillere dönüştürülmesi, hizalanması ve birleşik bir bağlamda işlenmesi üzerine kuruludur. Genel iş akışı dört temel aşamada ele alınmaktadır (Şekil 8).



Şekil 8: Görsel dil modellerde uçtan uca iş akışı: (1) görsel kodlama (patch→görsel belirteçler), (2) modalite hizalaması (bağlayıcı/projeksiyon), (3) LLM içinde birleşik işleme ve dikkat temelli çok modlu muhakeme, (4) görev türüne bağlı çıktı (otoregresif metin üretimi veya görüntü–metin eşleştirme/retrieval).

3.3.1. Girdi Temsili: Görüntü ve Metin

Görsel dil modeller tipik olarak bir görüntü (tek/çoklu) ve buna eşlik eden bir metin girdisi (soru, komut veya prompt) alır. Metin girdisi tokenizer aracılığıyla belirteçlere (token) ayrılarak modele beslenir. Görüntü tarafında ise modelin doğrudan piksel alanında çalışması yerine, görsel içerik önce öğrenilebilir temsillere dönüştürülür. Bu dönüşüm, özellikle Transformer tabanlı görsel kodlayıcılarda, görüntünün sabit boyutlu parçalara (patch) ayrılması ve her parçanın bir "görsel belirteç" gibi temsil edilmesiyle gerçekleştirilir.

3.3.2. Görsel Kodlama: Görsel Temsillerin Üretilmesi

Görsel kodlayıcı (çoğunlukla CNN veya ViT tabanlı), görüntüden ya tek bir global embedding ya da daha sık olarak görsel token dizisi üretir. Token dizisi yaklaşımı, görüntünün farklı bölgelerine ait bilgiyi ayrı ayrı taşıdığı için özellikle ayrıntı gerektiren görevlerde (ör. VQA, doküman yorumlama) daha zengin bir temsil sağlar. Görsel kodlayıcının çıktısı, modelin sonraki aşamalarda metinle ilişkilendireceği "görsel kanıt" (visual evidence) niteliğindedir (bkz. Şekil 8, 1. aşama).

3.3.3. Modalite Hizalaması: Bağlayıcı/Projeksiyon Katmanı

Görsel kodlayıcıdan gelen temsiller ile dil modelinin giriş uzayı çoğu zaman aynı boyutta değildir; ayrıca bu temsillerin dağılım özellikleri de farklı olabilir. Bu nedenle görsel dil modeller, görsel temsilleri dil modelinin beklediği uzaya taşıyan bir bağlayıcı (connector) / projeksiyon bileşeni kullanır. En basit formda bu bileşen doğrusal bir projeksiyon veya MLP iken, bazı tasarımlarda daha güçlü bir ara modül (ör. sorgu tabanlı dönüştürücüler) kullanılabilir. Amaç, görsel temsillerin dil modelinin bağlamına "yerleştirilebilir" hâle getirilmesi ve metinsel belirteçlerle aynı dizide birlikte işlenebilmesidir (bkz. Şekil 8, 2. aşama).

3.3.4. Birleşik İşleme: LLM İçinde Çok Modlu Muhakeme

Hizalanmış görsel belirteçler ve metinsel belirteçler, dil modeli gövdesine tek bir birleşik dizi olarak verilir. Bu aşamada dikkat (attention) mekanizması, metindeki sorgu ile görüntüdeki ilgili bölgeleri temsil eden görsel belirteçler arasında bağ kurar. Böylece model, metinsel bağlamı görsel kanıtla koşullandırarak çok modlu muhakeme yapabilir. Pratikte bu birleşim, modelin "hangi nesne", "nerede", "kaç tane", "hangi özellikte" gibi sorulara yanıt üretebilmesini veya görüntü hakkında tutarlı açıklamalar oluşturabilmesini sağlar (bkz. Şekil 8, 3. aşama).

3.3.5. Çıktı Üretimi: Retrieval ve Üretim Modları

Görsel dil modellerinin çıktısı, mimarinin hedeflediği görev türüne göre iki ana biçimde ele alınabilir:

- *Eşleştirme/Retrieval Odaklı Çıktı:* Dual-encoder benzeri tasarımlarda model, görüntü ve metin temsillerinin benzerliğini hesaplayarak bir skor üretir ve adaylar arasında sıralama yapar. Bu yaklaşım, metinden görsel bulma veya görselden metin bulma gibi erişim senaryolarında tercih edilir.
- *Üretim Odaklı Çıktı:* LLM-tabanlı görsel dil modellerde model, birleşik bağlam üzerinden otoregresif biçimde metin üretir; her adımda bir sonraki belirteci tahmin ederek yanıtı oluşturur. Bu mod, görüntü açıklama, VQA ve görsel diyalog gibi görevlerde kullanılır (bkz. Şekil 8, 4. aşama).

3.4. Görsel–Dil Modellerinin Mimari Türleri

VLM mimarileri genellikle görsel ve metinsel temsillerin hangi aşamada ve ne ölçüde etkileştiği, modelin benzerlik skoru mu yoksa metin mi ürettiği

ve buna karşılık gelen hesaplama maliyeti–performans dengesi açısından sınıflandırılmaktadır.

3.4.1. Çift Kodlayıcı (Dual-Encoder) Yaklaşımlar

Dual-encoder mimarilerde görüntü ve metin, iki ayrı kodlayıcı tarafından bağımsız olarak gömlemelere (embeddings) dönüştürülür ve ilişkilendirme genellikle son aşamada bir benzerlik ölçümü (ör. kosinüs benzerliği) üzerinden yapılır. Bu yaklaşım, özellikle büyük ölçekli veri tabanlarında görüntü–metin eşleştirme ve arama (retrieval) senaryolarında pratik avantaj sağlar; çünkü aday öğelerin temsil vektörleri önceden hesaplanabilir ve hızlı biçimde karşılaştırılabilir. Bununla birlikte, modaliteler arası etkileşim son aşamada sınırlı kaldığından, ayrıntılı hizalama (ör. kelime–bölge ilişkisi) veya çok adımlı görsel akıl yürütme gerektiren görevlerde tek başına yetersiz kalabilir. Temsilci örnekler arasında CLIP ailesi yer almaktadır.

3.4.2. Kodlayıcı–Çözücü (Encoder–Decoder) Yaklaşımlar

Encoder–decoder mimariler, görsel içeriği bir kodlayıcı ile temsile dönüştürdükten sonra bu temsili kullanan bir çözücü ile metin üretmeyi hedefler. Bu çerçeve, görüntü açıklama (captioning), açık uçlu görsel soru–cevaplama (VQA) ve görsel içerikten rapor/özet üretimi gibi üretim odaklı görevler için doğrudan bir çözüm sunar. Temsillerin çözücü tarafında otoregresif biçimde üretilmesi, daha zengin doğal dil çıktıları sağlarken; çıkarım süresini uzatabilen bir maliyet de doğurur. Derleme literatüründe BLIP/BLIP-2 benzeri tasarımlar, hem görsel temsil kalitesi hem de üretim performansı açısından sık atıf alan örneklerdendir.

3.4.3. Unified / Single-Stream (Tek Akışlı Birleşik Transformer) Yaklaşımlar

Tek akışlı modellerde görsel ve metinsel belirteçler (tokens), tek bir Transformer içinde ortak bir dizi olarak işlenir. Bu tasarım, modalitelerin "başından itibaren" birbirini görmesine izin verdiği için, ince taneli çapraz-modal bağıntıların öğrenilmesini kolaylaştırır. Buna karşılık, özellikle retrieval gibi çok sayıda aday çiftin değerlendirilmesini gerektiren senaryolarda her çift için ortak işleme maliyeti ortaya çıktığından, hesaplama verimliliği dual-encoder yaklaşımlara kıyasla düşebilir. Tek akışlı modellerin güçlü olduğu alanlar çoğunlukla VQA, görsel çıkarım ve karmaşık hizalama gerektiren görevlerdir.

3.4.4. Donmuş LLM Tabanlı (Frozen Backbone + Adapter/ Mapping) Yaklaşımlar

Son dönemde yaygınlaşan bir hat, güçlü bir dil modelini (çoğunlukla decoder-only bir LLM) büyük ölçüde donmuş halde tutup, görsel kodlayıcıdan çıkan temsilleri dil modeline taşıyan hafif köprü katmanları (projeksiyon/ adapter) üzerinden çok modlu yetenek kazandırmaktır. Bu yaklaşımda amaç, LLM'in dilsel akıcılığı ve talimat takip kabiliyetini korurken, görsel bağlamı modele düşük parametre maliyetiyle entegre etmektir. Böylece çok modlu diyalog, görsel talimat izleme ve uzun biçimli açıklama üretimi gibi senaryolarda güçlü sonuçlar elde edilebilir. LLaVA sınıfı modeller bu hattın temsilci örnekleri arasında gösterilebilir. Bununla birlikte, model boyutu ve çıkarım maliyeti pratik uygulamalarda belirleyici olabilir; ayrıca görsel bilginin "ne kadar derin" işlendiği, kullanılan köprü tasarımının kapasitesine bağlıdır.

3.4.5. Türler Arası Karşılaştırma ve Görev Uyumları

Derleme çalışmalarında genel eğilim; retrieval/eşleştirme için dual-encoder mimarilerin ölçeklenebilirlik avantajı sunduğu, üretim ve talimat takip senaryolarında encoder–decoder veya LLM tabanlı (adapter'lı) yaklaşımların daha uygun olduğu, ayrıntılı hizalama ve çapraz-modal muhakeme gerektiren görevlerde ise single-stream tasarımların öne çıktığı yönündedir. Bu nedenle mimari seçimi; hedef görev, veri ölçeği, gecikme kısıtı ve donanım kaynaklarına göre birlikte değerlendirilmelidir (Tablo 2).

Tablo 2. Görsel–Dil Modellerinde (VLM) Kullanılan Temel Mimari Yaklaşımların Karşılaştırılması

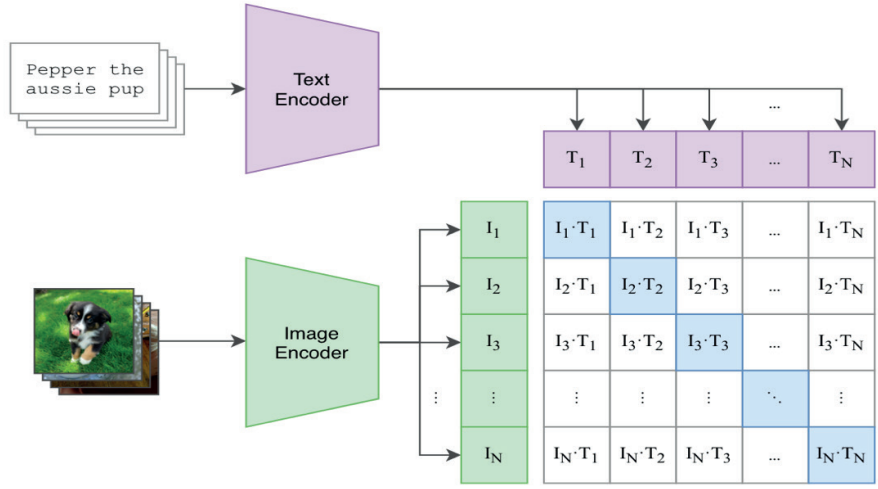
Mimari	Güçlü yön	Sınırlılık	Uygun görev
Çift Kodlayıcı	Hızlı, indekslenebilir, ölçeklenebilir	Derin görsel dil etkileşimi sınırlı	Retrieval / eşleştirme
Kodlayıcı–Çözücü	Metin üretiminde güçlü, görsele koşullu çıktı	Görev/etiketli veri bağımlılığı artabilir	Captioning, VQA
Tek Akışlı Birleşik Transformer	Erken ve yoğun etkileşim, muhakemede esneklik	Hesaplama maliyeti yüksek olabilir	Çok yönlü anlama-üretim
Donmuş LLM Tabanlı	Maliyet-etkin, hızlı uyarlama	Omurga kapasitesiyle sınırlanır	LLM'ye görsel yetenek ekleme

3.5. Görsel Dil Modellerinde Yaygın Olarak Kullanılan Modeller

Görsel–dil modelleri, mimari tasarımları ve hedefledikleri görevler açısından farklılaşmaktadır. Bu bölümde literatürde yaygın olarak kullanılan CLIP, BLIP, LLaVA, Kosmos ve Gemini modelleri kısaca incelenmektedir.

3.5.1. CLIP

CLIP, VLM alanında dual-encoder (çift kodlayıcı) yaklaşımının en bilinen örneklerinden biridir (Şekil 9). Model, görüntü ve metni iki ayrı kodlayıcıyla bağımsız biçimde temsil eder ve bu temsilleri ortak bir gömme uzayında hizalamayı hedefler. Eğitim sürecinde kontrastif öğrenme kullanılarak doğru görüntü–metin çiftlerinin yakınlaştırılması, yanlış eşleşmelerin ise uzaklaştırılması amaçlanır. Bu tasarım, özellikle görsel–metin eşleştirme ve çapraz erişim (retrieval) görevlerinde pratik avantaj sağlar; çünkü hem görseller hem de metinler ölçeklenebilir biçimde indekslenebilir. CLIP'in etkisi, sınıf etiketlerini doğal dil açıklamaları gibi ele alarak sıfır-atış (zero-shot) kullanım senaryolarında güçlü sonuçlar verebilmesiyle daha da belirginleşmiştir.

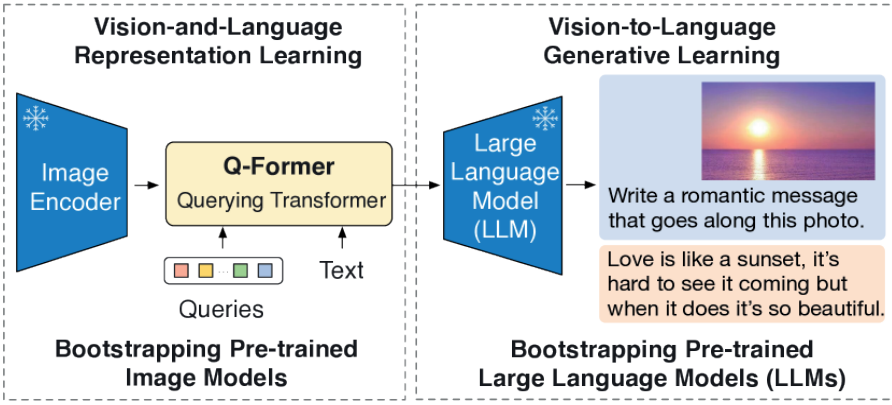


Şekil 9:CLIP modelinde metin ve görüntü girdilerinin ayrı kodlayıcılar tarafından ortak gömme uzayına projekte edilmesi ve tüm görüntü–metin çiftleri arasında benzerlik skorlarının hesaplanması.
Kaynak: OpenAI (2021), CLIP.

3.5.2. BLIP ve BLIP-2

BLIP ailesi, görsel dil modellerde hem anlama (ör. eşleştirme, VQA) hem de üretim (captioning) görevlerini destekleyen bir yaklaşım ortaya koyar (Şekil 10). BLIP'te görsel ve dil temsillerinin etkileşimi, görev ihtiyacına göre farklı düzenlemelerle ele alınabilir ve model, gürültülü web verisi gibi gerçek dünya kaynaklarından öğrenmeye uygun bir çerçeve sunmayı hedefler.

BLIP-2 ise daha verimli bir tasarım çizgisine yönelerek, güçlü bir görsel kodlayıcı ve güçlü bir dil modelini çoğunlukla dondurulmuş (frozen) biçimde kullanır; iki bileşeni bağlamak için hafif bir ara modül ile görsel bilginin LLM'ye aktarılmasını sağlar. Bu sayede tüm modeli baştan eğitmek yerine, sınırlı sayıda parametre ile etkili uyarlama yapılabilir. BLIP-2'nin bu yönü, "frozen backbone + bağlayıcı modül" yaklaşımının görsel dil modellerde neden popülerleştiğini gösteren iyi bir örnektir.

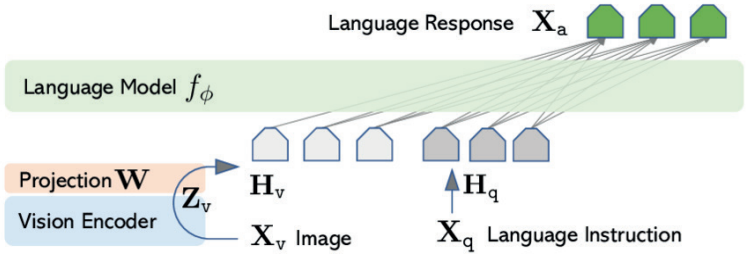


Şekil 10:BLIP-2 mimarisinde görsel kodlayıcıdan elde edilen temsillerin Q-Former (Querying Transformer) aracılığıyla büyük dil modeli (LLM) ile hizalanması ve görselden dile üretim süreci.

Kaynak: Salesforce BLIP-2, Weights & Biases raporu (2023).

3.5.3. LLAVA

LLaVA, CLIP tabanlı bir görsel kodlayıcıyı büyük bir dil modeliyle basit bir projeksiyon katmanı üzerinden bağlayan LLM-merkezli bir yaklaşımdır (Şekil 11). Görsel sohbet, talimat takibi ve açıklama üretimi gibi görevlerde başarılıdır. Bununla birlikte, küçük nesneler ve hassas görsel ayrıntılar konusunda performansı sınırlı olabilir.



Şekil 11: LLaVA mimarisinde görsel kodlayıcıdan elde edilen görsel temsillerin bir projeksiyon katmanı aracılığıyla büyük dil modeliyle hizalanması ve dil–görsel talimatlara dayalı yanıt üretim süreci.

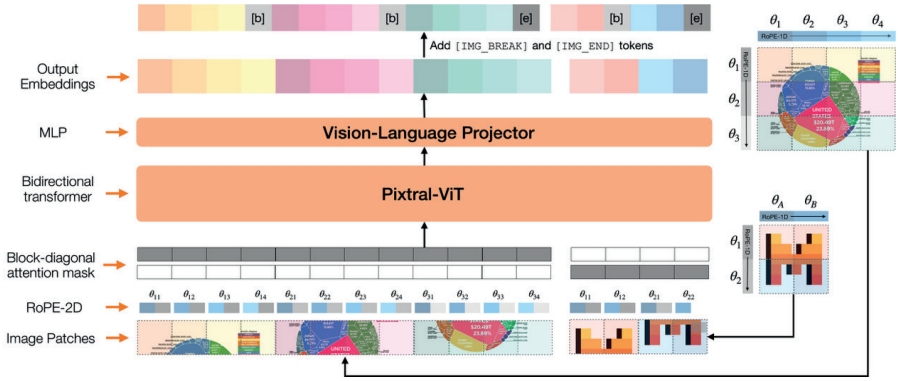
Kaynak: Encord (2023).

3.5.4. KOSMOS

Kosmos modelleri, görüntü ve metni tek bir Transformer mimarisi içinde işleyen birleşik (unified) bir yapıya sahiptir. Özellikle Kosmos-2 ile birlikte nesne konumlarını dilsel referanslarla ilişkilendiren grounding yeteneği öne çıkmıştır. OCR ve nesneye bağlı açıklamalarda güçlü olmakla birlikte, ölçek ve erişim açısından sınırlıdır.

3.5.5. PIXTRAL

Pixtral, Mistral AI tarafından geliştirilen ve görüntüyü dil modelinin bağlamına taşıyarak metin üreten (üretici) bir görsel–dil modelidir (Şekil 12). Görsel kodlayıcıdan elde edilen temsiller, büyük dil modeli (LLM) ile bütünleştirilerek görsel soru–cevaplama, görüntü açıklama ve özellikle doküman/şema yorumlama gibi görevlerde kullanılmaktadır. Uzun bağlam desteği ve çoklu görsel girdi senaryolarına uygun tasarımıyla dikkat çeken Pixtral, CLIP benzeri saf eşleştirme modellerine kıyasla daha yüksek hesaplama maliyetine sahiptir; performansı ise eğitildiği veri dağılımına (doküman ağırlıklı vs. doğal görüntüler) bağlı olarak değişkenlik gösterebilmektedir.



Şekil 12:Pixtral-Large mimarisinde görüntü yamalarının (image patches) Pixtral-ViT aracılığıyla işlenmesi, RoPE-2D ve blok-diyagonal attention maskesi ile konumsal bilginin modellenmesi ve görsel temsillerin görsel–dil projeksiyon katmanı üzerinden dil modeliyle hizalanması süreci.
Uyarlanmıştır: Encord (2024)

3.6. Görsel Dil Modellerinde Kullanılan Metrikler

Görsel dil modellerinin performansı tek bir metrikle temsil edilemez; çünkü modeller farklı görev sınıflarında (ör. eşleştirme/retrieval, açıklama üretimi, soru–cevaplama, grounding) farklı çıktı biçimleri üretir. Bu nedenle değerlendirme, görevin doğasına uygun ölçütlerle çok boyutlu biçimde ele alınmalıdır. Bu bölümde görsel dil modeller için literatürde yaygın kullanılan değerlendirme ölçütleri, görev ailelerine göre sınıflandırılarak özetlenmekte ve her bir metrik grubunun hangi tür çıktıları ölçmekte daha işlevsel olduğu kısaca açıklanmaktadır.

3.6.1. Görüntü Açıklama Görevleri için Metrikler (Image Captioning)

Bu metrikler, modelin bir resim için ürettiği metnin, insan tarafından yazılmış referans metne ne kadar benzediğini ölçer.

3.6.1.1. BLEU (Bilingual Evaluation Understudy)

Başlangıçta makine çevirisi için önerilmiş bir metriktir ve üretilen cümlenin referans cümlelerle paylaştığı *n*-gram sayısını ölçer. Aday cümledeki kaç kelimenin referans metinde (gerçek metin) geçtiğini dikkate alarak, hedef metnin referans metne göre doğruluğunu hesaplar.

$$BLEUScore = BP^{*}\exp\left(\sum_{i=1}^{N} W_i * \ln\left(p_i\right)\right)$$

- *BP(Brevity Penalty):* Aday cümlenin referans cümleye göre aşırı kısa olmasını cezalandıran terimdir.
- $w_i : i$. dereceden n-gram hassasiyetinin ağırlığıdır. Genellikle tüm n-gramlar için eşit seçilir:

$$w_i = \frac{1}{N}$$

- $p_i : i$. dereceden değiştirilmiş n-gram hassasiyetini (modified precision) ifade eder. Aday metindeki n-gram'ların referans metin(ler)deki karşılıklarıyla örtüşme oranını ölçer.
- N : Dikkate alınan maksimum n-gram derecesidir (çoğunlukla $N = 4$).

3.6.1.2. METEOR

Hassasiyet (precision) ve geri çağırma (recall) değerlerinin harmonik ortalamasını temel alan, ancak geri çağırma parametresine daha fazla ağırlık vererek bu bileşenleri bir ceza terimiyle çarpan bütüncül bir değerlendirme metriğidir. Tek bir parametreye odaklanan geleneksel metriklerin aksine, her iki veri kaynağından gelen bilgileri sentezleyerek daha kapsamlı bir analiz sunar. Eşanlamlı kelimeleri ve morfolojik varyasyonları dikkate alma yeteneği sayesinde cümle düzeyinde BLEU metriğine kıyasla daha esnek ve korelasyonu yüksek sonuçlar üretmekle birlikte; anlamsal eşdeğerliği tam olarak garanti edememesi, metriğin temel sınırlılığı olarak değerlendirilmektedir.

3.6.1.3. ROUGE-L

Otomatik metin özetlerini değerlendirmek için geliştirilmiştir. Bu metriğin bir varyasyonu olan ROUGE-L, en uzun ortak alt dizi (Longest Common Subsequence - LCS) yöntemine dayanarak metinler arasındaki yapısal örtüşmeyi hesaplamaktadır. Kapsam yönü oldukça güçlü olan bu yaklaşım, uzun cümleleri lehine değerlendirme eğilimi göstermekte ve serbest anlam değişimlerini (paraphrase) yakalamada belirli sınırlılıklar barındırmaktadır.

$$ROUGE-L = F\beta = \frac{\left(1+\beta^2\right) * P * R}{\beta^2 * P + R}$$

- *P(Precision):* En uzun ortak alt diziye (LCS) dayalı hassasiyet ölçüsüdür.
- *R (Recall):* En uzun ortak alt diziye (LCS) dayalı geri çağırma ölçüsüdür.

- β :Precision ve recall arasındaki dengeyi kontrol eden parametredir. Uygulamada genellikle $\beta = 1$ olarak seçilir ve bu durumda precision ile recall eşit ağırlıklandırılır

3.6.1.4. CIDEr (Consensus-based Image Description Evaluation)

Görüntü altyazılama (image captioning) görevlerine özgü olarak geliştirilmiş, konsensüs temelli bir değerlendirme metriğidir. Bu yöntem, referans altyazılardaki n-gram dizilerini TF-IDF ağırlıklandırma tekniğiyle analiz ederek, aday cümle ile referans kümesi arasındaki benzerliği kosinüs benzerliği (cosine similarity) üzerinden hesaplamaktadır. İnsan değerlendirmesiyle yüksek korelasyon sergilemesi ve ayırt edici terimleri ön plana çıkarması bakımından oldukça etkili bir ölçüt olan CIDEr; buna karşın, referans setinde yer almayan ancak görüntü içeriğiyle anlamsal olarak örtüşen alternatif ifadeleri düşük puanlandırma eğilimi göstermektedir.

3.6.1.5. SPICE (Semantic Propositional Image Caption Evaluation)

Metni nesne, öznitelik ve ilişkilerden oluşan sahne grafiklerine dönüştürerek anlamsal doğruluğu ölçen bir metriktir. İnsan yargılarıyla yüksek korelasyon (0,88) sergilemesiyle CIDEr ve METEOR gibi metriklerden ayrılan bu yöntem, içerik zenginliğini başarılı bir şekilde ödüllendirmektedir. Ancak dilbilgisel akıcılığı göz ardı etmesi ve yüksek hesaplama maliyeti sebebiyle, kapsamlı bir değerlendirme için genellikle BLEU gibi n-gram temelli metriklerle birlikte kullanılması tercih edilmektedir.

Değerlendirme metrikleri objektif bir ölçüt sunsa da BLEU, ROUGE ve CIDEr gibi n-gram temelli yöntemler, anlamsal eşdeğerliği ve görsel sadakati ölçmede sınırlı kalabilmektedir. Özellikle modellerin görüntüde bulunmayan nesneleri üretmesi (halüsinasyon) durumunda bu metrikler yetersiz kalabilmektedir. Bu nedenle, içerik tutarlılığını daha hassas analiz edebilmek için SPICE gibi anlamsal metriklerin veya doğrudan görüntü içeriğine dayalı doğruluk kontrollerinin kullanımı kritik önem arz etmektedir.

3.6.2. Görsel Soru-Cevaplama (VQA) için Metrikler

VQA görevlerinde model performansını çok boyutlu analiz edebilmek amacıyla, standart doğruluk ölçütlerinin yanı sıra görevin niteliğine (örn. metin okuma, anlamsal yakınlık) özgü farklı metrikler de kullanılmaktadır:

3.6.2.1. Accuracy (Doğruluk)

VQA literatüründeki en temel ölçüttür. Özellikle çoktan seçmeli veya kapalı uçlu sorularda, model çıktısının yer gerçekliği (ground truth) etiketiyle eşleşme durumunu ikili (binary) bir sistem üzerinden raporlar.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Burada TP (True Positive) ve TN (True Negative) sırasıyla doğru pozitif ve doğru negatif tahminleri; FP (False Positive) ve FN (False Negative) ise yanlış pozitif ve yanlış negatif tahminleri ifade etmektedir. Accuracy metriği, modelin tüm örnekler üzerindeki doğru sınıflandırma oranını ölçerek genel performansını özetler.

3.6.2.2. VQA Score (Consensus Accuracy)

Veri setinde tek bir "altın cevap" yerine birden fazla geçerli insan yanıtının bulunduğu durumlarda kullanılır. Modelin başarısı, ürettiği cevabın insan annotatörler arasındaki uzlaşı (konsensüs) kümesine olan yakınlığına göre değerlendirilir.

3.6.2.3. Exact Match (EM)

Model tarafından üretilen yanıtın referans yanıt ile karakter dizilimi bakımından birebir örtüşmesini şart koşar. Kısa ve kesin cevaplı görevlerde etkili bir ölçüt olmakla birlikte, morfolojik varyasyonlara veya eşanlamlı kullanımlara karşı toleransı düşüktür.

3.6.2.4. F1-Score

Yanıtın birden fazla kelimeden oluştuğu durumlarda, tahmin edilen ve referans metin arasındaki sözcük örtüşmesini hassasiyet (precision) ve geri çağırma (recall) metriklerinin harmonik ortalaması üzerinden dengeli bir şekilde ölçer.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Burada Precision, model tarafından üretilen sözcüklerin ne kadarının referans yanıtta yer aldığını; Recall ise referans yanıttaki sözcüklerin ne kadarının model tarafından doğru şekilde üretildiğini ifade eder.

3.6.2.5. ANLS (Average Normalized Levenshtein Similarity)

Özellikle görseldeki metni okumayı gerektiren (OCR-VQA) görevlerde tercih edilir. Tahmin ile referans metin arasındaki yazım benzerliğini Levenshtein düzenleme mesafesi üzerinden normalize ederek hesaplar; böylece küçük OCR hatalarına veya yazım yanlışlarına karşı esnek bir değerlendirme sağlar.

3.6.2.6. WUPS (Wu-Palmer Similarity)

Kelime düzeyindeki anlamsal eşdeğerliği ölçmek için WordNet taksonomisini kullanır. Model yanıtı ile referans arasındaki hiyerarşik anlamsal yakınlığı hesaplayarak, tam eşleşme olmasa dahi anlamsal açıdan doğru (eşanlamlı) yanıtları kısmi puanla ödüllendirir.

3.6.2.7. Top-k Accuracy

Geniş yanıt uzayına sahip veya olasılıksal dağılım üreten modellerde, doğru yanıtın modelin en yüksek olasılık atadığı ilk k tahmin içerisinde bulunma oranını ifade eder.

3.6.2.8. Human Evaluation (İnsan Değerlendirmesi)

Otomatik metriklerin anlamsal nüansları veya gerekçelendirmeyi (reasoning) yakalamada yetersiz kaldığı açık uçlu sorularda uygulanır. Yanıtlar; görsel dayanaklandırma (grounding), tutarlılık ve bağlamsal doğruluk gibi nitel kriterler üzerinden insan hakemler tarafından puanlanır.

3.6.3. Görsel-Metin Eşleştirme ve Retrieval Metrikleri

3.6.3.1. Recall@K (R@1, R@5, R@10)

Her sorgu için doğru eşleşmenin ilk K sonuç içinde yer alma oranını ölçer; retrieval çalışmalarında en sık raporlanan metriktir ve "ilk sonuçlar ne kadar isabetli?" sorusuna yanıt verir.

3.6.3.2. Median Rank (MedR)

Doğru eşleşmenin sıralamadaki konumlarının medyanıdır; "tipik olarak doğru eşleşme kaçıncı sırada geliyor?" bilgisini verir ve düşük olması daha iyidir.

3.6.3.3. Mean Rank (MR)

Doğru eşleşmenin sıralamadaki konumlarının ortalamasıdır; düşük değer daha iyidir ancak aykırı (çok geride kalan) örneklerden daha fazla etkilenir, bu yüzden çoğunlukla MedR ile birlikte raporlanır.

3.6.3.4. Mean Reciprocal Rank (MRR)

Doğru eşleşme üst sıralarda geldikçe daha yüksek puan verir (1/rank); özellikle her sorgu için tek doğru eşleşmenin olduğu senaryolarda sıralama kalitesini iyi özetler.

3.6.3.5. nDCG@K (Normalized Discounted Cumulative Gain)

İlgili sonuçları üst sıralarda bulmayı daha fazla ödüllendirir; bir sorgu için birden fazla ilgili örnek olduğunda R@K'ya göre daha anlamlı bir sıralama metriğidir.

3.6.3.6. mAP (mean Average Precision)

Sıralama boyunca ilgili sonuçları yakalama başarısını tek skorla özetler; çoklu ilgili örnek bulunan retrieval senaryolarında hem kapsama hem de sıralama kalitesini birlikte yansıtır.

3.6.3.7. Precision@K (P@K)

İlk K sonucun ne kadarının ilgili olduğunu ölçer; çoklu ilgili etiketlemesi varsa değerlidir, tek doğru eşleşme senaryosunda genellikle R@K daha yaygındır.

3.6.3.8. R-Precision

İlgili sonuç sayısı R ise, ilk R sonuçtaki precision değeridir; ilgili örnek sayısının sorguya göre değiştiği durumlarda sıralama kalitesini dengeli biçimde değerlendirmek için kullanılır.

3.6.4. Çok Modlu Üretim Görevlerinde Bütüncül Değerlendirme Metrikleri

Multimodal üretim modellerinin başarısı, günümüzde n-gram tabanlı yüzeysel ölçümlerin ötesine geçerek; anlamsal, yapısal ve olgusal tutarlılığı hedefleyen çok katmanlı bir değerlendirme rejimine dayanmaktadır. Literatürdeki temel yaklaşımlar şunlardır:

3.6.4.1. Gömme (Embedding) Tabanlı Metrikler

Üretilen içerik ile referans veri arasındaki ilişkiyi kelime örtüşmesi yerine, vektör uzayındaki anlamsal izdüşümler üzerinden analiz eder. BERTScore ve CLIPScore gibi yöntemler, sözdizimsel farklılıklara rağmen bağlamsal uyumu yakalayarak, modelin "anlama" kapasitesini ölçümler.

3.6.4.2. Yapısal ve İlişkisel Doğruluk

Dilbilgisi akıcılığından ziyade sahne içeriğine ve nesne ilişkilerine odaklanır. SPICE gibi metrikler, görseldeki varlıkların ve niteleyicilerin metne doğru aktarılıp aktarılmadığını bir "sahne grafiği" üzerinden denetleyerek, içerik bütünlüğünü ve olgusal doğruluğu sağlar.

3.6.4.3. Halüsinasyon ve Görsel Sadakat

Modelin görüntüde bulunmayan nesneleri üretme (halüsinasyon) sorununu ele alır. Klasik benzerlik testlerinin tespit edemediği bu fabrikasyon hataları, görsele sadakat (faithfulness) ilkesini temel alan özelleşmiş metriklerle (örn. CHAIR, POPE) saptanarak modelin güvenilirliği sınanır.

3.6.4.4. Model Tabanlı Değerlendirme (LLM-as-a-Judge)

Karmaşık ve açık uçlu görevlerde, gelişmiş dil modelleri (LLM) tanımlı rubrikler ışığında birer "hakem" olarak kullanılır. İnsan muhakemesine en yakın nüanslı sonuçları sunan bu yöntem, maliyet etkinliği ve ölçeklenebilirliği sayesinde modern araştırmalarda yaygınlaşmaktadır.

3.6.4.5. İnsan Değerlendirmesi

Otomatik metriklerin validasyonu için "altın standart" kabul edilir. Maliyet ve süre kısıtlarına rağmen; doğruluk, akıcılık ve bağlamsal nüansların en kesin ölçümü için hibrit değerlendirme süreçlerinin vazgeçilmez bir parçasıdır.

3.6.5. Görsel Dil Modelleri İçin Yeni Metrik Önerileri ve Benchmark Setleri

Görsel dil modellerinin değerlendirilmesinde klasik metrikler (ör. captioning'de BLEU/CIDEr, retrieval'da R@K) tek başına yeterli görülmemektedir; çünkü bu ölçütler yüksek skor üretse bile modelin görselde olmayan ayrıntıları eklemesi veya görsel kanıtla zayıf bağ kurması her zaman yakalanamayabilir. Bu nedenle literatürde eğilim, görsele sadakat/grounding ve muhakeme boyutlarını daha doğrudan ölçen değerlendirme kurulumlarına yönelmiştir.

Bu doğrultuda MMMU, MMBench ve MathVista gibi yeni nesil benchmark'lar modelin yalnızca görsel içerik algısını değil, çok-disiplinli bilgi kullanımı, sağlamlık ve matematiksel–mantıksal görsel muhakeme kapasitesini de sınamayı hedefler; HallusionBench ise halüsinasyon ve yanıltıcı görsel ipuçlarına karşı dayanıklılığı görünür kılar. Metrik tarafında POPE/H-POPE benzeri yaklaşımlar nesne/özellik düzeyinde halüsinasyonu daha kararlı biçimde ölçmeye odaklanırken, açık uçlu çıktılarda tek bir doğru tanımının zorlaşması

nedeniyle rubrik tabanlı model-hakemli değerlendirme (LLM/VLM-as-a-judge) kullanımı da yaygınlaşmaktadır. Genel olarak değerlendirme pratikleri, tek bir skor yerine doğruluk, sadakat ve muhakemeyi birlikte raporlayan çok boyutlu çerçevelere doğru evrilmektedir.

3.7. Görsel Dil Modellerinin Kullanım Alanları

3.7.1. Görsel Dil Modellerinin Otonom Sürüş Sistemlerindeki Kullanımı

3.7.1.1. Otonom Algı, Sahne Anlama ve Çevresel Farkındalık

Otonom sistemlerde güvenli ve bağlama duyarlı karar verme, çevrenin yalnızca geometrik olarak algılanmasını değil; sahnedeki nesnelerin, insanların ve çevresel koşulların anlamsal olarak yorumlanmasını gerektirmektedir. Bu gereksinim, görsel temsiller ile dil tabanlı muhakemeyi birleştiren Vision–Language Model (VLM) yaklaşımlarını otonom algı literatüründe önemli bir konuma taşımıştır. Son yıllarda yapılan çalışmalar, VLM'lerin 3B nesne tespiti, insan odaklı çevresel algı ve geniş ölçekli sokak görünümü analizlerinde açık sözcük dağarcığına dayalı, sıfır atışlı ve komutla yönlendirilebilir yetenekler sunduğunu göstermektedir (Sapkota et al., 2025; Greco et al., 2025; Peng et al., 2025).

3B nesne tespiti literatürünü VLM perspektifiyle sistematik biçimde inceleyen Sapkota ve arkadaşları (2025), geleneksel nokta bulutu ve voksel tabanlı yöntemlerin (PointNet++, VoteNet, PV-RCNN vb.) yoğun veri etiketleme gereksinimi, kapalı sınıf öğrenme yapısı ve semantik genelleme eksikliği gibi temel sınırlılıklarına dikkat çekmektedir. Derleme çalışması, bu sınırlamaları aşmaya yönelik geliştirilen yaklaşımları kapsamlı biçimde ele almaktadır. Dil–görsel hizalama temelli yöntemler (CLIP, PointCLIP v2), çok modlu büyük modeller (PaLM-E, BLIP-2, LLaVA, CogVLM) ve 3B sahnede dil temellendirme odaklı sistemler (3D-LLM, SpatialVLM, Cube R-CNN) bu çerçevede incelenmektedir. Bu modellerin ortak özelliği, görsel özellikleri dil uzayına projekte ederek referans ifadelerini çözümleyebilmesi ve açık sınıf tanıma yoluyla semantik temellendirmeyi güçlendirmesidir. OMNI3D (234K görüntü, 3M nesne etiketi), ScanNet, ScanRefer ve Nr3D/Sr3D veri setlerinde raporlanan sonuçlar; daha önce görülmemiş nesnelerin tespiti ve referans tabanlı 3B algıda anlamlı performans artışları göstermektedir. Örneğin CoDA çerçevesinde novel sınıflarda %80'e varan mAP artışı, 3DVLP modelinde ScanRefer üzerinde IoU@0.25 doğruluğunun %51.70'e ulaşması bu semantik genellemenin nicel göstergeleridir. Bununla birlikte çalışma, VLM tabanlı 3B sistemlerin yüksek hesaplama maliyeti, düşük kare hızları (örn. Instruct3D ≈

8 FPS) ve mekânsal halüsinasyon riski nedeniyle gerçek zamanlı otonom sürüş uygulamalarında doğrudan kullanımının sınırlı kaldığını vurgulamaktadır.

İnsan odaklı çevresel farkındalık alanında Greco ve arkadaşları (2025), yaya nitelik tanıma (Pedestrian Attribute Recognition, PAR) problemini VLM tabanlı bir çerçevede deneysel olarak değerlendirmiştir. Çalışmada çoklu nitelik çıkarımı klasik çok-etiketli sınıflandırma yaklaşımı yerine VQA formatında ele alınmıştır. Model, doğal dil soruları üzerinden aynı anda birden fazla niteliği çıkarabilmektedir. PAR 2023 değerlendirmelerinde MIVIA PAR, UPAR ve SYNTH-PEDES veri setleri üzerinde BLIP-2 tabanlı yaklaşım 0.921 ortalama doğruluk (mA) ile en yüksek performansı göstermiştir. Buna karşılık CNN tabanlı bir model 0.709 mA doğruluk üretmiş ancak 150 FPS hızına ulaşarak gerçek zamanlılık avantajı sağlamıştır. VLM tabanlı yaklaşımın <0.5 FPS işlem hızında kalması, doğruluk üstünlüğüne rağmen gerçek zamanlı video analizi açısından önemli bir sınırlılık oluşturmaktadır. Bu bulgular, görsel–dil ön-eğitiminin semantik nitelik çıkarımında avantaj sağladığını ancak hesaplama maliyetinin mühendislik açısından kritik olduğunu göstermektedir.

Daha geniş ölçekli çevresel farkındalık bağlamında Peng ve arkadaşları (2025), Street View Analytics literatürünü sistematik olarak incelemiş ve sokak görünümü görüntülerinin çok modlu analizine odaklanmıştır. Google Street View, Baidu Street View, Tencent Maps ve Mapillary gibi platformlardan elde edilen görüntüler; CLIP, BLIP-2, LLaVA, GPT-4V ve GPT-3.5 + görsel kodlayıcı kombinasyonları ile analiz edilerek kentsel güvenlik algısı, yürünebilirlik, ulaşım yoğunluğu ve sosyoekonomik göstergeler gibi üst düzey kavramların tahmini gerçekleştirilmiştir. Bu yaklaşımlar, görüntüye metin istemleri eşliğinde açık alan bilgisini entegre ederek yüksek seviyeli semantik çıkarım üretmektedir. Manuel etiketleme ihtiyacının azalması ve zengin semantik yorum üretimi önemli avantajlar sağlarken; verinin zamansal güncelliği, bağlamsal hassasiyet ve model kararlarının açıklanabilirliği temel sınırlılıklar olarak belirtilmektedir.

Genel olarak değerlendirildiğinde, Vision–Language Modelleri otonom algı, sahne anlama ve çevresel farkındalık alanlarında açık sözcük dağarcıklı genelleme, referans çözümleme ve insan merkezli semantik yorum üretimi açısından geleneksel yöntemlere kıyasla belirgin kazanımlar sağlamaktadır. Bununla birlikte yüksek hesaplama maliyeti, düşük kare hızları ve gecikme gibi mühendislik kısıtları nedeniyle bu modellerin kısa vadede bağımsız gerçek zamanlı algı modülleri olarak değil, üst seviye semantik muhakeme ve karar destek bileşenleri olarak konumlandırılması daha uygulanabilir bir yaklaşım olarak görünmektedir (Sapkota et al., 2025; Greco et al., 2025; Peng et al., 2025). Tablo 3' de Vision–Language Modellerinin otonom algı ve çevresel farkındalık görevlerinde kullanımının karşılaştırmalı analizi verilmiştir.

Tablo 3. Vision–Language Modellerinin otonom algı ve çevresel farkındalık görevlerinde kullanımı

Özellik	3D Nesne Tespiti ve Görsel Dil Modeller	Yaya Nitelik Tanıma (PAR)	Sokak Görünümü Analitiği (SVA)
İlgili Çalışma	*A Review of 3D Object Detection with Vision–Language Models* (Sapkota et al., 2025)	*An Experimental Evaluation of Smart Sensors for Pedestrian Attribute Recognition Using Multi-Task Learning and Vision Language Models* (Greco et al., 2025)	*VLM-enabled Street View Analytics: A Systematic Literature Review* (Peng et al., 2024)
Çalışmanın Amacı	Geleneksel nokta bulutu ve voksel tabanlı 3D tespit yöntemlerinin sınırlamalarını aşarak, açık sözcük dağarcığı (open-vocabulary), sıfır atışlı ve dil komutlarına duyarlı 3D nesne algısı sağlamak	Düşük çözünürlüklü güvenlik kamerası görüntülerinden yaya niteliklerini (cinsiyet, kıyafet rengi, aksesuar vb.) çoklu görevli ve VLM tabanlı yaklaşımlarla doğru biçimde tanımak	Sokak görünümü görüntülerini kullanarak kentsel güvenlik, yürünebilirlik, çevresel algı ve sosyoekonomik göstergeleri çok modlu biçimde analiz etmek
Kullanılan Model / Mimari	Geleneksel: PointNet++, VoteNet, PV-RCNN VLM tabanlı: PaLM-E, BLIP-2, LLaVA, CogVLM, CLIP, PointCLIP v2, 3D-LLM, SpatialVLM, Cube R-CNN	Kazanan yaklaşım: **BLIP-2 (VQA olarak kullanılmış)** Diğerleri: ResNet50 + PLIP, Swin Transformer varyantları, BeitV2, EfficientNetV2	GPT-4V, GPT-3.5 + görsel kodlayıcılar, CLIP, BLIP-2, LLaVA, LLaMA
Temel Teknik Yaklaşım	Dil–görsel hizalama ile 3D sahnelerde semantik temellendirme; sıfır atışlı tespit; metin tabanlı referans çözümleme (ör. "kırmızı sandalyenin yanındaki çanta")	Problemin görsel sınıflandırma yerine **VQA olarak formüle edilmesi;** çoklu niteliklerin tek modelle çıkarılması	Görüntü + metin istemleri ile yüksek seviyeli çevresel çıkarım; açık alan bilgisi (open-domain knowledge) ile kentsel yorumlama
Kullanılan Veri Setleri	KITTI, SUN RGB-D, Waymo Open Dataset, ScanNet, ScanRefer, Nr3D/Sr3D, **OMNI3D** (234K görüntü, 3M nesne etiketi)	Eğitim: MIVIA PAR (105.244 görüntü), UPAR, SYNTH-PEDES Test: 20.000 görüntülük gizli test seti	Google Street View (GSV), Baidu Street View, Tencent Maps, Mapillary + OpenStreetMap (OSM)
Değerlendirme Metrikleri	mAP, IoU@0.25 / IoU@0.5, doğruluk, FPS	Ortalama doğruluk (mA), standart sapma, FPS	Uygulamaya bağlı metrikler (sınıflandırma doğruluğu, korelasyon, nitel analiz)
Öne Çıkan Nicel Sonuçlar	CoDA: Novel object discovery'de %80'e varan mAP artışı 3DVLP: ScanRefer'da %51.70 doğruluk (IoU@0.25)	BLIP-2 kullanan iROC-ULPGC: **0.921 mA** (en yüksek doğruluk) Baseline CNN: 0.709 mA fakat 150 FPS	Görsel dil modeller, kentsel güvenlik ve yürünebilirlik gibi kavramları insan algısına yakın biçimde modelleyebilmiştir

Avantajlar	Sıfır atışlı öğrenme, dil komutlarına duyarlılık, anlamsal genelleme	Çok yüksek doğruluk, tek modelle çoklu nitelik çıkarımı	Manuel etiket ihtiyacını azaltma, zengin semantik açıklamalar
Sınırlılıklar	Yüksek hesaplama maliyeti, düşük FPS (ör. Instruct3D ≈ 8 FPS), mekansal halüsinasyonlar	Gerçek zamanlı kullanım için çok yavaş (<0.5 FPS)	Zamansal güncellik, bağlamsal hassasiyet ve karar sürecinde şeffaflık eksikliği

3.7.1.2. Otonom Karar Verme, Planlama ve Navigasyon

Otonom sürüş sistemlerinde karar verme, planlama ve navigasyon bileşenleri; algılanan çevresel bilginin güvenli, tutarlı ve açıklanabilir sürüş davranışlarına dönüştürülmesinden sorumludur. Geleneksel modüler yaklaşımlar, karmaşık ve nadir görülen (long-tail) senaryolarda insan benzeri muhakeme gerektiren durumlarda sınırlı kalmaktadır. Son yıllarda Vision–Language Model (VLM) ve Large Language Model (LLM) tabanlı yaklaşımlar, sürüş problemini yalnızca geometrik bir optimizasyon olarak değil; dil destekli akıl yürütme ve bağlamsal karar verme çerçevesinde ele alarak bu sınırlamaları aşmayı hedeflemektedir (Liu et al., 2023; Wang et al., 2024).

Bu doğrultuda öne çıkan VLM-AD yaklaşımının temel amacı, uçtan uca otonom sürüş modellerindeki muhakeme eksikliğini gidermektir (Chen et al., 2024). Önerilen mimaride, güçlü bir VLM (GPT-4o) doğrudan bir kontrolcü olarak değil, yalnızca eğitim aşamasında bir öğretmen model olarak konumlandırılır. Sürüş sahnesi, planlanan yörünge ve araç durumu üzerinden üretilen doğal dil açıklamaları ile yapılandırılmış sürüş eylemleri; bilgi damıtma (knowledge distillation) yöntemiyle UniAD ve VAD gibi uçtan uca sürüş modellerine aktarılır. Test aşamasında VLM'nin tamamen devre dışı bırakıldığı bu tasarım, sistemin gerçek zamanlı çalışmasını sağlarken dil tabanlı muhakeme bilgisinin sürüş davranışlarına yansıtılmasına olanak tanır.

Elde edilen deneysel sonuçlar, VLM-AD mimarisinin planlama doğruluğu ve güvenlik açısından anlamlı kazanımlar sağladığını göstermektedir. nuScenes veri setinde UniAD için raporlanan ortalama L2 planlama hatası 1.03 m iken, VLM-AD ile bu değer yaklaşık 0.88–0.89 m seviyesine düşmüş; çarpışma oranı ise %0.31'den %0.19–0.24 aralığına gerilemiştir (Chen et al., 2024). CARLA simülasyonlarında yapılan kapalı döngü deneyler de dil destekli damıtmanın daha kararlı sürüş davranışları ürettiğini ortaya koymuştur.

Bu tür görev-odaklı çalışmaların metodolojik arka planı literatürdeki kapsamlı derlemelerde de vurgulanmaktadır. *A Survey on Large Language Model-Powered Autonomous Driving* çalışması, büyük dil modellerinin süreçlere entegrasyonunu inceleyerek modelleri muhakeme yeteneklerine göre sınıflandırırken (Liu

et al., 2023); *Vision Language Models in Autonomous Driving: A Survey and Outlook* çalışması görsel dil modellerinin uçtan uca sürüşteki rollerini ve dil zenginleştirilmiş veri setlerinin önemini ele almaktadır (Wang et al., 2024). Sonuç olarak; yüksek hesaplama maliyeti, gecikme ve uzamsal halüsinasyonlar gibi donanımsal ve sistemsel darboğazlar göz önüne alındığında, VLM'lerin doğrudan kontrolcü yerine eğitim aşamasında rehberlik eden bileşenler olarak kullanılması kısa vadede en uygulanabilir mimari yaklaşım olarak öne çıkmaktadır (Chen et al., 2024; Liu et al., 2023; Wang et al., 2024). Tablo 4' de otonom karar verme, planlama ve navigasyon için VLM/LLM tabanlı çalışmaların teknik karşılaştırması sunulmuştur.

Tablo 4. Otonom Karar Verme, Planlama ve Navigasyon için VLM/LLM Tabanlı Çalışmaların Teknik Karşılaştırması

Özellik	VLM-AD	LLM-Powered AD Survey	VLM in AD Survey
Çalışmanın Amacı	Dil tabanlı muhakemeyi uçtan uca sürüşe aktarmak	Büyük dil modelleriin otonom sürüşteki rolünü sınıflandırmak	Görsel dil modellerinin AD'deki kullanım alanlarını özetlemek
Kullanılan Modeller	GPT-4o (öğretmen) + UniAD / VAD	GPT-Driver, DriveVLM, DriveGPT4	CLIP, LLaVA, GPT-4V
Dilin Rolü	Eğitimde açıklama + eylem damıtma	Karar verme ve açıklanabilirlik	Açık sözcük dağarcıklı algı
Veri Setleri	nuScenes, CARLA	nuScenes, BDD100K, CARLA	nuScenes-QA, Talk2Car, DriveLM
Performans Metrikleri	L2 hata, çarpışma oranı	Çeşitli (planlama, güvenlik)	Görev-odaklı doğruluk
Temel Sonuç	L2 hata ↓, çarpışma ↓	Long-tail senaryolarda güçlü	Algı–planlama entegrasyonu
Kaynak	Chen et al. (2024)	Liu et al. (2023)	Wang et al. (2024)

3.7.1.3. Otonom Sistemler için Veri Üretimi, Etiketleme ve Edge (Uç) Uygulamaları

Otonom sistemlerin gerçek dünya koşullarında güvenilir biçimde çalışabilmesi, yalnızca gelişmiş algı ve karar verme modellerine değil; bu modellerin eğitimi için gerekli yüksek kaliteli verinin ölçeklenebilir biçimde üretilmesine ve kaynak kısıtlı uç (edge) cihazlarda çalıştırılabilmesine de bağlıdır. Vision–Language Modelleri (VLM), bu iki gereksinimi birlikte ele alabilen yaklaşımlar sunarak literatürde giderek daha merkezi bir rol üstlenmektedir.

Otonom sürüş alanında en yüksek maliyetli süreçlerden biri olan 3D nesne etiketleme problemine çözüm getirmek amacıyla Ma ve arkadaşları, VLM

destekli yarı-otomatik bir veri üretim çerçevesi önermektedir (Ma et al., 2026). Önerilen sistemin teknik mimarisinde, stereo görüntülerden derinlik tahmini yapan PSMNet, görüntü kenar bilgisi ile seyrek LiDAR ölçümlerini birleştiren Frustum Graph Depth Correction (FGDC) modülü ve Transformer tabanlı TransFusion 3D dedektöründen oluşan çok aşamalı bir boru hattı (pipeline) kurgulanmıştır. Bu yapıda, seyrek LiDAR verileri kamera derinlik bilgisiyle zenginleştirilerek yoğun bir pseudo-LiDAR temsili oluşturulur. Elde edilen 3D sınırlayıcı kutular (bounding box), 2D görüntü düzlemine yansıtılarak GLM-4 tabanlı bir VLM'e ardışık soru-cevap istemleriyle (örneğin "kutu bir yaya içeriyor mu?") sunulur ve hatalı etiketler semantik olarak filtrelenir.

Bu sistemin deneysel sonuçları, VLM tabanlı doğrulama mekanizmasının manuel etiketlemeye kıyasla yaklaşık yirmi kat hızlanma sağladığını ve %99'un üzerinde filtreleme hassasiyetine ulaştığını kanıtlamaktadır. Ayrıca, otomatik üretilmiş bu etiketlerle eğitilen algı modellerinin özellikle araç tespitinde belirgin performans artışları sergilediği rapor edilmiştir (Ma et al., 2026).

Veri üretimindeki bu başarıların pratik sistemlere entegrasyonu, VLM'lerin uç cihazlarda çalıştırılabilmesine yönelik mühendislik çözümlerini zorunlu kılmaktadır. Sharshar ve arkadaşları tarafından sunulan kapsamlı derleme çalışması, büyük ölçekli görsel dil modellerinin IoT, mobil ve gömülü sistemler gibi kaynak kısıtlı platformlara nasıl uyarlanabileceğini sistematik olarak incelemeyi hedeflemektedir (Sharshar et al., 2024). Analiz sonuçlarına göre, büyük VLM'lerin doğrudan uç cihazlara taşınması gecikme, bellek ve enerji tüketimi açısından sürdürülebilir değildir. Bu darboğazı aşmak için kurgulanan uç cihaz mimarilerinde, bilgi damıtma, niceleme ve token azaltma gibi model sıkıştırma teknikleri merkeze alınmaktadır. Örneğin, MobileVLM-V2 gibi yapılar, görsel belirteç sayısını azaltan hafif projeksiyon katmanları sayesinde mobil donanımlarda kabul edilebilir gecikmelerle çok modlu çıkarım yapılmasına olanak tanır.

Ayrıca bu çalışma, sürüş verilerinin merkezi sunuculara aktarılmadan işlenmesini sağlayan federe öğrenme (federated learning) yaklaşımlarının gizlilik risklerini ve ağ gecikmelerini ortadan kaldırdığını; ancak uç cihaz mimarilerindeki bu yerel modellerin fiziksel saldırılara ve model zehirlenmelerine karşı hala savunmasız olduğunu vurgulamaktadır (Sharshar et al., 2024). Genel bir değerlendirmeyle; görsel dil modelleri bir yandan etiketleme süreçlerinde insan maliyetini düşüren semantik denetleyiciler olarak işlev görürken, diğer yandan kısa vadede tam ölçekli genel modeller yerine, sisteme entegre edilmiş, damıtılmış ve görev-özgü mimariler olarak uç cihazlarda gerçek zamanlı algı yetenekleri sunmaktadır (Ma et al., 2026; Sharshar et al., 2024). Otonom sistemlerde veri üretimi ve uç uygulamalar için Vision–Language Model yaklaşımlarının karşılaştırılması Tablo 5' de verilmiştir.

Tablo 5. Otonom Sistemlerde Veri Üretimi ve Uç Uygulamalar için Vision–Language Model Yaklaşımlarının Karşılaştırılması

Özellik	VLM Destekli Otomatik 3D Nesne Etiketleme	Uç Ağlar için Vision–Language Modelleri (VLM)
Kaynak	Ma et al. (2026)	Sharshar et al. (2024)
Çalışmanın Amacı	Otonom sürüş için yüksek maliyetli manuel 3D etiketleme sürecini azaltmak; kamera ve LiDAR verilerini kullanarak yarı-otomatik ve güvenilir 3D etiket üretmek	Büyük ölçekli görsel dil modellerinin kaynak kısıtlı uç (edge) cihazlarda düşük gecikme, düşük enerji tüketimi ve gizlilik korunarak çalıştırılabilmesini sağlamak
Temel Yaklaşım	Pseudo-LiDAR üretimi + Transformer tabanlı 3D tespit + VLM tabanlı semantik doğrulama ile kapalı döngü etiketleme	Model sıkıştırma, hafif mimariler ve federated learning kullanarak görsel dil modelleri uç cihazlara uyarlama
Kullanılan VLM / LLM	**GLM-4** (etiket doğrulama ve semantik filtreleme için)	**CLIP, MobileVLM-V2, Moondream2, MiniVLM, EdgeVL**
Görsel Algı Modelleri	PSMNet (stereo derinlik), FGDC (derinlik düzeltme), TransFusion (3D nesne tespiti)	CLIP-tabanlı görsel kodlayıcılar, tek-akışlı ve token-azaltmalı VLM mimarileri
Teknik Yenilik	• Stereo görüntülerden pseudo-LiDAR üretimi • 3D kutuların 2D projeksiyonu • VLM ile ardışık soru-cevap tabanlı etiket doğrulama	• Knowledge distillation (bilgi damıtma) • Quantization (8-bit ve altı) • Lightweight Downsampling Projection (LDP) • Federated learning
Kullanılan Veri Setleri	GA Dataset, TJ Dataset (gerçek dünya); KITTI (kıyaslama); nuScenes (farklı hava/gece koşulları)	Otonom sürüş, sağlık ve uzaktan algılama alanlarından çeşitli veri setleri (AerialVLN, MedBLIP, ChangeCLIP vb.)
Değerlendirme Metrikleri	Etiket doğrulama hassasiyeti (Precision), Ortalama Doğruluk (AP), Etiketleme süresi	Gecikme (latency), bellek kullanımı, enerji tüketimi, doğruluk, cihaz uyumluluğu
Nicel Sonuçlar	• Manuel etiketlemeye göre **~20× hızlanma** (300 sn → 15.2 sn) • VLM doğrulama hassasiyeti **>%99** • BEVDet performansı: büyük araçlarda **+%38 AP,** küçük araçlarda **+%15.6 AP**	• MobileVLM ve Moondream2 gibi modellerle uç cihazlarda gerçek zamanlı çıkarım mümkün • Bulut bağımlılığı ve ağ gecikmesi belirgin biçimde azaltılmış
Avantajlar	Yüksek kaliteli etiket üretimi, düşük insan maliyeti, ölçeklenebilir veri üretimi	Düşük gecikme, gizlilik korunumu, enerji verimliliği
Sınırlılıklar	Büyük görsel dil modeller sunucu tarafında çalışmak zorunda	Sıkıştırma sonrası anlamsal kapasite kaybı, güvenlik (model poisoning) riskleri
Otonom Sistemlere Katkı	Algı modelleri için ölçeklenebilir ve güvenilir eğitim verisi üretimi	Gerçek zamanlı, gizlilik dostu ve sahaya uyarlanabilir VLM tabanlı algı

3.7.2. Görsel Dil Modellerinin Robotik Sistemlerindeki Kullanımı

3.7.2.1. VLM/VLA Tabanlı Robot Kontrolü ve Gerçek Dünya Uygulamaları

Robotik alanda Vision–Language–Action (VLA) modelleri, görsel algı (vision), dil temsilleri (language) ve eylem üretimini (action) tek bir uçtan uca öğrenme çerçevesinde birleştirerek, robotların doğal dil ile tanımlanan görevleri yerine getirebilmesini hedeflemektedir. Bu yaklaşım, klasik modüler robotik mimarilerde ayrı ayrı ele alınan algı, planlama ve kontrol bileşenlerinin, yüksek seviyeli anlamsal temsiller aracılığıyla doğrudan ilişkilendirilmesini amaçlamaktadır.

Bu doğrultuda geliştirilen RT-2 (Robotics Transformer-2) modeli, büyük ölçekli görsel–dil modellerinin web verilerinden edindiği semantik bilginin robotik kontrol politikalarına aktarılabildiğini gösteren önemli bir örnek olarak rapor edilmiştir (Brohan et al., 2023). RT-2'de robot eylemleri, ayrık metin belirteçleri (tokens) olarak temsil edilmekte; böylece robot kontrol problemi, doğal dil üretimine benzer bir ardışık tahmin problemi şeklinde ele alınmaktadır. Bu kapsamda robotun altı serbestlik dereceli (6-DoF) konumu, yönelimi ve tutucu (gripper) durumu ayrıklaştırılarak modelin çıktı kelime dağarcığına dâhil edilmektedir (Brohan et al., 2023). Bu yaklaşım literatürde action-as-text tokenization olarak adlandırılmaktadır.

Mimari açıdan RT-2, PaLI-X (5B ve 55B parametre) ve PaLM-E (12B parametre) gibi büyük ölçekli Vision–Language Model (VLM) omurgaları üzerine inşa edilmiştir (Brohan et al., 2023). Eğitim sürecinde robotik trajektori verileri ile web ölçekli görsel–dil görevlerinin (ör. görsel soru-cevaplama ve görüntü altyazılama) birlikte kullanıldığı bir co-fine-tuning stratejisi uygulanmaktadır. Bu strateji, modelin robotik görevlerde uzmanlaşırken, web verilerinden öğrenilen genel semantik bilgiyi korumasını amaçlamaktadır.

Deneysel değerlendirmelerde RT-2'nin masaüstü manipülasyon görevlerinde (ör. pick-and-place, nesne ayırma ve hedefe taşıma) görülmemiş nesneler, arka planlar ve dilsel komutlar altında önceki modellere kıyasla daha yüksek başarı oranları elde ettiği rapor edilmiştir (Brohan et al., 2023). Çalışmada, bazı zero-shot senaryolarda başarı oranının önceki yaklaşımlara göre 2 ila 6 kat arttığı belirtilmektedir. Ayrıca modelin, robotik eğitim verisinde açıkça yer almayan bağlamsal ve ilişkisel talimatları yorumlayabildiği gözlemlenmiştir (Brohan et al., 2023).

VLA literatürünü daha geniş bir perspektiften ele alan Sapkota ve arkadaşları, bu modelleri tekil uygulama başarılarından ziyade genel amaçlı robotik ajanlara doğru evrilen bir mimari paradigma olarak konumlandırmaktadır (Sapkota et al., 2025). İlgili survey çalışmasında 80'den fazla VLA modeli incelenmiş

ve bu modeller erken füzyon (early fusion), çift sistemli (dual-system) ve kendi kendini düzelten (self-correcting) mimariler olmak üzere üç ana grupta sınıflandırılmıştır. Çift sistemli mimarilerde, hızlı tepki veren düşük seviyeli kontrol modülleri ile daha yavaş ancak muhakeme odaklı planlama modüllerinin ayrıştırıldığı ifade edilmektedir (Sapkota et al., 2025).

Aynı çalışmada, milyarlarca parametreye sahip VLA modellerinin pratik robotik sistemlerde çalıştırılabilmesi için LoRA, niceleme (quantization) ve düşük dereceli adaptasyon gibi parametre-verimli öğrenme tekniklerinin kritik öneme sahip olduğu vurgulanmaktadır (Sapkota et al., 2025). Bununla birlikte, otoregresif belirteç üretimine dayalı VLA modellerinin genellikle 3–5 Hz gibi düşük kontrol frekanslarında çalışabildiği ve bunun gerçek zamanlı ve güvenli robot kontrolü açısından önemli bir sınırlama oluşturduğu belirtilmektedir.

Gerçek dünya uygulamalarına odaklanan Kawaharazuka ve arkadaşları ise VLA modellerini veri toplama stratejileri, robot donanımı ve eğitim paradigmaları ile birlikte ele almaktadır (Kawaharazuka et al., 2023). Bu çalışmada sensörimotor modeller, dünya modelleri ve sağlar (affordance) temelli yaklaşımlar karşılaştırılmış; özellikle sürekli eylem uzaylarında difüzyon tabanlı politika başlıklarının daha pürüzsüz ve kararlı kontrol sağladığı rapor edilmiştir (Kawaharazuka et al., 2023). Ayrıca, farklı robot platformları arasında öğrenilen temsillerin doğrudan aktarılmasının performans kaybına yol açabildiği ve embodiment transfer probleminin hâlen açık bir araştırma konusu olduğu vurgulanmaktadır.

Genel olarak değerlendirildiğinde, VLA modellerinin robotik sistemlere dil temelli muhakeme, görev genellemesi ve açıklanabilirlik kazandırdığı görülmektedir. Ancak bu kazanımların; hesaplama maliyeti, düşük kontrol frekansı, güvenlik garantileri ve donanım bağımlılığı gibi mühendislik kısıtlarıyla dengelenmesi gerektiği literatürde açıkça ifade edilmektedir (Brohan et al., 2023; Sapkota et al., 2025; Kawaharazuka et al., 2023). VLA modellerinin robotik alandaki karşılaştırmalı analizi Tablo 6' da verilmiştir.

Tablo 6. Vision–Language–Action (VLA) Modellerinin Robotik Alandaki Karşılaştırmalı Özeti

Özellik	VLM Destekli Otomatik 3D Nesne Etiketleme	Uç Ağlar için Vision–Language Modelleri (VLM)
Kaynak	Ma et al. (2026)	Sharshar et al. (2024)
Çalışmanın Amacı	Otonom sürüş için yüksek maliyetli manuel 3D etiketleme sürecini azaltmak; kamera ve LiDAR verilerini kullanarak yarı-otomatik ve güvenilir 3D etiket üretmek	Büyük ölçekli Görsel dil modellerinin kaynak kısıtlı uç (edge) cihazlarda düşük gecikme, düşük enerji tüketimi ve gizlilik korunarak çalıştırılabilmesini sağlamak
Temel Yaklaşım	Pseudo-LiDAR üretimi + Transformer tabanlı 3D tespit + VLM tabanlı semantik doğrulama ile kapalı döngü etiketleme	Model sıkıştırma, hafif mimariler ve federated learning kullanarak görsel dil modelleri uç cihazlara uyarlama
Kullanılan VLM / LLM	**GLM-4** (etiket doğrulama ve semantik filtreleme için)	**CLIP, MobileVLM-V2, Moondream2, MiniVLM, EdgeVL**
Görsel Algı Modelleri	PSMNet (stereo derinlik), FGDC (derinlik düzeltme), TransFusion (3D nesne tespiti)	CLIP-tabanlı görsel kodlayıcılar, tek-akışlı ve token-azaltmalı VLM mimarileri
Teknik Yenilik	• Stereo görüntülerden pseudo-LiDAR üretimi • 3D kutuların 2D projeksiyonu • VLM ile ardışık soru-cevap tabanlı etiket doğrulama	• Knowledge distillation (bilgi damıtma) • Quantization (8-bit ve altı) • Lightweight Downsampling Projection (LDP) • Federated learning
Kullanılan Veri Setleri	GA Dataset, TJ Dataset (gerçek dünya); KITTI (kıyaslama); nuScenes (farklı hava/gece koşulları)	Otonom sürüş, sağlık ve uzaktan algılama alanlarından çeşitli veri setleri (AerialVLN, MedBLIP, ChangeCLIP vb.)
Değerlendirme Metrikleri	Etiket doğrulama hassasiyeti (Precision), Ortalama Doğruluk (AP), Etiketleme süresi	Gecikme (latency), bellek kullanımı, enerji tüketimi, doğruluk, cihaz uyumluluğu
Nicel Sonuçlar	• Manuel etiketlemeye göre **~20× hızlanma** (300 sn → 15.2 sn) • VLM doğrulama hassasiyeti **>%99** • BEVDet performansı: büyük araçlarda **+%38 AP,** küçük araçlarda **+%15.6 AP**	• MobileVLM ve Moondream2 gibi modellerle uç cihazlarda gerçek zamanlı çıkarım mümkün • Bulut bağımlılığı ve ağ gecikmesi belirgin biçimde azaltılmış
Avantajlar	Yüksek kaliteli etiket üretimi, düşük insan maliyeti, ölçeklenebilir veri üretimi	Düşük gecikme, gizlilik korunumu, enerji verimliliği
Sınırlılıklar	Büyük görsel dil modeller sunucu tarafında çalışmak zorunda	Sıkıştırma sonrası anlamsal kapasite kaybı, güvenlik (model poisoning) riskleri
Otonom Sistemlere Katkı	Algı modelleri için ölçeklenebilir ve güvenilir eğitim verisi üretimi	Gerçek zamanlı, gizlilik dostu ve sahaya uyarlanabilir VLM tabanlı algı

3.7.2.2. Görsel Dil Modellerinin Robotik Algı, Anlama ve Görev Yönlendirmede Kullanımı

Robotik alanda Vision–Language–Action (VLA) modelleri, görsel algı, dil temsilleri ve eylem üretimini tek bir uçtan uca öğrenme çerçevesinde birleştirerek doğal dil ile tanımlanan görevlerin doğrudan robot kontrolüne aktarılmasını hedeflemektedir. Bu yaklaşım, klasik modüler robotik mimarilerde ayrı ele alınan algı, planlama ve kontrol bileşenlerini yüksek seviyeli semantik temsiller üzerinden bütünleştirmeyi amaçlamaktadır (Tablo 7).

Brohan ve arkadaşları (2023) tarafından geliştirilen RT-2 (Robotics Transformer-2), büyük ölçekli görsel–dil modellerinden elde edilen semantik bilginin robotik kontrol politikalarına aktarılabileceğini göstermektedir. RT-2'de robot eylemleri ayrık metin belirteçleri olarak temsil edilmekte ve kontrol problemi, doğal dil üretimine benzer otoregresif bir tahmin süreci şeklinde modellenmektedir. Altı serbestlik dereceli (6-DoF) konum, yönelim ve tutucu durumu ayrıklaştırılarak modelin çıktı kelime dağarcığına dâhil edilmekte; bu yaklaşım literatürde action-as-text tokenization olarak adlandırılmaktadır. Mimari olarak RT-2, PaLI-X (5B ve 55B parametre) ve PaLM-E (12B parametre) omurgaları üzerine inşa edilmiştir. Eğitim sürecinde robotik trajektori verileri ile web ölçekli görsel–dil görevleri birlikte kullanılarak co-fine-tuning uygulanmış; böylece robotik uzmanlaşma sağlanırken genel semantik bilginin korunması amaçlanmıştır. Deneylerde modelin masaüstü manipülasyon görevlerinde görülmemiş nesneler ve komutlar altında daha yüksek başarı oranları elde ettiği, bazı zero-shot senaryolarda başarının önceki yöntemlere göre 2–6 kat arttığı rapor edilmiştir (Brohan et al., 2023).

VLA literatürünü kapsamlı biçimde inceleyen Sapkota ve arkadaşları (2025), bu modelleri genel amaçlı robotik ajanlara doğru evrilen bir mimari paradigma olarak konumlandırmaktadır. Seksenin üzerinde VLA modeli; erken füzyon, çift sistemli ve kendi kendini düzelten mimariler olarak sınıflandırılmıştır. Çift sistemli yapılarda hızlı düşük seviyeli kontrol modülleri ile daha yavaş ancak muhakeme odaklı planlama modüllerinin ayrıştırıldığı belirtilmektedir. Ayrıca milyarlarca parametreli VLA modellerinin pratik sistemlerde uygulanabilmesi için LoRA, niceleme ve düşük dereceli adaptasyon gibi parametre-verimli öğrenme tekniklerinin kritik olduğu vurgulanmaktadır. Bununla birlikte otoregresif belirteç üretimine dayalı modellerin genellikle 3–5 Hz gibi düşük kontrol frekanslarında çalışabildiği ve bunun gerçek zamanlı, güvenli robot kontrolü açısından önemli bir sınırlama oluşturduğu ifade edilmektedir (Sapkota et al., 2025).

Gerçek dünya uygulamalarına odaklanan Kawaharazuka ve arkadaşları (2023), VLA modellerini veri toplama stratejileri, robot donanımı ve eğitim

paradigmaları bağlamında değerlendirmiştir. Çalışmada sensörimotor modeller, dünya modelleri ve sağlar (affordance) temelli yaklaşımlar karşılaştırılmış; özellikle sürekli eylem uzaylarında difüzyon tabanlı politika başlıklarının daha pürüzsüz ve kararlı kontrol sağladığı rapor edilmiştir. Ayrıca farklı robot platformları arasında öğrenilen temsillerin doğrudan aktarımında performans kaybı yaşanabildiği ve embodiment transfer probleminin hâlen açık bir araştırma alanı olduğu belirtilmiştir (Kawaharazuka et al., 2023).

Genel olarak VLA modelleri, robotik sistemlere dil temelli muhakeme, görev genellemesi ve açıklanabilirlik kazandırmaktadır. Ancak bu kazanımlar; yüksek hesaplama maliyeti, düşük kontrol frekansı, güvenlik garantileri ve donanım bağımlılığı gibi mühendislik kısıtları ile dengelenmek zorundadır (Brohan et al., 2023; Sapkota et al., 2025; Kawaharazuka et al., 2023).

Tablo 7. Görsel Dil Modellerinin Robotik Algı, Anlama ve Görev Yönlendirmede Kullanımı

Özellik	VLM-Social-Nav	NaVid	Human-Guided Mobile Robot Navigation (HSM)	Planning with Vision-Language Models
Kaynak	Song et al., 2024	Zhang et al., 2024	Wang et al., 2025	Hu et al., 2023
Çalışma Alanı	Sosyal farkındalıklı robot navigasyonu	Haritasız görsel–dil navigasyonu (VLN-CE)	İnsan odaklı akıllı üretim ortamları	Robot destekli öğretim ve görev planlama
Temel Amaç	İnsan–robot etkileşimlerini dikkate alarak güvenli ve sosyal açıdan uygun rota seçimi	Harita, derinlik veya odometri olmadan yalnızca RGB video ve dil ile navigasyon	Gürültülü ve yapılandırılmamış üretim ortamlarında insan komutlarıyla güvenilir hareket	Doğal dil talimatlarını sıralı robot eylemlerine dönüştürmek
Temel Yaklaşım	VLM ile aday navigasyon yollarını skorlayarak seçim	Video tabanlı uçtan uca VLM; bir sonraki eylemi doğrudan üretme	Modüler mimari: 3B rekonstrüksiyon + VLM algı + LLM planlama	VLM destekli görev ayrıştırma ve planlama
Kullanılan VLM / LLM	CLIP-türevi VLM + skorlayıcı yapı	LLaMA-VID tabanı, Vicuna-7B	LSeg (CLIP tabanlı) + GPT-3.5	Genel amaçlı VLM + LLM
Görsel Kodlayıcı	CLIP tabanlı encoder	EVA-CLIP	ViT-L/16 (LSeg)	CLIP / ViT tabanlı
Dil Modeli	VLM içi dil bileşeni	Vicuna-7B	GPT-3.5	GPT-türevi LLM
Mimari Özellik	Sosyal bağlam farkındalığı	Instruction-queried & instruction-agnostic tokenlar	3 aşamalı modüler yapı (rekonstrüksiyon–algı–planlama)	Dil → görev → eylem zinciri
Girdi Türü	RGB görüntü + doğal dil	Monoküler RGB video + doğal dil	RGB-D video + doğal dil	Görsel gözlem + doğal dil

Çıktı Türü	En uygun navigasyon rotası	Düşük seviyeli yürütülebilir eylemler	Python tabanlı robot komutları	Sıralı görev ve eylem planı
Veri Setleri	Sosyal navigasyon senaryoları	R2R, RxR, VLN-CE	Özel fabrika verisi + AI Habitat	Eğitim ve öğretim senaryoları
Değerlendirme Metrikleri	Başarı oranı, güvenlik ölçütleri	SR, SPL	Piksel doğruluğu, navigasyon başarı oranı	Görev tamamlama oranı
Öne Çıkan Sonuçlar	Sosyal açıdan daha güvenli navigasyon	Gerçek dünyada %66 başarı; çapraz veri setinde %41.7 SR artışı	Simülasyonda %92.5 navigasyon başarısı	Eğitim senaryolarında güvenilir planlama
Güçlü Yan	İnsan merkezli navigasyon	Sensörsüz ve haritasız genelleme	Endüstriyel ortamlara uygunluk	Dil tabanlı esnek görev planlama
Sınırlılıklar	Hesaplama maliyeti	Gerçek zamanlılık sınırlı	Sistem karmaşıklığı	Fiziksel algıdan ziyade planlama odaklı

3.7.2.3. Görsel Dil Modelleri ile Algı, Grounding, Affordance ve Düşük-Seviyeli Robotik Biliş

Robotik sistemlerde algıdan eyleme uzanan bilişsel süreçler, yalnızca nesne tanıma veya sahne sınıflandırma çıktılarıyla sınırlı değildir. Robotların çevreyle güvenli ve etkili biçimde etkileşime girebilmesi; mekânsal temellendirme (grounding), eylem uygunluğu (affordance), çevresel durum algısı ve düşük-seviyeli fiziksel özelliklerin doğru biçimde yorumlanmasını gerektirmektedir. Bu bağlamda görsel dil modelleri, görsel gözlemleri dilsel ve kavramsal temsillerle ilişkilendirerek, klasik denetimli algı yaklaşımlarının ötesinde daha esnek ve genellenebilir robotik biliş çerçeveleri sunmaktadır (Tablo 8).

Düşük-seviyeli robotik algıya odaklanan Osada ve arkadaşları (2024), yakınlık sensörleri açısından kritik olan kızılötesi yansıtma (reflectance) değerinin, yalnızca fiziksel ölçümle değil dağılımsal semantik bilgi üzerinden de tahmin edilebileceğini göstermiştir. Metin tabanlı tahmin için GPT-3.5 ve GPT-4, görüntü tabanlı tahmin için CLIP temelli bir VLM kullanılmıştır. Sonuçlar, GPT-4'ün yalnızca nesne adından %14.7 ortalama hata oranı ile görüntü tabanlı ResNet modellerini geride bıraktığını; CLIP tabanlı yaklaşımın ise %11.8 hata oranı ile en düşük hatayı elde ettiğini göstermektedir. Bu bulgular, dil modellerinde gömülü semantik bilginin düşük-seviyeli robotik görevlerde kullanılabileceğini ortaya koymaktadır.

Algısal çıktının doğrudan eyleme dönüştürülmesini amaçlayan bir diğer yaklaşım olan RoboPoint (Zhu ve ark., 2024), robotik manipülasyon ve navigasyon görevlerinde gerekli eylem noktalarını, sınırlayıcı kutular yerine

doğrudan görüntü düzleminde (x, y) koordinatları olarak tahmin etmeyi hedeflemektedir. Model mimarisi, Vicuna-v1.5-13B tabanlı bir dil modeli ile CLIP ön-eğitimli görsel kodlayıcı üzerine inşa edilmiş olup, komut ince ayarı (instruction fine-tuning) ve ölçeklenebilir sentetik veriyle eğitilmiştir. Sonuçlar, RoboPoint modelinin gerçek dünya robot görevlerindeki başarı oranını %30'un üzerinde artırdığını ve diğer VLM tabanlı yaklaşımlara kıyasla daha yüksek doğruluk sağladığını göstermektedir.

Çevresel durum farkındalığına yönelik Sharshar ve arkadaşları (2023), ön-eğitimli VLM'leri kara kutu optimizasyon yöntemleriyle birleştirmiştir. Kapı durumu, musluk akışı veya ortam temizliği gibi durumlar VQA ve görüntü–metin eşleştirme görevleri üzerinden tanınmıştır. BLIP-2, OFA, CLIP ve ImageBind gibi modeller için oluşturulan çoklu metin istemleri, genetik algoritma tabanlı optimizasyonla ayarlanmıştır. Optimize edilmiş istemlerle tek bir VLM kullanılarak farklı çevresel durumların %90'ın üzerinde doğrulukla tanınabildiği ve özellikle şeffaf yüzeyler gibi zorlayıcı senaryolarda performansın arttığı rapor edilmiştir.

Mekânsal temellendirmeyi üç boyuta taşıyan VLM-Grounder yaklaşımı (Xu et al., 2024), sıfır atış 3B görsel temellendirme problemine odaklanmaktadır. Sistem, 3B nokta bulutuna ihtiyaç duymadan yalnızca 2B görüntüler ve doğal dil sorguları üzerinden hedef nesnenin 3B sınırlayıcı kutusunu tahmin etmektedir. GPT-4o ajan olarak kullanılmış; Grounding DINO, SAM ve YOLOv8-World gibi 2B algı bileşenleri dinamik birleştirme, geri bildirimli temellendirme ve çok görünümlü projeksiyon mekanizmalarıyla entegre edilmiştir. ScanRefer ve Nr3D veri setlerinde önerilen yöntemin önceki zero-shot yaklaşımlara kıyasla anlamlı doğruluk artışı sağladığı ve bazı gözetimli yöntemlerle karşılaştırılabilir performans sunduğu bildirilmiştir.

Bu çalışmalar, görsel dil modellerinin robotikte yalnızca yüksek-seviyeli planlama değil; düşük-seviyeli fiziksel özellik tahmini, eylem noktası çıkarımı ve çok boyutlu mekânsal temellendirme gibi görevlerde de etkili olduğunu göstermektedir. Ortaya çıkan eğilim, robotik bilişin tekil algı modülleri yerine görsel, dilsel ve fiziksel bilgiyi ortak bir temsil uzayında bütünleştiren çok modlu sistemler üzerinden ele alınması yönündedir.

Tablo 8. Görsel–Dil Modellerinin Robotik Algı, Mekânsal Temellendirme ve Eylem Uygunluğu Görevlerindeki Kullanımı

Özellik	Yansıtma Tahmini ile Yakınlık Algılama	RoboPoint: Mekânsal Eylem Noktası Tahmini	Çevresel Durum Tanıma (VLM + Optimizasyon)	VLM-Grounder: Sıfır Atış 3B Temellendirme
Kaynak	Osada et al., 2024	Zhu et al., 2024	Sharshar et al., 2023	Xu et al., 2024
Temel Amaç	Nesnelerin IR yansıtma değerlerini tahmin ederek yakınlık algısını iyileştirmek	Robot eylemleri için doğrudan (x, y) eylem noktaları tahmin etmek	Çevresel durumları doğal dil üzerinden tanımak	2B görüntülerden 3B hedef nesne temellendirmek
Robotik Problem Türü	Düşük-seviyeli algı ve kavrama hassasiyeti	Mekânsal uygunluk ve hassas manipülasyon	Durum farkındalığı	3B mekânsal temellendirme
Kullanılan Modeller	GPT-3.5, GPT-4, CLIP; VGG16, ResNet101 (kıyas)	Vicuna-13B + CLIP	BLIP-2, OFA, CLIP, ImageBind	GPT-4o, Grounding DINO, SAM
Temel Teknikler	Few-shot prompting, CoT, CLIP embedding regresyonu	Instruction fine-tuning, sentetik veri üretimi	VQA, ITR, genetik algoritma ile istem optimizasyonu	Dinamik stitching, geri bildirim, çok görünümlü projeksiyon
Veri Setleri	54 nesne, 324 görüntü-yansıtma çifti	RoboRefIt, WHERE2PLACE	Özel çevresel durum veri setleri	ScanRefer, Nr3D
Öne Çıkan Sonuçlar	CLIP %11.8 hata, GPT-4 %14.7 hata	%30+ başarı artışı	%90+ doğruluk	Acc@0.25'te %51.6

3.7.3. Görsel Dil Modellerinin Sağlık Alanındaki Kullanımı

3.7.3.1. Radyoloji ve Klinik Görüntülemede Görsel Dil Modelleri

Radyoloji alanında Görsel Dil Modellerinin (VLM) kullanımı, klasik serbest metin rapor üretiminin ötesine geçerek; klinik bilginin yapılandırılması, halüsinasyonların azaltılması ve karar destek süreçlerine entegrasyonu hedefleyen bütüncül sistemlere evrilmektedir. Son dönemdeki çalışmaların temel amacı, radyoloji görüntüleri ve raporlarının birlikte ele alındığı çok modlu çerçeveler aracılığıyla metin tabanlı yaklaşımların bağlam sınırlamalarını ve klinik tutarsızlık sorunlarını aşmaktır (Zhang et al., 2024; Li et al., 2024; Zhong et al., 2025; Wang et al., 2023).

Bu doğrultuda öne çıkan VLM-KG, radyoloji görüntüleri ile serbest metin raporlarını birlikte kullanarak yapılandırılmış bilgi grafikleri üretmeyi hedeflemektedir (Zhang et al., 2024). Önerilen mimaride, uzun raporları işleyebilmek için 32K bağlam uzunluğunu destekleyen Qwen1.5-0.5B dil modeli ve görsel tarafta kontrastif biçimde eğitilmiş MedCLIP kodlayıcısı kullanılmıştır. Görsel ve dilsel temsiller arasındaki boyut uyumsuzluğu çok

katmanlı transformatör tabanlı bir projektörle giderilmiş ve RadGraph şemasına dayalı varlık–ilişki–varlık üçlüleri üzerinden denetimli öğrenme uygulanmıştır. Deneysel sonuçlar, görsel bağlamın entegrasyonuyla bilgi grafiği üretimindeki halüsinasyonların anlamlı biçimde azaldığını ve performans metriklerinde belirgin iyileşmeler sağlandığını göstermektedir.

Radyoloji rapor üretiminde güvenilirliği artırmayı amaçlayan SERPENT-VLM (Li et al., 2024), görüntüde bulunmayan bulguların rapora eklenmesi sorununu çözmek için kendi kendini iyileştiren (self-refining) bir eğitim stratejisi hedeflemektedir. Sistemin mimarisinde, dondurulmuş Swin-Transformer V2 görsel kodlayıcısı ile LLaMA2-7B dil modeli entegre edilmiş; standart dil modeli kaybına ek olarak, üretilen rapor ile görsel temsil arasındaki tutarlılığı zorlayan bir "self-refining loss" tanımlanmıştır. IU-Xray ve ROCO veri setleri üzerindeki sonuçlar, bu yaklaşımın rapor doğruluğunu artırırken halüsinasyonları sistematik olarak azalttığını kanıtlamaktadır.

Görsel dil modellerinin klinik iş akışına derin entegrasyonunu hedefleyen ve raporlamayı tanı/prognoz tahminiyle birleştirmeyi amaçlayan bir diğer sistem, CT Pulmonary Angiogram (CTPA) odaklı ajan tabanlı çerçevedir (Zhong et al., 2025). Bu kapsamlı mimaride, pulmoner emboli ile ilişkili 32 anormalliği tespit eden çok etiketli bir sınıflandırıcının ardından; CT-CHAT, RadFM ve M3D gibi modeller "okuma ajanı" olarak konumlandırılmıştır. Bölgesel bulgular LLaMA 3 tabanlı bir yazma ajanı tarafından yapılandırılmış raporlara dönüştürülmekte ve son aşamada görüntü, klinik değişkenler ve rapor çıktıları Cox Proportional Hazards modeli kullanılarak sağkalım tahmini yapılmaktadır. Çok merkezli veri setlerinden elde edilen sonuçlar, bu çok ajanlı yaklaşımın hem rapor kalitesini hem de klinik sonuç tahminini anlamlı ölçüde iyileştirdiğini ortaya koymaktadır.

Önceden eğitilmiş bir VLM mimarisinin klinik alana çok aşamalı ince ayar stratejileriyle uyarlanmasını amaçlayan ClinicalBLIP modelinde ise (Wang et al., 2023); InstructBLIP tabanında CLIP görsel kodlayıcısı, Flan-T5-XL dil modeli ve Q-Former yapısı bir araya getirilmiş, CheXBERT ile çıkarılan tıbbi etiketler modele istem (prompt) olarak entegre edilmiştir. IU-Xray ve MIMIC-CXR veri setlerindeki sonuçlar, klinik bağlama uyarlanan bu çok modlu modellerin rapor doğruluğu ve tutarlılığında önemli kazanımlar sağladığını göstermektedir. Genel bir değerlendirmeyle; bu çalışmalar görsel dil modellerinin radyolojide salt rapor üretimi sınırlarını aşarak yapılandırılmış bilgi çıkarımı ve klinik karar destek mekanizmaları için vazgeçilmez hâle geldiğini doğrulamaktadır. Tablo 9'da radyoloji ve klinik görüntülemede VLM tabanlı yaklaşımların karşılaştırılması verilmiştir.

Tablo 9. Radyoloji ve Klinik Görüntülemede VLM Tabanlı Yaklaşımların Karşılaştırılması

Özellikler	VLM-KG	SERPENT-VLM	ClinicalBLIP	CTPA-VLM
Kaynak	Multimodal Radiology Knowledge Graph Generation	Self-Refining Radiology Report Generation Using VLM	Vision-Language Model for Generating Textual Descriptions From Clinical Images	VLM for Report Generation and Outcome Prediction in CT Pulmonary Angiogram
Çalışma Alanı	Radyoloji bilgi çıkarımı ve klinik bilgi yapılandırma	Güvenilir radyoloji rapor üretimi	Klinik görüntüden rapor üretimi	Toraks BT (CTPA) tabanlı tanı ve prognoz
Temel Amaç	Görüntü + rapor kullanarak bilgi grafiği üretmek ve halüsinasyonları azaltmak	Görüntüyle uyumsuz rapor üretimini azaltmak	Klinik bağlama uyumlu rapor üretimini güçlendirmek	Tanı, yapılandırılmış rapor ve klinik sonuç tahminini entegre etmek
Temel Yaklaşım	Görsel dil temsillerini bilgi grafiği triplet'lerine dönüştürme	Self-refining kayıp ile görsel–metin hizalaması	Çok aşamalı klinik ince ayar ve önbilgi entegrasyonu	Anormallik-rehberli ajan tabanlı raporlama
Kullanılan VLM / LLM	Qwen1.5-0.5B	LLaMA2-7B	Flan-T5-XL	LLaMA-3
Görsel Kodlayıcı	MedCLIP	Swin Transformer V2 (frozen)	CLIP	3D I3D
Dil Modeli	Qwen1.5-0.5B	LLaMA2-7B	Flan-T5-XL	LLaMA-3
Mimari Özellik	Transformer tabanlı projektör + visual instruction tuning	LoRA destekli eşleme + Self-Refining Loss	Q-Former + InstructBLIP mimarisi	Reading Agent + Writing Agent + CoxPH
Girdi Türü	Radyoloji görüntüsü + serbest metin rapor	Radyoloji görüntüsü	Klinik görüntü	CTPA hacimleri + klinik değişkenler
Çıktı Türü	Bilgi grafiği triplet'leri	Serbest metin radyoloji raporu	Klinik rapor / açıklama	Tanı, yapılandırılmış rapor, prognoz
Veri Setleri	MIMIC-CXR, IU-Xray	IU-Xray, ROCO	IU-Xray, MIMIC-CXR	BUH, JHU, INSPECT
Değerlendirme Metrikleri	BLEU-1/2/3/4, ROUGE-L	BLEU-4, ROUGE-L, BERTScore	BLEU-A, METEOR, ROUGE-L	AUROC, F1, BERT-F1, BLEU-4
Öne Çıkan Sonuçlar	Daha düşük halüsinasyon ve daha yüksek yapısal doğruluk	Halüsinasyonların belirgin biçimde azalması	Klinik tutarlılığı yüksek raporlar	Tanı ve prognoz başarısında anlamlı artış
Güçlü Yan	Görsel bağlamla güvenilir bilgi çıkarımı	Görsel–metin uyumunun doğrudan optimize edilmesi	Klinik bağlama güçlü adaptasyon	Uçtan uca klinik iş akışına entegrasyon
Sınırlılıklar	Bilgi grafiği şemasına bağımlılık	Hesaplama maliyeti ve eğitim karmaşıklığı	Gerçek zamanlı kullanım için ağır mimari	Yüksek hesaplama ve veri gereksinimi

3.7.3.2. Uzmanlık Alanına Özgü Klinik Görsel Dil Model Uygulamaları: VQA, Tanı Destekli Yorumlama ve Görev Odaklı Modeller

Uzmanlık alanına özgü klinik görsel dil modeli uygulamaları, genel amaçlı rapor üretiminden farklı olarak belirli klinik görevlerin doğrudan desteklenmesine odaklanmaktadır. Bu yaklaşımlar, özellikle uzman eksikliğinin belirgin olduğu alanlarda tanı destek süreçlerinin güvenilirliğini ve erişilebilirliğini artırmayı amaçlamaktadır.

Dermatoloji alanında çok dilli Görsel Soru-Cevaplama (VQA) görevine odaklanan ve IKIM ekibi tarafından MEDIQA-M3G 2024 yarışmasında sunulan çalışma, bu vizyonun başarılı bir örneğidir (Liu et al., 2024). Çalışmanın amacı, dermatolog eksikliği olan senaryolarda VLM tabanlı sistemlerin tanı destek potansiyelini değerlendirmek ve özellikle Çince veriler üzerinde en ideal modelleme stratejisini belirlemektir. Önerilen sistem mimarisinde, biyomedikal alan için özelleştirilmiş LLaVA-med modeli temel alınmış ve yalnızca Çince verilerle minimal düzeyde bir ince ayar (fine-tuning) sürecinden geçirilmiştir. Modelden alınan Çince çıktılar ise Mixtral-8×7B-instruct büyük dil modeli kullanılarak İngilizce ve İspanyolcaya aktarılmıştır. Eğitim mimarisine dair en dikkat çekici strateji; modelin yalnızca tek bir epoch (epok) boyunca (öğrenme oranı = $2e^{-5}$, yığın boyutu = 4, gradyan birikimi = 16) eğitilmesi ve Fitzpatrick17k ile DermNet gibi harici veri setlerinin aşırı uyum (overfitting) riski nedeniyle eğitim dışı bırakılmasıdır. Deneysel sonuçlar, her dil için ayrı bir model eğitmek yerine, tek dilde optimize edilmiş bir modelin çıktılarını LLM tabanlı çeviri ile çoğaltmanın çok daha başarılı olduğunu kanıtlamıştır. Bu stratejinin bir sonucu olarak IKIM ekibi, yarışmanın Çince ve İspanyolca kategorilerinde birincilik, İngilizce kategorisinde ise üçüncülük elde etmiştir (Liu et al., 2024).

Benzer şekilde görev odaklı bir VLM yaklaşımıyla önkol ultrason görüntülerinden el hareketi çözümleme problemini ele alan bir diğer çalışma, klasik yoğun etiketli sınıflandırma yöntemlerini Vision-Language çerçevesiyle değiştirmeyi amaçlamaktadır (Bimbraw et al., 2024). Kurgulanan mimaride, görsel dil modellerinin bağlam içi öğrenme (in-context learning - ICL) yetenekleri ve dikkatle tasarlanmış istem (prompt) yapıları kullanılarak; tam bir ince ayara gerek kalmadan, yalnızca kullanıcıya özgü birkaç örnekle (few-shot) sistemin uyarlanabilmesi hedeflenmiştir. Özellikle klinik açıdan gerçeğe en yakın olan oturumlar arası (cross-session) ve kullanıcılar arası (cross-subject) senaryolarda gerçekleştirilen deneylerin sonuçları, few-shot örnek sayısı arttıkça modelin genelleme performansının belirgin biçimde yükseldiğini göstermektedir. Bu bulgular, ultrason gibi operatöre ve bireysel anatomiye son derece duyarlı olan, veri kıtlığının yaşandığı durumlarda dahi

VLM tabanlı mimarilerin rekabetçi doğruluk oranlarına ulaşabildiğini ortaya koymaktadır (Bimbraw et al., 2024).

Genel bir çerçeveden değerlendirildiğinde, uzmanlık alanına özgü klinik VLM uygulamalarının başarısının karmaşık çoklu model mimarilerinden ziyade; doğru görev tanımlamasına, hedef odaklı ve sınırlı ince ayar stratejilerine, ayrıca model yetenekleriyle örtüşen doğru veri kullanımına bağlı olduğu sonucuna varılmaktadır (Bimbraw et al., 2024; Liu et al., 2024). Uzmanlık alanına özgü klinik VLM uygulamalarının karşılaştırılması Tablo 10' da verilmiştir.

Tablo 10. Uzmanlık Alanına Özgü Klinik Görsel Dil Model(VLM) Uygulamalarının Karşılaştırılması

Özellik	Multilingual Visual Question Answering for Dermatology	Hand Gesture Decoding from Forearm Ultrasound Images
Kaynak	Liu et al. (2024)	Bimbraw et al. (2024)
Çalışma Alanı	Dermatoloji	Rehabilitasyon, protez ve insan–makine etkileşimi
Temel Amaç	Dermatoloji görüntülerinden klinik sorulara çok dilli ve doğru yanıt üretmek	Önkol ultrason görüntülerinden el hareketlerini veri kıtlığı koşullarında çözümlemek
Temel Yaklaşım	VQA tabanlı klinik tanı destek sistemi + LLM tabanlı çeviri	Vision–Language çerçevesinde few-shot jest sınıflandırma
Kullanılan VLM / LLM	LLaVA-med + Mixtral-8×7B-instruct	GPT-Sonography (VLM tabanlı yaklaşım)
Görsel Kodlayıcı	LLaVA-med görsel omurgası	Ultrason görüntülerine uyarlanmış VLM görsel kodlayıcısı
Dil Modeli	LLaVA-med (temel) + Mixtral (çeviri)	VLM'nin entegre dil bileşeni
Mimari Özellik	Tek dilde fine-tuning + çok dilli çıktı üretimi	Fine-tuning olmadan in-context learning (ICL)
Girdi Türü	Dermatoloji görüntüsü + klinik soru	Ultrason görüntüsü + görev tanımlı prompt
Çıktı Türü	Klinik VQA yanıtı (çok dilli)	El jesti sınıf etiketi
Veri Setleri	MEDIQA-M3G (Çince, İngilizce, İspanyolca); Fitzpatrick17k, DermNet (deneysel)	Özel toplanmış ultrason jest veri seti
Değerlendirme Senaryosu	MEDIQA-M3G yarışması	Cross-session ve cross-subject testler
Değerlendirme Metrikleri	Yarışma sıralaması, doğruluk	Sınıflandırma doğruluğu
Öne Çıkan Sonuçlar	Çince & İspanyolca: 1.'lik; İngilizce: 3.'lük	Few-shot örnek sayısı arttıkça doğruluk belirgin artıyor
Güçlü Yan	Çok dilli klinik tanı desteği, düşük eğitim maliyeti	Hızlı kişiselleştirme, veri kıtlığına dayanıklılık
Sınırlılıklar	Harici veri ve çoklu görüntü girdisi performansı düşürüyor	Jest çeşitliliği arttıkça performans sınırlanabiliyor

3.7.3.3. Temel ve Geniş Ölçekli Medikal VLM Yaklaşımları: Foundation Modeller, Etiket Kıtlığı ve Büyük Veri ile Öğrenme

Medikal Görsel Dil Modelleri (VLM) literatüründe son yılların temel eğilimi, etiketleme maliyetlerini minimize etmeyi ve sağlık sistemi ölçeğinde genellenebilirliği hedefleyen "foundation" (temel) model yaklaşımlarıdır. Patoloji, çok modlu klinik görevler ve nörogörüntüleme alanlarına odaklanan bu çalışmalar; etiket kıtlığı, görevler arası çatışma ve büyük ölçekli klinik verilerin entegrasyonu gibi temel sorunlara yenilikçi teknik çözümler sunmayı amaçlamaktadır.

Etiketleme sürecini tamamen ortadan kaldırmayı hedefleyen VLM-CPL yöntemi, patoloji görüntülerinin sınıflandırılması problemine odaklanmaktadır (Zhang et al., 2023). Önerilen mimaride, önceden eğitilmiş VLM'lerin sıfır-atış (zero-shot) çıkarım yetenekleri kullanılarak otomatik sözde etiketler (pseudo-labels) üretilir. Eğitim ve hedef veri dağılımları arasındaki uyumsuzluktan (domain gap) kaynaklanan etiket gürültüsünü filtrelemek amacıyla, PLIP ve BioMedCLIP modelleriyle üretilen etiketler çok aşamalı bir yarı denetimli öğrenme çerçevesinden geçirilir ve nihai sınıflandırma ResNet50 ile UNI kodlayıcıları üzerinden yapılır. Bu mimarinin merkezinde yer alan Çoklu Görünüm Uzlaşması (Multi-View Consensus - MVC) mekanizması, farklı veri artırma görünümleri arasındaki tahmin tutarlılığını ölçerek güvenilir örnekleri seçer. Sınıf dengesizliğini hafifletmek için Sınıfa Duyarlı MVC (CMVC), etiket güvenilirliğini artırmak için ise Prompt-Feature Consensus (PFC) ve High-confidence Cross Supervision (HCS) modülleri sisteme entegre edilmiştir. Tüm slayt görüntüleri (WSI) düzeyinde ise ilgisiz yamaları eleyen Open-Set Prompting (OSP) mekanizması kullanılmıştır. Elde edilen deneysel sonuçlar, mekânsal ve renk dönüşümlerinin birlikte kullanımının MVC başarısını katladığını (örneğin HPH veri setinde doğruluğun 0.645'ten 0.904'e çıktığını) ve VLM-CPL yaklaşımının doğrudan sıfır-atış kullanımına kıyasla ortalama %18.8 doğruluk artışı sağladığını kanıtlamaktadır (Zhang et al., 2023).

Tıbbi görsel anlama (comprehension) ve görsel üretim (generation) görevlerini tek bir model çatısında birleştirmeyi amaçlayan bir diğer foundation yaklaşım HealthGPT'dir (Li et al., 2024). Bu çalışma, anlama görevlerinin gerektirdiği soyut/semantik temsiller ile üretim görevlerinin ihtiyaç duyduğu ayrıntılı/yüksek frekanslı görsel bilgi arasındaki görev çatışmasını çözmeyi hedeflemektedir. Geliştirilen Heterojen Düşük Dereceli Adaptasyon (Heterogeneous Low-Rank Adaptation - H-LoRA) mimarisi, görevlere özgü bilgileri izole edilmiş LoRA eklentilerinde saklayarak dinamik yönlendirme ile görevler arası negatif etkileşimi sınırlar. Sistem mimarisinde

görsel kodlayıcı olarak CLIP-L/14 kullanılmış (derin katmanlar anlama, sığ katmanlar üretim için); dil modeli tarafında Phi-3-mini ve Phi-4 tercih edilmiş ve VQGAN tabanlı ayrık jetonlaştırma uygulanmıştır. Üç Aşamalı Eğitim Stratejisi (TLS) ile optimize edilen modelin deneysel sonuçları, VL-Health veri setinde OmniMedVQA testi için 68.5 puan ve CT-MRI dönüşüm görevinde 79.38 SSIM skoru elde edildiğini göstermektedir. Ayrıca H-LoRA yapısının, MoELoRA yaklaşımına kıyasla eğitim süresini yaklaşık %33 oranında azalttığı rapor edilmiştir (Li et al., 2024).

Sağlık sistemi ölçeğinde nörogörüntüleme alanına odaklanan Prima modeli ise, gerçek dünya klinik MRI verilerini tüm sekanslarıyla işleyerek genellenebilir bir yapay zeka temeli oluşturmayı amaçlamaktadır (Peng et al., 2024). Prima mimarisi, MRI verisinin hiyerarşik doğasına tam uygun olarak tasarlanmış; hacim düzeyinde sıkıştırma için 3D VQ-VAE tabanlı bir volume tokenizer, ardından sekans ve çalışma düzeyinde temsil üretimi için iki aşamalı bir Vision Transformer (ViT) yapısı kurgulanmıştır. Eğitim sürecinde, MRI görüntüleri ile GPT-3.5 tarafından özetlenen radyoloji raporları (önyargıyı azaltmak amacıyla) CLIP hedefi üzerinden hizalanmıştır. UM-220K veri setiyle (220.000'den fazla MRI çalışması) eğitilen modelin prospektif sonuçları, 52 farklı radyolojik tanıda ortalama %92.0 AUROC değerine ulaştığını ve sıfır-atış senaryolarında dahi CLIP, LLaVA, BioMedCLIP ve Med-Flamingo gibi modelleri geride bıraktığını ortaya koymaktadır. Klinik uygulama açısından modelin, radyolog iş listesi önceliklendirmesinde gerçek şiddet skorlarıyla yüksek korelasyon (Pearson r = 0.69) ve sevk önerisi görevlerinde %85 üzeri AUROC sağladığı kanıtlanmıştır (Peng et al., 2024). Tablo 11' de temel ve geniş ölçekli medikal VLM yaklaşımlarının karşılaştırılması sunulmuştur.

Tablo 11. Temel ve Geniş Ölçekli Medikal VLM Yaklaşımlarının Karşılaştırılması

Özellikler	VLM-CPL	HealthGPT	Prima
Kaynak	Zhang et al. (2023)	Li et al. (2024)	Peng et al. (2024)
Çalışma alanı	Dijital patoloji	Çok modlu klinik görsel zekâ	Nörogörüntüleme (MRI)
Temel amaç	İnsan etiketlemesini ortadan kaldırarak etiketsiz patoloji verilerinden yüksek doğruluklu sınıflandırma yapmak	Tıbbi görsel anlama ve üretim görevlerini tek bir modelde görev çatışması olmadan birleştirmek	Sağlık sistemi ölçeğinde MRI verileriyle genellenebilir bir foundation model geliştirmek
Temel yaklaşım	Zero-shot VLM tabanlı pseudo-label üretimi + yarı denetimli öğrenme	Çok görevli birleşik öğrenme ve görev-özgü adaptasyon	Büyük ölçekli contrastive vision–language öğrenme

Kullanılan VLM / LLM	PLIP, BioMedCLIP (pseudo-label üretimi)	HealthGPT; LLM: Phi-3-mini (3.8B), Phi-4 (14B)	Prima (CLIP hedefli vision–language model)
Görsel kodlayıcı	PLIP / BioMedCLIP image encoder'ları	CLIP-L/14 (katman bazlı ayrıştırma)	Hiyerarşik Vision Transformer (Volume → Sequence → Study)
Dil modeli	prompt tabanlı VLM çıkarımı	Phi-3-mini, Phi-4	— (rapor–görüntü hizalaması için CLIP hedefi)
Mimari özellikler	MVC, CMVC, PFC, HCS, Open-Set Prompting (OSP)	H-LoRA, Hierarchical Visual Perception (HVP), Three-Stage Learning Strategy	3D VQ-VAE volume tokenizer, iki aşamalı ViT
Girdi türü	Patoloji yama görüntüleri (patch), WSI	Klinik görüntüler + doğal dil	Çok sekanslı MRI hacimleri + radyoloji raporları
Çıktı türü	Sınıf etiketleri (patoloji)	Metinsel cevaplar, üretilmiş tıbbi görüntüler	Tanı skorları, önceliklendirme puanları
Veri setleri	Patch: HPH, LC25K, NCT-CRC-HE-100K; WSI: DigestPath, TCGA-RCC	VL-Health: PubMedVision, LLaVA-Med, PathVQA, MIMIC-CXR-VQA, IXI, SynthRAD2023	UM-220K: 220K+ MRI çalışma, 5.6M sekans, 362M kesit
Değerlendirme metrikleri	Accuracy	VQA score, SSIM	AUROC, Pearson korelasyonu
Öne çıkan sonuçlar	Ortalama %18.8 doğruluk artışı; HPH'de 0.645 → 0.883; MVC + ColorJitter ile 0.904	OmniMedVQA'da 68.5; CT→MRI dönüşümünde 79.38 SSIM; MoELoRA'ya kıyasla %33 daha hızlı eğitim	52 tanıda %92.0 AUROC; sevk önerilerinde %85+ AUROC; iş listesi önceliklendirmede r = 0.69
Güçlü yan	Gürültüye dayanıklı pseudo-label temizleme ve etiketsiz öğrenme	Görev çatışmasını mimari düzeyde azaltan birleşik yapı	Gerçek dünya klinik verilerinde yüksek genellenebilirlik
Sınırlılıklar	Domain shift'e duyarlılık ve patch–WSI geçişinde karmaşıklık	Yüksek hesaplama maliyeti ve karmaşık eğitim süreci	Büyük bellek gereksinimi ve MRI odaklı sınırlı modalite

3.7.4. Görsel -Dil Modellerinin Tarım Alanındaki Kullanımı

3.7.4.1. Uzaktan Algılama ve Mekânsal-Zamansal Tahmin ile Tarımsal Planlama

Tarımsal üretimin planlanması ve mahsul veriminin güvenilir biçimde tahmin edilmesi, yüksek boyutlu uzaktan algılama verilerinin çevresel ve iklimsel bağlamla bütünleştirilmesini gerektirmektedir. Bu doğrultuda Görsel Dil Modelleri (VLM), uydu görüntülerini meteorolojik zaman serileri ve kavramsal temsillerle harmanlayarak tarımsal karar destek süreçlerini optimize etmeyi amaçlamaktadır.

Bu bağlamda geliştirilen MMST-ViT modeli, iklim değişikliği etkilerini dikkate alan çok modlu ve mekânsal-zamansal bir mahsul verim tahmin çerçevesi sunmayı hedeflemektedir (Zhou et al., 2023). Sistemin mimarisi üç temel bileşenden oluşmaktadır. İlk bileşen olan çok modlu transformatör; Sentinel-2 uydu görüntülerindeki görsel girdileri (Pyramid Vision Transformer - PVT omurgasıyla) ve kısa vadeli meteorolojik değişkenleri, Çok Modlu Çoklu Başlık Dikkati (MM-MHA) mekanizması aracılığıyla ortak bir temsilde birleştirir. İkinci bileşen olan uzaysal transformatör, ilçe bazlı tarım alanları arasındaki mekânsal bağımlılıkları modellerken; üçüncü bileşen olan zamansal transformatör, uzun vadeli tarihsel hava verileri üzerinden iklim değişikliğinin mahsul üzerindeki kümülatif etkilerini yakalar (Zhou et al., 2023). Modelin eğitiminde ABD'deki 200'den fazla ilçeyi kapsayan 2017–2022 dönemine ait uydu, HRRR ve USDA verileri kullanılmıştır.

Deneysel sonuçlar, MMST-ViT modelinin soya fasulyesi verim tahmininde RMSE = 3.9, R^2 = 0.843 ve Pearson korelasyonu = 0.918 değerlerine ulaşarak ConvLSTM, CNN-RNN ve GNN-RNN gibi önceki uzaysal-zamansal modelleri belirgin şekilde geride bıraktığını göstermektedir. Pamuk mahsulünde (birim farklılıklarından kaynaklanan yüksek varyans nedeniyle) daha yüksek RMSE değerleri gözlense de, yüksek R^2 ve korelasyon oranları modelin güçlü açıklayıcılığını koruduğunu kanıtlamaktadır (Zhou et al., 2023).

Uzaktan algılama verilerinin tarıma özgü uyarlanmasını amaçlayan bir diğer çalışma olan AgriCLIP, genel amaçlı CLIP yapılarının tarımsal ince ayrıntıları yakalamadaki yetersizliğini aşmayı hedeflemektedir (Wang et al., 2024). AgriCLIP mimarisinde, kontrastif öğrenme yoluyla elde edilen küresel anlamsal temsiller ile DINO tabanlı kendi kendine denetimli görsel kodlayıcıdan alınan yerel ayrıntıya duyarlı özellikler, öğrenilebilir bir dönüşüm üzerinden (Ortalama Kare Hatası - MSE minimize edilerek) birbirine hizalanır. Modelin eğitimi, bitkisel üretim ve hayvancılık alanlarını kapsayan ve GPT-4

destekli metinlerle oluşturulmuş yaklaşık 600.000 görüntü-metin çiftinden oluşan ALive veri seti üzerinde gerçekleştirilmiştir.

Modelin daha önce hiç görülmemiş veriler üzerindeki sıfır-atış (zero-shot) sınıflandırma sonuçları, standart CLIP modeline kıyasla mutlak %9.07 oranında bir doğruluk artışı sağlandığını ve bu kazanımın özellikle ince taneli ayrımlar gerektiren görevlerde çok daha belirginleştiğini ortaya koymaktadır (Wang et al., 2024).

Uzaktan algılama alanındaki VLM literatürünü inceleyen kapsamlı derlemeler (Xie et al., 2024), bu modellerin güçlü görsel temsiller sunduğunu, ancak karmaşık akıl yürütme ve alana özgü bilgi entegrasyonu konularında hâlen geliştirilmeye açık olduğunu vurgulamaktadır. Genel bir değerlendirmeyle; veri heterojenliği ve iklimsel belirsizlikler gibi zorluklara rağmen, VLM tabanlı mimariler makro ölçekli tarımsal tahmin ve planlama problemlerinde kritik bir potansiyel sunmaktadır. Uzaktan algılama ve mekânsal-zamansal tahmin ile tarımsal planlama alanındaki çalışmaların karşılaştırılması Tablo 12' de verilmiştir.

Tablo 12. Uzaktan Algılama ve Mekânsal-Zamansal Tahmin ile Tarımsal Planlama

Özellikler	**AgriCLIP (Wang et al., 2024)**	**MMST-ViT (Zhou et al., 2023)**	**Vision–Language Modeling Meets Remote Sensing (Xie et al., 2024)**
Kaynak / Makale	*Adapting CLIP for Agriculture and Livestock via Domain-Specialized Cross-Model Alignment*	*Climate Change-Aware Crop Yield Prediction via Multi-Modal Spatial–Temporal Vision Transformer*	*Vision–Language Modeling Meets Remote Sensing: Models, Datasets and Perspectives*
Çalışma Alanı	Tarım ve hayvancılık	Uzaktan algılama tabanlı mahsul verim tahmini	Uzaktan algılama (genel inceleme)
Temel Amaç	Tarım ve hayvancılık alanına özgü ince taneli görsel ayrımları zero-shot olarak iyileştirmek	İklim değişikliği etkilerini dikkate alarak mahsul verimini mekânsal-zamansal bağlamda tahmin etmek	Uzaktan algılama alanında VLM yaklaşımlarını sistematik biçimde sınıflandırmak
Temel Yaklaşım	Alan-özgü CLIP adaptasyonu ve cross-model alignment	Çok modlu mekânsal-zamansal Vision Transformer mimarisi	Kontrastif, instruction-based ve generation-based VLM yaklaşımlarının karşılaştırmalı analizi
Kullanılan VLM / LLM	CLIP / OpenCLIP (kontrastif); DINO (self-supervised)	VLM tabanlı çok modlu Transformer	CLIP türevleri, instruction-tuned görsel dil modeller

Görsel Kodlayıcı	CLIP image encoder + DINO	PVT-Tiny (PVT-T/4)	Çeşitli (CLIP, RemoteCLIP vb.)
Dil Modeli	CLIP text encoder (GPT-4 yalnızca prompt üretimi için)	—	LLM tabanlı instruction modeller (inceleme kapsamında)
Mimari Özellikler	Üç aşamalı öğrenme: kontrastif öğrenme, self-supervised öğrenme, cross-model alignment (MSE)	Multi-Modal Transformer (MM-MHA), Spatial Transformer (S-MHA), Temporal Transformer (T-MHA)	Literatür temelli mimari sınıflandırma
Girdi Türü	Tarım ve hayvancılığa ait görüntüler + metin istemleri	Sentinel-2 uydu görüntüleri + kısa ve uzun vadeli HRRR meteorolojik veriler	Çeşitli RS veri setleri (inceleme)
Çıktı Türü	Zero-shot sınıf tahminleri	Sürekli mahsul verim tahminleri	Performans karşılaştırmaları ve metodolojik çıkarımlar
Veri Setleri	**ALive**: ~600K image–text çifti (25 veri seti)	MS, LA, IA, IL eyaletleri; corn, cotton, soybean, winter wheat	AID, RSICD ve diğer RS veri setleri
Değerlendirme Metrikleri	Top-1 Accuracy (zero-shot)	RMSE, R^2, Pearson korelasyonu	Top-1 Accuracy, Recall@K
Öne Çıkan Sonuçlar	20 veri setinde %48.27 zero-shot doğruluk; adapted CLIP'e göre +%9.07 mutlak artış	Soybean için RMSE = 3.9, R^2 = 0.843, Corr = 0.918	RemoteCLIP'in AID'de %87.9 doğruluk; PIR-CLIP'in RSICD'de SOTA
Güçlü Yan	İnce taneli tarımsal ayrımlarda güçlü zero-shot genelleme	İklimsel ve mekânsal bağımlılıkları birlikte modelleme	Alanın bütüncül ve sistematik çerçevesini sunma
Sınırlılıklar	Yüksek veri hazırlama maliyeti; alan dışı genellenebilirlik sınırlı	Bölgesel veri bağımlılığı ve hesaplama maliyeti	Deneysel katkı yerine metodolojik özet sunması

3.7.4.2. Saha ve Bitki Düzeyi Tanılama ve Karar Destek

Tarımsal üretimde hastalık ve zararlı tespiti ile bitki gelişim evrelerinin belirlenmesi, salt görsel benzerliklere dayalı sınıflandırma problemlerinin ötesindedir. Bu görevler, ince taneli görsel ipuçlarının bitki türü, gelişim aşaması ve çevresel bağlamla birlikte yorumlanmasını gerektirir. Bu nedenle

Görsel Dil Modelleri (VLM), tarımsal tanı problemlerini bağlama duyarlı ve açıklanabilir karar destek sistemleri çerçevesinde ele almayı hedeflemektedir.

Bu doğrultuda geliştirilen AgriVLM, tarımsal görüntü ve metin verilerini çapraz modlu (cross-modal) biçimde işleyerek hastalık tanıma, gelişim evresi belirleme ve tarımsal soru-cevap görevlerini tek bir model çatısı altında birleştirmeyi amaçlamaktadır (Zhang et al., 2023). Modelin temel mimarisi; görsel özellik çıkarımı için Vision Transformer (ViT) tabanlı bir görüntü kodlayıcı, bu görsel temsilleri dil modelinin anlamsal uzayına aktaran bir Q-Former bileşeni ve temel dil modeli olarak ChatGLM'den oluşmaktadır. Tarımsal alana uyarlama sürecinde sınırlı veriyle aşırı uyum (overfitting) riskini azaltmak amacıyla, LoRA tabanlı parametre-verimli ince ayar (parameter-efficient fine-tuning) stratejisi uygulanmıştır. Elde edilen deneysel sonuçlar, beş farklı veri seti üzerinde değerlendirilen modelin (doğruluk, kesinlik, duyarlılık ve F1 skoru metriklerine göre) tüm görevlerde %90'ın üzerinde doğruluk oranına ulaştığını ve ince taneli ayrım gerektiren görevlerde genel amaçlı modellere kıyasla çok daha tutarlı bir performans sunduğunu göstermektedir (Zhang et al., 2023).

Tarımsal hastalık ve zararlı tanısına odaklanan bir diğer güncel çalışma olan LLMI-CDP ise, görsel tanı çıktıları ile alana özgü önleme önerilerini tek bir çerçevede birleştirmeyi hedeflemektedir (Liu et al., 2024). Önerilen sistem mimarisi, VisualGLM temeli üzerine inşa edilmiştir. Bu yapıda, modelin genel görsel-dil hizalama yeteneğini korumak ve tarımsal bilgi aktarımını sağlamak amacıyla görsel ve dil bileşenleri dondurulmuş, yalnızca LoRA katmanları eğitilmiştir. Toplam 15 mahsul türü ve 141 hastalık/zararlı kategorisini kapsayan 2.498 görüntülük veri seti üzerinde yapılan değerlendirme sonuçları, LLMI-CDP'nin hastalık tanımada %78.8 ve zararlı tanımada %86.7 doğruluk oranlarına ulaşarak temel VisualGLM modelini belirgin şekilde geride bıraktığını kanıtlamaktadır (Liu et al., 2024).

Karşılaştırma ölçütü olarak kullanılan genel amaçlı VisualGLM (Du et al., 2023), büyük ölçekli ön-eğitime sahip olmasına rağmen tarıma özgü ince ayar içermediği için bu tür spesifik görevlerde sınırlı bir performans sergilemektedir. Bu durum, tarımsal tanı problemlerinde alan-özgül veri entegrasyonunun ve doğru uyarlama mimarilerinin kritik önemini açıkça ortaya koymaktadır. Tablo 13'te tarımsal saha ve bitki düzeyi tanılama için VLM tabanlı yaklaşımların karşılaştırılması sunulmuştur.

Tablo13. Tarımsal Saha ve Bitki Düzeyi Tanılama İçin VLM Tabanlı Yaklaşımların Karşılaştırılması

Özellikler	AgriVLM	LLMI-CDP	VisualGLM (Baseline)
Kaynak	Zhang et al. (2023)	Liu et al. (2024)	Du et al. (2023)
Çalışma Alanı	Tarımsal saha ve bitki düzeyi tanılama	Tarımsal hastalık ve zararlı tanısı	Genel amaçlı multimodal diyalog
Temel Amaç	Tarımsal görüntü ve metinlerden çoklu tanı ve karar destek üretmek	Hastalık/zararlı tanısı ve önleme önerilerini birleştirmek	Genel görsel dil etkileşimi sağlamak
Temel Yaklaşım	Çapraz modlu görsel dil temsil öğrenimi	Alan-özgül LoRA uyarlaması	Genel amaçlı VLM ön eğitimi
Kullanılan VLM / LLM	ChatGLM tabanlı AgriVLM	VisualGLM + ChatGLM	ChatGLM-6B
Görsel Kodlayıcı	ViT tabanlı encoder	VisualGLM image encoder	VisualGLM image encoder
Dil Modeli	ChatGLM	ChatGLM	ChatGLM-6B
Mimari Özellik	Q-Former, LoRA ince ayar	Q-Former, dondurulmuş backbone + LoRA	Standart multimodal hizalama
Girdi Türü	Bitki ve hayvan görüntüleri + metin	Tarımsal görüntüler + metin	Genel görüntü + metin
Çıktı Türü	Tanı etiketleri, büyüme evresi, VQA yanıtları	Tanı etiketleri ve önleme önerileri	Görsel açıklama ve diyalog
Veri Setleri	5 görev veri seti + 5.593 tarımsal S-C örneği	2.498 görüntü, 141 sınıf	30M Çince + 300M İngilizce image–text
Değerlendirme Metrikleri	Accuracy, Precision, Recall, F1	Accuracy, insan & GPT-4 değerlendirmesi	Accuracy, VQA puanı
Öne Çıkan Sonuçlar	Tüm görevlerde >%90 doğruluk	Hastalık %78.8, zararlı %86.7	Hastalık %67.4, zararlı %72.4
Güçlü Yan	İnce taneli tarımsal görevlerde yüksek doğruluk	Tanı + öneri üretimini tek sistemde birleştirmesi	Güçlü genel dil yeteneği
Sınırlılıklar	Veri setleri görece küçük	Veri dili ağırlıklı olarak Çince	Tarımsal uzmanlık sınırlı

3.7.4.3. Tarım Odaklı Görsel Dil Model Tasarımı, Alan Uyarlaması ve Değerlendirme

Tarım alanında Görsel Dil Modellerinin (VLM) etkin biçimde kullanılabilmesi, genel amaçlı büyük modellerin ötesine geçerek; bitki hastalıkları, zararlı türleri ve yabani otlar gibi birbirine son derece benzeyen sınıflar için ince taneli görsel ayrımların ve uzmanlık gerektiren kavramsal etiketlerin modele kazandırılmasını gerektirmektedir. Sadece görsel benzerliğe dayalı genellemenin yetersiz kaldığı bu alanda literatür, tarımsal görevler için güvenilir değerlendirme ortamlarının oluşturulması ve tarıma özgü bilgiyle eğitilmiş alan uyarlamalı modellerin geliştirilmesi olmak üzere iki tamamlayıcı eksen etrafında şekillenmeyi hedeflemektedir (Shinoda et al., 2024; Zhou et al., 2024).

Bu doğrultuda birinci ekseni temsil eden AgroBench, tarım odaklı görsel dil modellerinin performansını gerçek dünya senaryoları üzerinden sistematik biçimde ölçmeyi amaçlayan kapsamlı bir değerlendirme çalışmasıdır (Shinoda et al., 2024). Sentetik ve sınırlı veri setlerinin ötesine geçmeyi hedefleyen bu kıyaslama mimarisi; 203 mahsul türü, 682 hastalık, 134 zararlı, 108 yabani ot ve 98 tarımsal makine kategorisini kapsayan 4.218 görüntü ile 4.342 soru-cevap çiftinden oluşturulmuştur. Yedi temel görev altında gruplanan AgroBench üzerinde yapılan değerlendirme sonuçları, kapalı kaynaklı modellerde GPT-4o'nun, açık kaynaklılarda ise QwenVLM-72B'nin en yüksek performansı sergilediğini göstermektedir. Ancak hata analizleri, model başarısızlıklarının büyük ölçüde alan bilgisi eksikliğinden (%51.92) ve algısal hatalardan (%32.69) kaynaklandığını; özellikle görsel benzerliklerin çok yüksek olduğu yabani ot tanımlama görevinin modeller için en zorlayıcı test olduğunu kanıtlamıştır (Shinoda et al., 2024).

AgroBench'in ortaya koyduğu bu alan bilgisi eksikliğini doğrudan çözmeyi amaçlayan AgroGPT ise, tarımsal uzmanlık kazandırılmış bir VLM geliştirme yaklaşımı sunmaktadır (Zhou et al., 2024). Zengin görüntü-metin çiftlerinin yokluğunu aşmak hedefiyle, genel amaçlı modeller ve uzman kaynaklar kullanılarak yaklaşık 70.000 diyalogluk talimat tabanlı bir eğitim veri seti (AgroInstruct) oluşturulmuştur. Modelin temel mimarisi, LLaVA yapısı üzerine inşa edilmiş olup görsel kodlayıcı olarak CLIP yerine SigLIP kullanmaktadır. AgroGPT-3B ve AgroGPT-7B olarak iki ölçekte tasarlanan bu mimari; görsel-dil hizalama, genel görsel talimat ayarlama ve tarımsal uzmanlık kazandırmayı amaçlayan "expert tuning" (uzman ince ayarı) olmak üzere üç basamaklı bir eğitim sürecinden geçirilmiştir.

Elde edilen deneysel sonuçlar, AgroGPT'nin özellikle ince taneli hastalık ve zararlı tanımlama görevlerinde genel amaçlı büyük modellere kıyasla belirgin

bir üstünlük sağladığını ortaya koymaktadır. Hastalık tanımlama görevinde AgroGPT %51.37 doğruluğa ulaşırken, ChatGPT (%30.82) ve temel LLaVA modeli (%10.27) oldukça geride kalmıştır. Ayrıca insan uzmanlarla yapılan kör değerlendirme sonuçlarında, AgroGPT çıktılarının %86–%96 gibi ezici bir oranla daha fazla tercih edildiği rapor edilmiştir (Zhou et al., 2024).

Özetle; AgroBench tarımsal görevlerde mevcut modellerin sınırlarını sistematik biçimde ortaya koyarken, AgroGPT bu sınırların temel nedeni olan alan bilgisi eksikliğine yönelik uygulanabilir ve başarılı bir mimari çözüm sunmaktadır (Tablo 14).

Tablo 14. Tarım Odaklı VLM Tasarımı ve Değerlendirmesi

Özellikler	AgroBench	AgroGPT
Kaynak	Shinoda et al. (2024)	Zhou et al. (2024)
Çalışma Türü	Benchmark / değerlendirme	Model + veri üretimi
Temel Amaç	Tarımsal görsel dil modeller için uzman onaylı standart test ortamı oluşturmak	Tarımsal uzmanlık kazandırılmış verimli VLM geliştirmek
Temel Yaklaşım	Gerçek görüntüler ve uzman doğrulamalı QA ile kıyaslama	Instruction tuning + expert tuning
Kullanılan VLM / LLM	Mevcut görsel dil modeller (GPT-4o, QwenVLM vb.)	AgroGPT-3B, AgroGPT-7B
Görsel Kodlayıcı	—	SigLIP
Dil Modeli	—	LLaVA tabanlı LLM
Mimari Özellik	Model üretmez, performans analizi yapar	LLaVA tabanlı, üç aşamalı eğitim
Girdi Türü	Tarımsal görüntüler + görsel QA	Tarımsal görüntüler + talimatlar
Çıktı Türü	Doğruluk, hata analizi	Tanı ve açıklamalı yanıtlar
Veri Setleri	4.218 görüntü, 4.342 QA, 7 görev	AgroInstruct (~70k), AgroEvals
Değerlendirme Metrikleri	Accuracy	Accuracy, insan tercihi
Öne Çıkan Sonuçlar	En büyük hata kaynağı bilgi eksikliği (%51.92)	Küçük modellerle yüksek tarımsal uzmanlık
Güçlü Yan	Güvenilir değerlendirme standardı	Alan-özgü bilgi entegrasyonu
Sınırlılıklar	Model üretmez	Sentetik talimatlara bağımlılık

3.7.5. Video Anlama ve Zaman-Bağımlı Görsel Dil Modelleri Uygulamaları

Görsel Dil Modellerinin (VLM) son yıllardaki gelişimi, bu yapıların tekil görüntülerden çıkarak zaman boyutu içeren video verileri üzerinde de kullanılabilirliğini gündeme getirmiştir. Video verisi; zaman içindeki değişimleri, hareket örüntülerini ve sahne sürekliliğini barındırması nedeniyle tek görüntüye dayalı algıdan çok daha karmaşık bir temsil ve akıl yürütme problemi sunmaktadır. Bu nedenle güncel literatür, doğrudan zamansal modelleme yapmak yerine, video verisinin dönüştürülerek uygun temsiller aracılığıyla VLM'lere sunulmasını hedeflemektedir (Chen et al., 2024; Wang et al., 2024).

Bu bağlamda geliştirilen IG-VLM (Image Grid-based Vision–Language Model) yaklaşımı, video anlama problemini özel bir video modeli eğitmeye gerek kalmadan çözmeyi amaçlamaktadır (Chen et al., 2024). Sistemin temel mimarisinde, videodan zamansal olarak örneklenen kareler tek bir birleşik görüntü ızgarası (image grid) hâlinde düzenlenerek önceden eğitilmiş (CogAgent, LLaVA v1.6 ve GPT-4V gibi) bir VLM'e tek girdi olarak sunulur. Böylece video boyunca gerçekleşen olayların zamansal özeti, modelin güçlü görüntü-metin hizalama ve muhakeme yetenekleri üzerinden dolaylı olarak modellenir. MSVD-QA, MSRVTT-QA, ActivityNet-QA ve TGIF-QA veri setlerindeki deneysel sonuçlar; kare sayısının, sıralamasının ve ızgara yerleşiminin performansı doğrudan etkilediğini kanıtlamaktadır. Bu bulgular, video için özel olarak eğitilmemiş genel modellerin dahi uygun görsel temsil stratejileriyle zamansal akıl yürütme gerektiren görevleri başarıyla yerine getirebildiğini ortaya koymaktadır (Chen et al., 2024).

Video tabanlı VLM kullanımının gerçek dünya uygulamalarına yönelik bir diğer yenilikçi yaklaşımı, araç gözetimi ve trafik denetimi alanında plaka (ALPR) ile marka-model tanıma süreçlerini hedeflemektedir (Wang et al., 2024). Sabit trafik kameralarının yüksek kurulum maliyetlerine karşın, akıllı telefonlardan veya araç içi kameralardan elde edilen hareketli ve düşük kaliteli videoların kullanımını amaçlayan bu sistem; GPT-4o, Llama-3.2-Vision, LLaVA ve MiniCPM-V gibi modelleri hiçbir özel eğitim gerektirmeden sıfır-atış (zero-shot) biçimde kullanmaktadır. Kurulan yenilikçi mimaride, video karelerinin kalitesini değerlendirmek ve en bilgilendirici olanları seçmek için CLIP-IQA ve BRISQUE tabanlı ölçütler kullanılmıştır. Ayrıca, araç marka ve model tanıma sürecinde modelin ilk tahminini web'den toplanan referans görüntülerle karşılaştıran ve gerekirse sorguyu düzelten bir öz-yansıtma (self-reflection) modülü geliştirilmiştir.

UFPR-ALPR ve gerçek dünya akıllı telefon video veri setleri üzerindeki değerlendirme sonuçları, plaka tanımada %90'ın üzerinde doğruluk elde edildiğini ve marka-model tespitinde geleneksel OCR tabanlı yöntemlere kıyasla belirgin bir üstünlük sağlandığını göstermektedir (Wang et al., 2024). Bu durum, görsel dil modellerinin düşük çözünürlük, hareket bulanıklığı ve açı değişimi gibi zorlu koşullarda oldukça dayanıklı çözümler üretebildiğini doğrulamaktadır.

Genel bir çerçeveden değerlendirildiğinde; görsel dil modellerinin video anlama görevlerinde doğrudan zamansal modelleme yerine temsil dönüştürme, kare seçimi ve test-zamanı muhakeme mimarileriyle son derece etkili biçimde kullanılabildiği görülmektedir (Tablo 15). Bu yaklaşımlar, VLM'lerin yalnızca statik algı sistemleri olmaktan çıkarak, zaman-bağımlı karar destek bileşenlerine evrildiğini açıkça ortaya koymaktadır (Chen et al., 2024; Wang et al., 2024).

Tablo 15. Video Anlama ve Zaman-Bağımlı VLM Uygulamaları

Özellikler	IG-VLM	Self-Reflective VLM (Araç Gözetimi)
Kaynak	Chen et al. (2024)	Wang et al. (2024)
Çalışmanın Amacı	Video-QA görevlerini video için özel model eğitmeden VLM ile çözmek	Gerçek dünya videolarından ölçeklenebilir trafik denetimi sağlamak
Problem Tanımı	Uzun video dizilerinde zamansal bilginin görsel dil modellerde kaybolması	Düşük kalite ve hareketli videolarda ALPR başarısızlığı
Temel Yaklaşım	Videoyu image grid temsiline dönüştürerek VLM'e girdi vermek	Zero-shot VLM + öz-yansıtma ile doğrulama
Kullanılan Görsel Dil Modeller	CogAgent, LLaVA v1.6, GPT-4V	GPT-4o, Llama-3.2-Vision, LLaVA, MiniCPM-V
Zamansal Bilgi Temsili	Kare örnekleme + grid düzeni	Kare seçimi + çoklu sorgulama
Kare Seçimi Stratejisi	Sabit aralıklı örnekleme	CLIP-IQA, BRISQUE
Ek Modüller	Grid ve reasoning guidance promptları	Self-reflection, SAM tabanlı arka plan temizleme
Veri Setleri	MSVD-QA, MSRVTT-QA, ActivityNet-QA, TGIF-QA	UFPR-ALPR, gerçek dünya akıllı telefon videoları
Değerlendirme Görevleri	Video Question Answering	ALPR, araç marka/model tanıma
Değerlendirme Metrikleri	Accuracy, VQA Score	Top-1 Accuracy

Öne Çıkan Sonuçlar	Zero-shot görsel dil modellerle etkili video-QA	ALPR'de %90+ doğruluk
Güçlü Yanlar	Video modeli gerektirmemesi	Zorlu koşullara dayanıklılık
Sınırlılıklar	Uzun olay zincirlerinde grid kapasitesi	Aşırı bulanıklık hâlen zorlayıcı

3.7.6. İnsan Odaklı Görsel Dil Modelleri Uygulamaları

İnsan odaklı görsel anlama (Human-Centric Vision), insan vücudu ve yüzüne ait görsel ipuçlarından poz (human pose), yüz ifadesi (facial expression) ve davranış (human behavior) gibi yüksek seviyeli semantik bilgilerin çıkarılmasını amaçlamaktadır. Geleneksel yöntemlerin yalnızca görsel özniteliklere dayalı sınırlılıklarını aşmayı hedefleyen güncel yaklaşımlar, Görsel Dil Modellerinin (VLM) anlamsal ve bağlamsal yorumlama yeteneklerini sürece entegre etmektedir.

İnsan pozu tahmini alanında (Human Pose Estimation) önerilen "Vision-Language Model Guided Pose Knowledge Mining" yaklaşımı, poz bilgisinin otomatik olarak çıkarılmasını ve bu bilginin poz tahmin modellerine aktarılmasını hedeflemektedir (Zhang et al., 2023). Önerilen mimaride, CLIP tabanlı bir görsel-dil modeli kullanılarak poz içeren görüntüler; poz türü, vücut duruşu ve hareket bağlamını tanımlayan metinsel açıklamalarla hizalanır. Elde edilen bu yüksek seviyeli anlamsal bilgiler, klasik poz tahmin ağlarına yardımcı denetim (auxiliary supervision) sinyali olarak entegre edilir. Böylece sistemin yalnızca eklem noktaları arasındaki geometrik ilişkileri değil, pozun bütüncül anlamsal yapısını da öğrenmesi sağlanır. Deneysel sonuçlar, özellikle etiketli verinin sınırlı olduğu (low-label regime) ve karmaşık pozların yer aldığı senaryolarda bu yaklaşımın tahmin performansını anlamlı biçimde artırdığını göstermektedir (Zhang et al., 2023).

Yüz ifadesi analizi (Facial Expression Analysis) alanında öne çıkan SMILE-VLM yaklaşımı ise, 3B ve 4B yüz ifadesi tanıma problemini kendinden denetimli öğrenme (self-supervised learning) ve çoklu görünüm temsili (multi-view representation) perspektifinden çözmeyi amaçlamaktadır (Li et al., 2024). Bu sistemin mimarisinde VLM'ler, farklı kamera açıları ve zaman adımlarında elde edilen yüz görüntülerinin ortak ve tutarlı anlamsal temsillerini öğrenmek üzere konumlandırılmıştır. Görsel-dil hizalaması sayesinde, yüz kaslarının uzamsal ve zamansal değişimleri ile duygusal durumlar arasındaki karmaşık ilişkiler çok daha etkili biçimde modellenir. Elde edilen sonuçlar, etiketlenmiş 3B/4B verilerin kısıtlı olduğu durumlarda bile bu yaklaşımın tam denetimli

yöntemlerle rekabet edebilir bir performans ve yüksek genellenebilirlik sunduğunu kanıtlamaktadır (Li et al., 2024).

Bu çalışmalar birlikte değerlendirildiğinde, görsel–dil modellerinin insan odaklı görsel analiz görevlerinde doğrudan bir tahmin edici bileşen olmaktan ziyade, bilgi çıkarımı (knowledge extraction) ve temsil zenginleştirme (representation enrichment) aracı olarak konumlandığı görülmektedir. Poz ve yüz ifadesi gibi insan davranışını yansıtan karmaşık görsel yapıların VLM destekli yaklaşımlarla modellenmesi, yalnızca görsel bilgilere dayanan yöntemlerin sınırlamalarını tamamlayıcı nitelikte, çok modlu ve bağlama duyarlı bir analiz çerçevesi sunmaktadır. Tablo 16' da insan odaklı Görsel–Dil Modelleme (Human-Centric Vision–Language Modeling) çalışmaları detaylı olarak analiz edilmiştir.

Tablo 16. İnsan Odaklı Görsel–Dil Modelleme (Human-Centric Vision–Language Modeling) Çalışmaları

Özellikler	VLM-Guided Pose Knowledge Mining	SMILE-VLM
Kaynak	Zhang et al. (2023)	Li et al. (2024)
Problem Alanı	İnsan pozu tahmini (Human Pose Estimation)	3B / 4B yüz ifadesi tanıma
Temel Amaç	Pozlara ait yüksek seviyeli anlamsal bilginin çıkarılması ve poz tahminine aktarılması	Etiket gerektirmeden çoklu görünüm ve zamansal yüz temsilleri öğrenmek
Model Türü	CLIP tabanlı VLM + klasik pose estimation ağı	CLIP tabanlı VLM + kendinden denetimli öğrenme
Görsel Temsil	Poz içeren görüntülerden çıkarılan görsel gömüler	Çoklu kamera ve zamansal yüz projeksiyonları
Dil Bileşeni	Poz, duruş ve hareketi tanımlayan metinsel açıklamalar	Yüz ifadelerine dair anlamsal dil temsilleri
Temel Yaklaşım	Poz bilgisi madenciliği ve yardımcı denetim	Self-supervised multi-view öğrenme + VLM hizalaması
Öğrenme Stratejisi	Anlamsal bilginin aşağı akış poz tahmin görevine transferi	Ortak görsel–dil anlamsal uzay öğrenimi
Kullanılan Veri Setleri	COCO, MPII	BU-3DFE, BU-4DFE, Bosphorus, BP4D
Değerlendirme Senaryosu	Sınırlı etiketli ve karmaşık poz içeren görüntüler	Etiketsiz 3B/4B yüz ifadesi tanıma
Başlıca Bulgular	Karmaşık ve az etiketli durumlarda doğruluk artışı	Denetimli yöntemlerle karşılaştırılabilir performans
Uygulama Alanları	İnsan hareketi ve davranış analizi	Affective computing, HCI, mental sağlık

4. Değerlendirme ve Sonuç

İncelenen makaleler, büyük dil modelleri ve görsel dil modellerinin farklı disiplinlerde önemli ilerlemeler sağladığını ancak her alanın kendine özgü yapısal eksiklikler ve çözülmemiş sorunlar barındırdığını göstermektedir.

Görsel Dil Modelleri (VLM) ve bu modellerin eylem boyutuyla genişletilmiş versiyonları olan Vision–Language–Action (VLA) yaklaşımları, yapay zekâ literatüründe yalnızca görsel algıyı iyileştirmekle kalmamış; bilişsel muhakeme, bağlamsal yorumlama ve karar verme süreçlerini de kökten değiştirerek önemli bir paradigma dönüşümü yaratmıştır. İncelenen tüm çalışmalar geniş bir perspektiften değerlendirildiğinde; otonom sistemlerden robotiğe, sağlık bilimlerinden tarıma ve insan odaklı video analizlerine kadar pek çok alanda, salt geometrik ve modüler algı mimarilerinin yerini daha bütüncül, açıklanabilir, çok modlu ve bağlama duyarlı yapıların aldığı açıkça görülmektedir.

Otonom sistemler ve robotik alanında, klasik nokta bulutu veya sınırlayıcı kutu (bounding box) temelli algı sistemleri, VLM'lerin açık sözcük dağarcıklı (open-vocabulary) ve sıfır-atışlı (zero-shot) yetenekleri sayesinde yepyeni semantik bir derinlik kazanmıştır. Otonom sürüşte dil destekli bilgi damıtma stratejileri ve robotikte eylemlerin metinsel belirteçler (action-as-text) olarak modellenmesi, makinelerin görevleri sadece geometrik bir optimizasyonla değil, insan benzeri bir muhakemeyle yorumlamasını sağlamıştır. Bu sayede düşük seviyeli fiziksel özellik tahmininden sosyal navigasyona kadar uzanan geniş bir yelpazede, çevrenin çok daha bütüncül bir bilişsel çerçevede algılanması mümkün kılınmıştır.

Sağlık ve tarım gibi yüksek uzmanlık gerektiren alanlarda ise bu modellerin spesifik ihtiyaçlara göre evrildiği görülmektedir. Sağlık alanında VLM tabanlı sistemler, basit rapor üretiminin ötesine geçerek yapılandırılmış bilgi grafiği oluşturma, kendi kendini iyileştirme (self-refining) mekanizmalarıyla halüsinasyon azaltma ve çok ajanlı karar destek çerçevelerine dönüşmüştür. Etiket kıtlığı yaşanan biyomedikal alanlarda "foundation" (temel) modellerin sıfır-atışlı etiket üretimi büyük avantajlar sunarken; tarım ve çevresel planlamada genel amaçlı modellerin ince taneli sınıflarda (zararlı, hastalık vb.) yetersiz kaldığı tespit edilmiştir. Bu bağlamda, AgroGPT gibi alan-özgü ince ayarlı (expert-tuned) modellerin ve AgroBench gibi uzman onaylı değerlendirme standartlarının, sistem başarısında model ölçeğinden bile daha belirleyici olduğu kanıtlanmıştır. Benzer şekilde video analizi ve insan odaklı görevlerde, ağır zamansal modellemeler yerine görsel ızgara (image grid) gibi temsil dönüştürme stratejileri ve VLM destekli bilgi madenciliği, karmaşık davranışsal göstergeleri başarıyla semantik düzleme taşımaktadır.

Elde edilen tüm bu umut verici metodolojik kazanımlara rağmen literatür, bu modellerin doğrudan gerçek dünya sistemlerine aktarımında kritik mühendislik darboğazlarına işaret etmektedir. Milyarlarca parametreli modellerin düşük frekanslarda çalışması ve yüksek hesaplama maliyetleri, otonom sürüş veya robotik kontrol gibi anlık tepki gerektiren güvenlik-kritik senaryolarda doğrudan kullanımı sınırlandırmaktadır. Ek olarak, mekânsal ve anlamsal halüsinasyon riskleri, medikal ve endüstriyel ortamlar gibi hataya sıfır toleranslı alanlarda VLM'lerin tek başına karar alıcı olarak konumlandırılmasını oldukça riskli hâle getirmektedir. Ayrıca bulut bağımlılığı ve donanım kısıtları, beraberinde veri gizliliği ve ağ gecikmesi problemlerini getirmektedir.

Bu tabloyu göz önünde bulundurduğumuzda, VLM ve VLA tabanlı sistemlerin geleceğinin yalnızca "daha büyük modeller" üretmekten ibaret olmadığı düşünülmektedir. Aksine geleceğin vizyonu; bilgi damıtma, niceleme (quantization) ve LoRA gibi parametre-verimli tekniklerle sıkıştırılmış, uç cihazlarda (edge) çalışabilen, göreve özgü uyarlanmış ve çok ajanlı (multi-agent) mimarilerle desteklenen hibrit sistemlere işaret etmektedir. Kısa ve orta vadede bu modeller, bağımsız kontrolcü sistemler olmaktan ziyade; üst düzey semantik rehberlik sağlayan, süreçlerin açıklanabilirliğini artıran ve insan operatörlerle uyum içinde çalışan "bilişsel karar destek bileşenleri" olarak konumlanacaktır. Görsel dil modellerinin başlattığı bu güçlü metodolojik dönüşümün güvenli, etik ve sürdürülebilir biçimde ilerlemesi; ancak alan-özgü uyarlama, sıkı denetim mekanizmaları ve güvenilirlik odaklı tasarım ilkelerinin bütüncül bir yaklaşımla benimsenmesine bağlıdır.

Kaynaklar

Alammar, J. (2024). Hands-on large language models: Language understanding and generation. O'Reilly Media.

Bimbraw, K., Singh, A., Patel, R., & Ghosh, S. (2024). Hand gesture decoding from forearm ultrasound images via vision–language models. IEEE Transactions on Medical Imaging. https://arxiv.org/abs/2407.10870

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., & Chandra, V. (2024). An introduction to vision-language modeling. arXiv. https://arxiv.org/abs/2405.17247

Brohan, A., Brown, N., Carbajal, J., et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv. https://arxiv.org/abs/2307.15818

Cai, H., et al. (2025). Unilaw-R1: A large language model for legal reasoning with reinforcement learning and iterative inference. arXiv.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2022). LexGLUE: A benchmark dataset for legal language understanding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL).

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), Article 39. https://doi.org/10.1145/3641289

Chen, Y., Ding, Z.-H., Wang, Z., Wang, Y., Zhang, L., & Liu, S. (2024). Asynchronous large language model enhanced planner for autonomous driving (AsyncDriver). arXiv. https://arxiv.org/abs/2406.14556

Chen, X., Li, Y., Wang, H., & Zhang, Z. (2024). An image grid can be worth a video: Zero-shot video question answering using a vision–language model. arXiv.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2023). VisualGLM: Visual and language model for multimodal dialogue. arXiv. https://arxiv.org/abs/2304.08857

Encord. (2023). LLaVA: Large language and vision assistant. https://encord.com/blog/llava-large-language-vision-assistant/

Gkournelos, C., et al. (2024). An LLM-based approach for enabling seamless human–robot collaboration in assembly. IEEE Robotics and Automation Letters.

Google DeepMind. (2024). AutoRT: Embodied foundation models for large-scale orchestration of robotic agents. https://deepmind.google/discover/blog/

Greco, A., et al. (2025). An experimental evaluation of smart sensors for pedestrian attribute recognition using multi-task learning and vision–language models.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., & Chilton, L. (2023). LegalBench: A benchmark for measuring legal reasoning in large language models. arXiv. https://arxiv.org/abs/2308.11462

He, K., et al. (2025). A survey of large language models for healthcare.

Ho, H.-T., Nguyen, L. V., Pham, M.-T., Pham, Q.-H., Tran, Q.-D., Huy, D. N. M., & Nguyen, T.-H. (2025). A review on vision-language-based approaches: Challenges and applications. Computers, Materials & Continua, 82(2), 1733–1760. https://doi.org/10.32604/cmc.2025.060363

Hu, Y., et al. (2023). Planning with vision-language models and a use case in robot-assisted teaching.

Hugging Face. (n.d.). What are LLMs? In Agents Course. https://huggingface.co/learn/agents-course/unit1/what-are-llms

Jadhav, A., & Mirza, V. (2025). Large language models in equity markets. Frontiers in Artificial Intelligence, 8, 1608365. https://doi.org/10.3389/frai.2025.1608365

Jia, F., Mao, W., Liu, Y., Zhao, Y., Wen, Y., Zhang, C., Zhang, X., Wang, T., et al. (2023). ADriver-I: A general world model for autonomous driving. arXiv. https://arxiv.org/abs/2311.13549

Jung, K. H. (2025). Large language models in medicine: Clinical applications, limitations, and ethics.

Kawaharazuka, K., Iwase, M., & Sugita, N. (2023). Vision-language-action models for robotics: A review towards real-world applications. arXiv.

Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). Large language models for financial and investment management: Applications and benchmarks.

Li, B., et al. (2025). CoLE: A collaborative legal expert prompting framework for large language models in law.

Li, J., et al. (2025). Guiding multiple remote users in physical tasks with language-driven robotic telepresence. ACM Human–Computer Interaction.

Li, Q., Zhao, Y., Sun, Z., & Wang, X. (2024). SMILE-VLM: Self-supervised multi-view representation learning using vision-language models for 3D/4D facial expression recognition. arXiv.

Li, T., Wang, H., Tan, J., Kong, L., Zhang, H., Pan, D., & Zhao, Z. (2025). Intelligent quality assessment of concrete vibration using computer vision and large language models. Automation in Construction, 180, 106507. https://doi.org/10.1016/j.autcon.2025.106507

Lin, C., et al. (2025). Roles and potential of large language models in healthcare.

Lin, T., Zhang, W., Li, S., Yuan, Y., Yu, B., Li, H., He, W., Jiang, H., Li, M., Song, X., Tang, S., Xiao, J., Lin, H., Zhuang, Y., & Ooi, B. C. (2025). HealthGPT: A medical large vision-language model for unifying comp-

rehension and generation via heterogeneous knowledge adaptation. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025) (Proceedings of Machine Learning Research, Vol. 267, pp. 37975–37995). PMLR. https://proceedings.mlr.press/v267/lin25a.html

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Multilingual visual question-answering for dermatology through vision–language model fine-tuning and large language model translations. In Proceedings of the MEDIQA-M3G 2024 Shared Task.

Liu, J., Wang, Y., Zhang, Q., Zhao, S., & Li, H. (2024). A large language model for multimodal identification of crop diseases and pests. arXiv.

Ma, Y., Yao, Z., Liu, X., Xiong, Z., He, X., & Wu, X. (2026). Efficient 3D object annotation via vision-derived pseudo-LiDAR and vision–language model validation. Transportation Research Part C: Emerging Technologies, 182, 105429. https://doi.org/10.1016/j.trc.2025.105429

Macdonald, J., et al. (2023). Language, camera, autonomy! Prompt-engineered robot control for rapidly evolving deployment. arXiv.

Mei, T., Zhang, W., & Yao, T. (2019). Vision and language: From visual perception to content creation. Signal Processing, 164, 22–35. https://doi.org/10.1016/j.sigpro.2019.06.008

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). Large language models: A survey. arXiv. https://arxiv.org/abs/2402.06196

Navarro, H. J., et al. (2025). Large language models in medicine: A systematic review.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). A comprehensive overview of large language models. arXiv. https://arxiv.org/abs/2307.06435

Ong, J. C. L., et al. (2025). Ethical and regulatory challenges of large language models in healthcare.

OpenAI. (2021). CLIP: Connecting vision and language. https://openai.com/index/clip/

Osada, M., Garcia Ricardez, G. A., Suzuki, Y., & Taniguchi, T. (2024). Reflectance estimation for proximity sensing by vision-language models: Utilizing distributional semantics for low-level cognition in robotics. Advanced Robotics, 38(18), 1287–1306. https://doi.org/10.1080/01691864.2024.2393408

Patil, R., & Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (LLMs). Applied Sciences, 14(5), 2074. https://doi.org/10.3390/app14052074

Peng, Y., et al. (2024). Vision-language model-enabled street view analytics: A systematic literature review.

Peng, H., Zhang, L., Wu, J., & Patel, V. M. (2024). PRIMA: A foundation vision-language model for large-scale neuroimaging analysis. arXiv.

Pollini, A., et al. (2024). Reducing latency in LLM-based robot control via ROS 2 integration.

Polo Club. (n.d.). Transformer explainer: LLM transformer model visually explained. https://poloclub.github.io/transformer-explainer

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML).

SaM Solutions. (n.d.). LLM architecture: A comprehensive guide. https://sam-solutions.com/blog/llm-architecture

Sapkota, M., et al. (2025). A review of 3D object detection with vision–language models.

Sapkota, S., Zhang, Y., Wang, L., et al. (2025). Vision-language-action models: Concepts, progress, applications and challenges. arXiv.

Shao, M., Basit, A., Karri, R., & Shafique, M. (2024). Survey of different large language model architectures: Trends, benchmarks, and challenges (arXiv:2412.03220). arXiv. https://arxiv.org/abs/2412.03220

Sharshar, M., Kanaan, H., & Abou-Zeid, H. (2023). Robotic environmental state recognition with pre-trained vision-language models and black-box optimization. Robotics and Autonomous Systems, 168, 104504.

Sharshar, A., Khan, L. U., Ullah, W., & Guizani, M. (2025). Vision-language models for edge networks: A comprehensive survey. IEEE Access. Advance online publication. https://arxiv.org/abs/2502.07855

Shekhar, A. C., et al. (2025). Use of a large language model for ambulance dispatch and triage.

Shinoda, R., Inoue, N., Kataoka, H., Onishi, M., & Ushiku, Y. (2024). AgroBench: Vision–language model benchmark in agriculture. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://arxiv.org/abs/2403.08819

Shu, D., et al. (2024). LawLLM: Law large language model for the US legal system.

Song, X., et al. (2024). VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models.

Ullah, E., et al. (2024). Challenges and barriers of large language models in digital pathology.

Wang, J. (2024). Hallucination reduction and optimization for large language model-based autonomous driving. Symmetry, 16(9), 1196.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345.

Wang, Y., Li, J., Xu, R., & Zhao, D. (2024). Vision language models in autonomous driving: A survey and outlook.

Wang, Y., et al. (2025). Vision-language model-based human-guided mobile robot navigation in an unstructured environment for human-centric smart manufacturing.

Xia. (2025, June 15). How vision language models are trained: A deep dive into the VLM training process. Medium. https://medium.com/@xiaxiami/how-vision-language-models-are-trained-a-deep-dive-into-the-vlm-training-process-1ba1d8704bb0

Xu, Z., Wang, Y., Chen, X., Liu, J., & Shi, B. (2024). VLM-Grounder: A vision-language model agent for zero-shot 3D visual grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 19,845–19,855).

Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. arXiv. https://arxiv.org/abs/2306.06031

Yang, R., et al. (2023). Large language models in health care.

Zhang, C., et al. (2024). NaVid: Video-based vision-language model plans the next step for vision-and-language navigation.

Zhang, J., Xu, C., & Li, B. (2024). ChatScene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. arXiv. https://arxiv.org/abs/2405.14062

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://arxiv.org/abs/2304.00685

Zhang, K., et al. (2025). Revolutionizing healthcare with large language models.

Zhang, P., Lin, K., Li, D., Fu, Z., Cai, Y., Li, B., Yu, H., & Li, M. (2025). DAPlanner: Dual-agent framework with multimodal large language model for autonomous driving motion planning. Applied Soft Computing.

Zhang, Y., et al. (2024). Agentic LLM-based robotic systems for real-world applications: A survey.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2025). A survey of large language models (Version 16). arXiv. https://arxiv.org/abs/2303.18223

Zhong, L., Huang, Z., Liu, Y., Liao, W., Zhang, S., Wang, G., & Zhang, S. (2025). VLM-CPL: Consensus pseudo-labels from vision-language models for annotation-free pathological image classification. IEEE Transactions on Medical Imaging, 44(10), 4023–4036. https://doi.org/10.1109/TMI.2025.3595111

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9), 2337–2348. https://arxiv.org/abs/2109.01134

Zhou, Y., Wang, H., Li, X., Zhang, Y., & Sun, J. (2024). AgroGPT: Efficient agricultural vision–language model with expert tuning. arXiv.

Zhu, H., Zhang, Y., Li, Z., & Zhu, Y. (2024). RoboPoint: A vision-language model for spatial affordance prediction for robotics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 14,732–14,741).