2026 34th Signal Processing and Communications Applications Conference (SIU) | 979-8-3195-1046-4/26/$31.00 ©2026 IEEE | DOI: 10.1109/SIU71813.2026.11636432

Görüntü-Dil-Aksiyon Modellerinin Çapraz Dil Performans Analizi: OpenVLA'da Türkçe Komut Anlayışının Değerlendirilmesi Cross-Lingual Performance Analysis of Vision-Language-Action Models: Evaluating Turkish Command Understanding in OpenVLA

Hasan Tatar[1], Umut Gökmen[2], Ilgın Akkoyun[3], Sevcan Kahraman[4]
[1,2,4]*İstanbul Ticaret Üniversitesi, İstanbul, Türkiye* [3]*Yıldız Teknik Üniversitesi, İstanbul, Türkiye*
[1]y23lbim801@istanbulticaret.edu.tr, [2]umut.gokmen@istanbulticaret.edu.tr, [3]ilgin.akkoyun@std.yildiz.edu.tr, [4]skahraman@ticaret.edu.tr

***Özetçe*—Görüntü-Dil-Aksiyon (VLA) modelleri, görsel bilgiyi dil anlayışıyla ve eylem üretimiyle bir araya getirmektedir. Ancak bu modeller genellikle İngilizce verilerle eğitildiği için diğer dillere göre daha iyi performans sergilemektedir. Bu çalışmada, OpenVLA-7B modelinin İngilizce ve Türkçe komutlarla ürettiği robot eylemleri karşılaştırılmıştır. Aynı 50 manipülasyon komutunu iki dilde de uygulayarak elde edilen ölçümlerde önemli farklar gözlemlenmiştir: Diller arasında ortalama L2 hata değeri 0,501 olarak belirlenirken, kosinüs benzerliği ise 0,372 seviyesinde kalmıştır. Her iki ölçüm de istatistiksel olarak anlamlı bulunmuştur. Komut karmaşıklığı ile performans düşüşü arasında doğrusal bir ilişki bulunamamıştır. Türkçe komutlar yüzde 30,8 daha fazla token üretse de, çıkarım süresinde anlamlı bir gecikme gözlenmemiştir. Elde edilen sonuçlar, şu anki VLA modellerinin çok dilli eylem ürettiklerinde önemli bir eksiklik barındırdığını göstermektedir.**

***Anahtar Kelimeler*—Görüntü-Dil-Aksiyon modelleri, çok dilli değerlendirme, robotik manipülasyon, OpenVLA, çapraz dil analizi.**

***Abstract*—Vision-Language-Action (VLA) models unify visual perception, language comprehension, and action generation under a single architecture, yet they are trained almost exclusively on English data. We evaluated the OpenVLA-7B model by issuing the same 50 manipulation commands in both Turkish and English and comparing the resulting action vectors. The paired measurements revealed substantial cross-lingual divergence: the mean L2 error reached 0.501 and cosine similarity dropped to 0.372, both statistically significant. No linear relationship between command complexity and performance degradation was observed. Although Turkish commands incurred a 30.8% token overhead, no meaningful latency difference was recorded. Our findings highlight a deep-seated deficiency in multilingual action generation within current VLA architectures.**

***Keywords*—Vision-Language-Action models, multilingual evaluation, robotic manipulation, OpenVLA, cross-language analysis.**

I. Giriş

İstanbul'da bir lojistik deposunda çalışan, Türkçe konuşan bir kat amiri tarafından emir alan bir robotu hayal edelim. Amir "kırmızı kutuyu rafa koy" dediğinde, robot kolunun başka bir yerde savrulduğunu ya da tutucusunun havada açıldığı görülmektedir. Bu tablo, çalışmanın çıkış noktasını oluşturmuştur. RT-2 [2] ve OpenVLA [1] gibi VLA modelleri, hâlâ çoğunlukla İngilizce verilerle tasarlanıp test edilmiştir ve uçtan uca robotik denetleyicilerin son sürümü olarak kabul edilmektedir.

Türkçe sondan eklemeli bir dil olduğu için, "koyduramayacaklarmış" gibi tek bir kelime, İngilizcede sekiz ya da dokuz kelimeyle ifade edilen bilgiyi sıkıştırabilir [13]. Bu noktada, OpenVLA'nın Türkçe verilere göre anlam çıkarıp çıkaramayacağı henüz test edilmemiştir. Deney, 50 manipülasyon komutuyla gerçekleştirilmiştir; aynı komutlar İngilizce ve Türkçe olarak, aynı görsel ortamda, ağırlıkları dondurulmuş bir modele sunulmuştur. Türkçe çıktılar, İngilizce çıktılarla niteliksel olarak farklılaşırken, kosinüs benzerliği −0,33 seviyesine kadar düşmüştür.

Dört ana katkı sunulmaktadır: eşleştirilmiş çok dilli manipülasyon ölçüt takımı, L2 ile kosinüs sapma ölçümleri ve istatistiksel açıklaması, karmaşıklığın öngörülemeyeceğinin ispatı ve LLaMA 2 BPE tokenizörünün Türkçe için yüzde 30,8 oranında ek maliyeti içeren tokenizasyon analizi.

II. İlgili Çalışmalar

RT-2 [2], görüntü-dil ön eğitiminin robotik manipülasyona aktarılabilirliğini, SayCan'ın [11] dayandırdığı destekleme yaklaşımını takip ederek ilk kez göstermiştir. OpenVLA [1], bu süreci açık kaynaklı bir yapıda yeniden yapmıştır. SigLIP ve DinoV2 adlı görsel kodlayıcıları, LLaMA 2 ile birleştirerek, Open X-Embodiment [5] veri kümesinde bulunan 970 bin manipülasyon olayından yararlanmıştır. Söz konusu modellerin

paylaştığı kritik eksiklik, hiçbirinin İngilizce dışında bir dilde incelenmemiş olmasıdır.

Görüntü-dil alanındaki VHELM [6], VLM'ler için çok dilli bir karşılaştırma sunarken, LLaVA-OneVision [10] görsel görevler için uzantılar sunmakta, XT-VQA [7] ise İngilizce dışındaki görsel sorularda doğruluk kaybı bildirmektedir. XLM-R [12], çok dilli NLP aktarımını büyük ölçüde geliştirmiş olsa da bu gelişmeler VLA mimarisine aktarılamamıştır.

Tokenizasyon başlı başına bir sorun oluşturmaktadır. Rust vd. [8], İngilizce merkezli BPE tokenizörlerin, Türkçe sözcükleri anlamca daha az zenginleştirilmiş şekilde parçaladığını göstermiştir. Bir Türkçe kelime, zaman, kişi, olumsuzluk ve rivayet gibi dilbilgisi işaretçilerini hepsi birlikte son ek olarak içine alabilir; rastgele yapılan bölünmeler bu anlamlı bilgileri ortadan kaldırmaktadır. LoRA [9] ile ince ayar bu sorunu hafifletebilir, ancak İngilizce olmayan VLA bağlamında bu konu henüz araştırılmamıştır.

III. Yöntem

A. Model ve Eylem Uzayı

Çalışmamızda, anlambilimsel ve mekânsal temellendirme için sırasıyla SigLIP ve DinoV2 çift kodlayıcısını, öğrenilmiş bir projeksiyon katmanını ve LLaMA 2 7B [4] dil omurgasını içeren OpenVLA-7B modeli [1] kullanılmıştır. Eylemler, LLaMA 2 tokenizörünün düşük frekanslı sözcük sembollerini yeniden düzenleyerek her boyut için 256 kutuya bölünmektedir. İleri geçiş, yedi serbestlik derecesini içermektedir: x, y, z, döndürme, eğim, kayma ve tutucu durumu. Tutucu dışındaki tüm boyutlar süreklidir; tutucu ise iki değer alabilmektedir. Model, BitsAndBytes ile 4-bit NF4 formatına dönüştürülmüş ve Modal platformunda bulunan 15,6 GB VRAM kapasiteli T4 GPU üzerinde çalıştırılmıştır.

B. Veri Seti ve Çeviri

BridgeData V2 [3] kaynaklarından 20 al-ve-yerleştir, 15 itme ve 15 yönlendirme olmak üzere toplam 50 farklı manipülasyon komutu derlenmiştir. Bu komutlar üç karmaşıklık düzeyine ayrılmıştır. Basit görevler 18 isimfiil çiftinden oluşmaktadır (örneğin "Kırmızı küpü kaldır"). Orta seviye görevler 16 adet isim, fiil ve mekânsal bilgi içermektedir (örneğin, "Bardağı tabağa koy"). Zor görevler en az iki mekânsal ya da ilişkisel kelime içeren 16 isim-fiil çiftinden oluşmaktadır (örneğin "En yakın nesneyi al ve ortaya yerleştir").

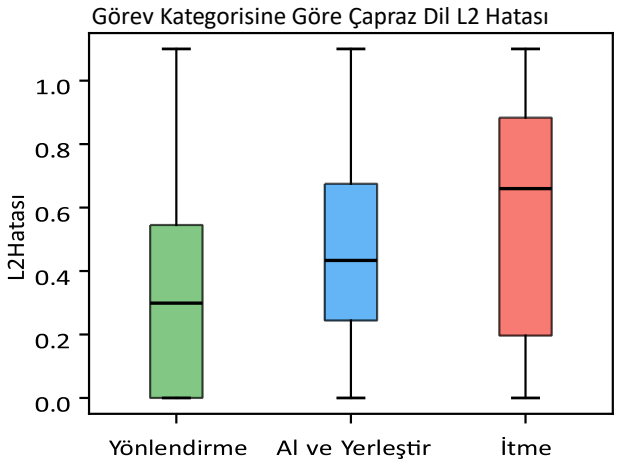
Çeviri, Türkçe anlayışlı ve İngilizceyi çok iyi bilen iki araştırmacı tarafından gerçekleştirilmiştir. Makine çevirisi araçlarından yararlanılmamıştır; çünkü 3B uzayda edat ve sonek farklarının manipülasyon görevlerinde ne kadar etkili olduğu bilinmektedir. Örneğin, "üstüne" ve "yanına" gibi ifadeler, masa üstü manipülasyonda 20 cm'lik bir konum farkına karşılık gelebilir.

C. Deney Düzeni

Her İngilizce ve Türkçe komut çifti, aynı sabit sentetik görüntüye (224×224) eşleştirilmiştir:

Tablo I: Görev Kategorisine Göre Çapraz Dil Sapması

Kategori	N	$N_{göz}$	L2 Hatası	Kos. Ben.
Yönlendirme	15	45	0,404±0,449	0,622±0,481
Al ve Yerleştir	20	60	0,488±0,468	0,244±0,525
İtme	15	45	0,614±0,459	0,294±0,490
Toplam	**50**	**150**	**0,501±0,464**	**0,372±0,528**



Şekil 1: Görev kategorisine göre L2 hata dağılımı.

$$A_{EN} = f_\theta(\text{img}, \text{prompt}_{EN}), \quad A_{TR} = f_\theta(\text{img}, \text{prompt}_{TR}) \quad (1)$$

Burada f_θ, donmuş ağırlıklara sahip OpenVLA-7B modelidir. Deneyini 0,3 sıcaklık değerinde üç defa tekrar ettiğimizde, toplamda 50 görev × 2 dil × 3 tekrar = 300 çıkarım elde etmek mümkün oldu. Her dil için üç tekrarın ortalaması hesaplanmadan önce eşleştirilmiş metrikler alınmıştır.

İstatistiksel analizde, L2 hatalarını karşılaştırmak için tek örneklem t-testi ve Wilcoxon işaretli sıralar testi kullanılmıştır. Görev kategorilerini karşılaştırmak için tek yönlü ANOVA ve karmaşıklık ile hata arasındaki ilişkiyi incelemek için Spearman sıra korelasyonu uygulanmıştır.

IV. Sonuçlar

A. Genel Çapraz Dil Sapması

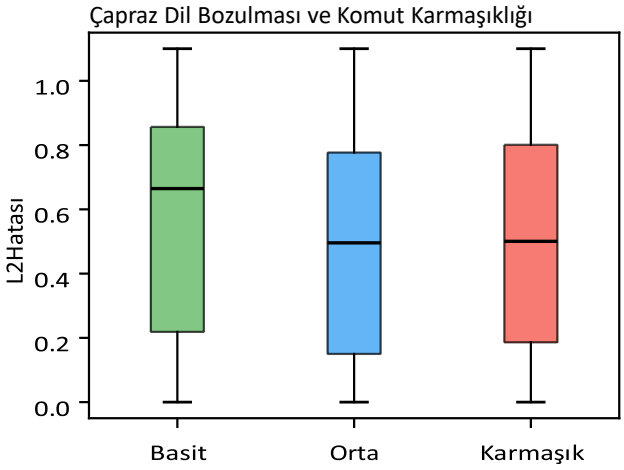
Tablo I'de gösterildiği üzere, Türkçe komutlarla yapılan eylem vektörleri, İngilizce karşılıklarına göre çok fazla farklılık göstermiştir. 50 eşleştirilmiş L2 hatası üzerinde tek örneklem t-testi uygulandığında $t(49) = 9{,}60$ ($p < 0{,}001$, Cohen's $d = 1{,}36$) sonucu bulunmuştur; Wilcoxon testi de $W = 0{,}0$ ($p < 0{,}001$) ile bu sonucu desteklemiştir.

İkili bir Eylem Başarı Oranı (EBO) tanımlanmıştır; bu oran, L2 hatasının 0,99 eşiğini aştığı görevleri başarısız saymaktadır. Bu eşik, pratikte tutucunun ters yönde hareket etmesi durumunu ifade eder. 50 görevin 36 tanesi bu kriteri karşılamıştır ve EBO

yüzde 72 olarak bulunmuştur. Dikkat çeken, 0,501 olan L2 hatasının yanı sıra ortalama kosinüs benzerliğinin sadece 0,372'ye düşmesi ve bazı görevlerde −0,33'e kadar inmesidir. Bu durum, eylem vektörünün sadece zayıflamadığını, aksine ters yöne döndüğünü göstermektedir.

Tablo II: Komut Karmaşıklığına Göre L2 Hatası

Karmaşıklık	$N_{görev}$	$N_{göz}$	L2 Hatası
Basit	18	54	0,551±0,471
Orta	16	48	0,462±0,462
Karmaşık	16	48	0,482±0,462



Şekil 2: L2 hatasının komut karmaşıklığına göre dağılımı (1 =basit, 2 =orta, 3 =karmaşık). Korelasyon gözlemlenmemiştir (ρ = 0,002).

Yüksek L2 hatası ve yüksek kosinüs benzerliği birlikte görülürse kazanç ölçekleme sorunu söz konusu olurdu; ancak burada yönlerdeki temsilin bozulduğuna dair bir durum söz konusudur.

B. Kategori ve Karmaşıklık Analizi

En yüksek L2 hatasını 0,614 ortalama ile "itme" kategorisi göstermiştir; bunu 0,488 ile "al-ve-yerleştir" ve 0,404 ile "yönlendirme" izledi. Ancak ANOVA, kategoriler arasında L2 hatalarında istatistiksel olarak anlamlı bir fark bulamadı ($F(2,147) = 2,37$, $p = 0,097$). Yönlendirme kategorisindeki görece düşük hata oranının, "döndür" gibi dönüş fiillerinin, çok dilli bir web korpusunda "ittir" gibi özel kelimelerle karıştırılarak daha sık karşılaşıldığından dolayı olduğu değerlendirilmektedir.

Çok uzun cümlelerin kısa cümlelere göre daha kötü sonuç vereceği varsayılmıştı. Bu varsayım bütünüyle çürütüldü. Tablo II'de görüldüğü üzere L2 hatası, basit talimatlar için 0,551, orta düzey talimatlar için 0,462 ve karmaşık talimatlar için 0,482 olarak bulunmuştur. Spearman korelasyonu $\rho = 0,002$ ($p = 0,981$) sonucu elde edildi ve bu değer sıfıra çok yakın olarak bulunmuştur.

C. Tokenizasyon ve Gecikme

LLaMA 2'nin BPE tokenizasyonu ile Türkçe komutlar ortalama 28,7 tokene ayrılmakta iken, İngilizce komutlar için bu sayı 21,9 olarak gerçekleşmektedir; dolayısıyla Türkçe talimatlar ortalamada yüzde 30,8 daha fazla token gerektirmektedir. "Kırmızı küpü kaldır" ifadesi 8 tokene bölünürken, İngilizce karşılığı 6 token olarak kalmaktadır. Ancak iki dili de kullanarak yapılan çıkarım süresi neredeyse

Tablo III: Tokenizasyon ve Gecikme Karşılaştırması

Metrik	İngilizce	Türkçe
Ort. token/komut	21,9	28,7
Token ek yükü	—	+%30,8
Ort. gecikme (ms)	810	808
Gecikme farkı	%0,2	—

aynıdır: İngilizce için 810 ms, Türkçe için 808 ms ve yüzde −0,2'lik bir fark bulunmuştur. Bu durum, 224×224 boyutundaki görüntüde SigLIP ve DinoV2 kodlayıcılarının ileri geçiş süresini bastırmak suretiyle açıklanmaktadır.

V. Tartışma

A. Başarısızlığın Doğası

Başarısızlığın doğası, performansta yumuşak bir aşınma değil, bir boşluk saptadığımızdır. L2 hatası 0,99'u aşan 14 görevi tek tek inceledikten sonra, her görevde tutucu boyutunun önemli bir etkisi olduğu fark edilmiştir. Tutucu ikilik bir boyuttur (0 kapat, 1 aç) ve tek bir uyumsuzluk, L2 hatasını 1,0'a yaklaştırmaya yetebilir. Gerçek hayatta bu, robota "kutuyu kaldır" komutu verildiğinde tutucunun kapanmaması, nesnenin yerden düşmesi ya da daha tehlikeli bir durumda kolun çalışma alanına çarpması anlamına gelir. Bu, manipülasyon görevlerinde görülebilecek en tehlikeli hata türüdür.

B. Karmaşıklığın Neden Etkisiz Olduğu

İlginç bulgu, talimatın zorluk seviyesi ile modelin performansına zarar vermesi arasında hiçbir bağlantı bulunmamasıdır. "Küpü kaldır" gibi basit bir komutla, çok sayıda bileşene sahip bir talimat aynı oranda etkilenmektedir. Sebep, LLaMA 2'nin "Kaldır" kelimesini tokenlere ayırmak istediğinde onu "Kal" ve "dır" olarak bölmektedir. "Kal" kelimesi Türkçede "beklemek" manasına gelmektedir; bu yüzden tokenizör, sadece zayıf bir ipucu değil, tamamen yanlış bir ipucu üretmektedir. Rust vd.'nin [8] bildirdiği, iyi temsil edilmeyen diller için ortaya çıkan bu sorun, bulgularımızla örtüşmektedir.

C. Pratik Çıkarımlar

İngilizce dışı bir VLA modelinin sahada konuşlandırılması için farklı alternatifler düşünülebilir. En doğrudan yol, Türkçe'den İngilizce'ye gerçek zamanlı bir makine çevirisi katmanı eklemektir. Günümüzdeki çeviri sistemleri genelde kısa ve sade emir cümlelerini iyi anlar; ancak edatlarda bulunan son eklerin farklılığı, ciddi konumlama hatasına neden olabilir. "Üstüne koy" ve "içine koy" arasındaki fark, sözcüğün son

eklerine bağlıdır ve yanlış çeviri, üç boyutlu uzayda büyük hata yapmaya neden olabilir.

Daha iyi bir çözüm, Türkçe fiil-çatı çiftlerine LoRA [9] temelli parametre verimli ayarlar yapmaktır. Bu yöntem, modelin Türkçe alt kelimelerinin doğrudan daha iyi hale gelmesine yardımcı olmaktadır.

D. Kısıtlamalar

Çalışmamız sadece bir dil çiftini içerdiği ve gerçek zamanlı video yerine sentetik görüntülerle yapıldığı için sınırlıdır. Başarısızlık örüntüleri; Çince gibi yalıtımlı, Arapça gibi çekimli ya da İnuitçe gibi çok sentezlemeli dillerde oldukça farklı şekillerde olabilir. NF4 ile 4-bit nicemlemenin dil token dağılımı arasında oluşturduğu karmaşık etkileşimler, kesinlik sınamalarını etkileyebilir. Sonunda, tüm komut çiftleri için aynı görsel girdi kullanılarak dil değişkeni ayrıştırılmış olsa da, görsel girdinin karmaşıklığının sonuçlar üzerindeki etkisi hâlâ araştırılmamıştır.

VI. Sonuç

OpenVLA-7B modeline Türkçe girdi verildiğinde performansı hafifçe azalmakla kalmayıp ciddi şekilde düşmektedir. 50 görev çifti ve 300 çıkarım üzerinde yapılan ortalama L2 hatası $0{,}501 \pm 0{,}464$ seviyesine çıkmış, kosinüs benzerliği de 0,372'ye düşmüştür. Bu sonuçlar, parametrik ve parametrik olmayan tüm testlerde $p < 0{,}001$ olarak anlamlıdır ve Cohen's d değeri 1,36'dır. Talimat karmaşıklığı ile bozulmanın arasında bir ilişki bulunmaması ($\rho = 0{,}002$), sorunun söz dizimi çözümlemesinden ziyade token düzeyinde temsil ettiği düşüncesiyle uyumludur.

LLaMA 2'nin BPE tokenizörünün yüzde 30,8'lik ek gecikmesi, zaman kaybı olarak önemsiz olmakla birlikte, Türkçe gibi bazı kelimelerin baştan itibaren yanlış şekilde temsil edildiği bir dilde ciddi anlam kaybına neden olmaktadır. 50 görevden 14'ünde tutucu durumu tersine dönmüş durumda olup, bu durum gerçek donanımda düşme ve çarpışma riski doğurmaktadır. Türkiye'de ve İngilizce konuşulmayan diğer bölgelerdeki robotik uygulamalarda, sistemi yerel dillerde çalışır hale getirmek yalnızca bir tercih değil, bir zorunluluktur.

Kaynaklar

[1] M. J. Kim ve diğ., "OpenVLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[2] A. Brohan ve diğ., "RT-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[3] H. Walke ve diğ., "BridgeData V2: A dataset for robot learning at scale," *arXiv preprint arXiv:2308.12952*, 2023.

[4] H. Touvron ve diğ., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[5] A. Padalkar ve diğ., "Open X-Embodiment: Robotic learning datasets and RT-X models," *arXiv preprint arXiv:2310.08864*, 2023.

[6] H. Lee ve diğ., "VHELM: A holistic evaluation of vision language models," *NeurIPS*, 2024.

[7] S. Changpinyo ve diğ., "Cross-lingual text-rich visual comprehension," *AAAI*, 2025.

[8] P. Rust ve diğ., "How good is your tokenizer? On the monolingual performance of multilingual language models," *ACL*, 2021.

[9] E. J. Hu ve diğ., "LoRA: Low-rank adaptation of large language models," *ICLR*, 2022.

[10] D. Zhu ve diğ., "LLaVA-OneVision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[11] M. Ahn ve diğ., "Do as I can, not as I say: Grounding language in robotic affordances," *CoRL*, 2022.

[12] A. Conneau ve diğ., "Unsupervised cross-lingual representation learning at scale," *ACL*, 2020.

[13] K. Oflazer, "Turkish and its challenges for language processing," *Language Resources and Evaluation*, c. 48, no. 4, ss. 531–556, 2014.