

İleri Görev Gösterimlerinden Görsel-Dil Modelleri Aracılığıyla Ters Planlama Backwards Planning from Onward Task Demonstrations via Vision-Language Models

Ahmet Fırat Gamsız*, Deniz Bilge Akkoç*, Yiğit Yıldırım ve Emre Uğur

Bilgisayar Mühendisliği

Boğaziçi Üniversitesi

İstanbul, Türkiye

{ahmet.gamsiz, deniz.akkoc}@std.bogazici.edu.tr, {yigit.yildirim, emre.ugur}@bogazici.edu.tr

*Bu yazarlar eşit katkıda bulunmuştur.

Özetçe—Bu makalede, görsel-dil modellerini (VLM) kullanarak geriye dönük planlama alanında yenilikçi bir yöntem önermekteyiz. Geri planlama üzerine önceki çalışmalar, manipülasyon görevlerinin anlamsal boyutunu göz ardı eden geleneksel yöntemleri uygulamıştır. Önerdiğimiz sistem, ileri görev yürütmelerini analiz ederek geriye dönük planları çıkarsamak için VLM’lerin anlamsal kavrama ve fiziksel muhakeme yeteneklerinden yararlanır. Yöntemimiz, bu modellerin kendi başlarına kullanımını keşfetmekte ve yaygın kapalı kaynak VLM’lerin planlama yeteneklerini karşılaştıran kapsamlı bir ablasyon çalışması sunmaktadır. Sistemimiz iki robotik manipülasyon görevinde %80 başarı oranına ulaşmaktadır. Ayrıca, son teknoloji VLM’lerin, görsel anlamada önemli ölçüde zorlandığı gözlemlenmiştir. Bu kısıtlama genelgeçer yürütme için hala harici somutlaştırmalara ihtiyaç duyulduğunu gösterirken, gözlemlenen planlama yetenekleri, etkili geri planlamanın karmaşık mimarilere ihtiyaç duymayabileceğini göstermektedir.

Anahtar Kelimeler—Robotik, Görev Planlama, Görsel-Dil Modelleri, Gösterimden Öğrenme

Abstract—In this paper, we propose a novel method of backward planning using visual-language models (VLM). Previous work on backward planning applied traditional methods that ignore the semantic meaning of manipulation tasks. Our proposed framework utilizes VLMs’ semantic understanding and physical reasoning capabilities to infer backward plans by analyzing onward task executions. Our method explores the barebone usage of these models and provides a comprehensive ablation study comparing the planning capabilities of common closed-source VLMs. We demonstrate that our system reaches an 80% success rate in two robotic manipulation tasks. We also observe that several state-of-the-art VLMs struggle significantly in visual understanding. This limitation still necessitates external embodiments for robust execution. However, the observed planning capabilities suggest that effective backward planning may not require highly complex architectures.

Keywords—Robotics, Task Planning, Vision-Language Models, Learning from Demonstration

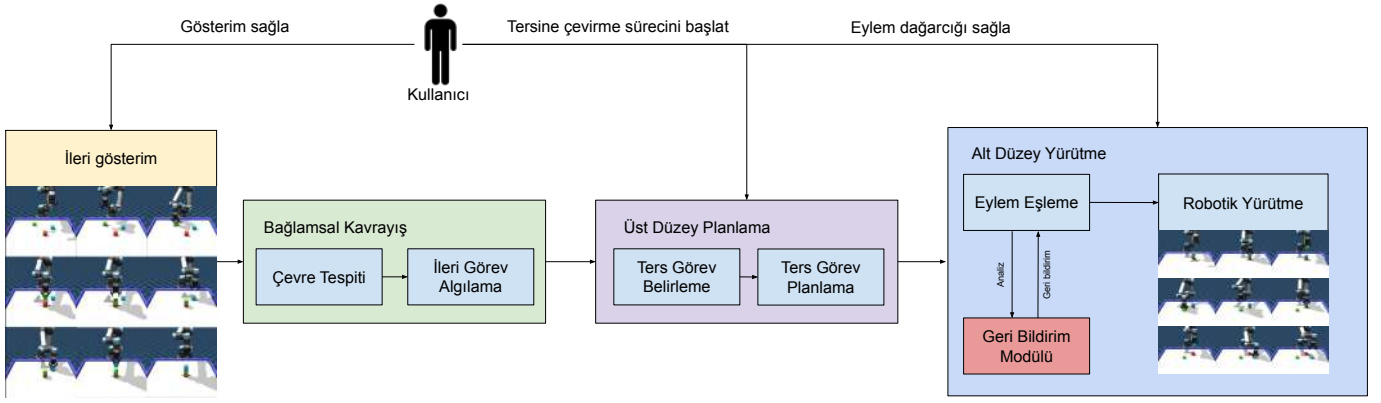
I. GİRİŞ

Planlama, klasik sembolik planlamadan modern öğrenme tabanlı yöntemlere kadar uzanan yaklaşımlarıyla, robotikte

temel bir uğraşı olmuştur. Geleneksel planlama sistemleri genellikle önceden tanımlanmış sembolik gösterimlere ve durum-eylem uzaylarına [1] [2] dayanır; bu da durum gösterimlerinin karmaşık ve belirsiz olduğu gerçek dünya senaryolarında zayıf kalabilmektedir [3]. Öğrenme tabanlı planlamadaki son gelişmeler bu karmaşıklığı ele almada umut verici olmuştur [4]–[6], ancak genellikle kapsamlı eğitim verisi gerektirirler [7] ve yeni senaryolara genelleme konusunda zorlanırlar [8].

İleri görev planlaması kapsamlı bir şekilde çalışılmış olsa da, ters planlama – bir dizi eylemi geri almak için gereken adımları belirleme yeteneği – daha az ilgi görmüştür. Ters planlama, montaj adımlarının geri alınması gerekebileceği üretim süreçleri veya nesnelerin orijinal konumlarına geri döndürülmesi gereken ev ortamları gibi birçok pratik senaryoda kritik öneme sahiptir. Ayrıca ters plan üretme yeteneği, bir sistemin istenmeyen bir durumdan geri dönmesi gereken hata kurtarma durumları için de oldukça değerlidir. Robotlar için zorlu montaj görevlerindeki başarının, görece daha basit demontaj süreçlerinin keşfedilip tersine çevrilmesiyle artırılabilir [9]. Direkt terse çevirme, yalnızca hareket yörüngesi seviyesinde direkt olarak terse çevrilebilen görevler için kullanılabilen bir yaklaşım olup hedeflenen becerinin tersine dair anlamsal muhakemeden yoksundur. Bu yaklaşım, tersine planların sağlayacağı faydaların yalnızca belirli bir alt kümesini hedeflemektedir. Öte yandan, ters görevlerdeki zamanı tersine çevirerek Pekiştirmeli Öğrenme (PÖ) modelleri eğitmek, ileri ve ters görevler arasındaki bağlantıdan daha derinlemesine faydalanmayı amaçlayan farklı bir yaklaşımdır [10]. PÖ kullanımı ise, akılcıca bir kapsamlı arama yaparak çözüm üretmekte ve çok kıymetli bağlam bilgisinden faydalanmamaktadır [11]. Bu durum, öğrenme tabanlı sistemlerden beklenen genelleme kapasitesinin önünde bir engel oluşturur.

Büyük dil modellerinin (LLM’ler) ortaya çıkışı, robotik planlamada yeni ufuklar açmıştır. Bu modeller, bağlam bilgilerini ve görev tanımlarını anlama [12], sıralı planlar oluşturma ve eylemler ile sonuçları hakkında akıl yürütme [13] konularında dikkat çekici yetenekler sergilemiştir. Son çalışmalar, LLM’lerin karmaşık görevleri, uygulanabilir adımlara etkili bir şekilde ayırabildiğini [14] ve fiziksel kısıtlamaları ve



Şekil 1: Önerilen sistemin mimarisi. Kullanıcının ileri gösterim sağlamasıyla başlayan sistem bağlamsal kavrayış modülünün çevre ve görevi algılamasıyla devam eder. Daha sonra üst düzey planlama modeline sağlanan bu bilgiler kullanıcının süreci başlatmasıyla ters görev ve plana dönüştürülür. Devamında alt düzey eylemlere eşlenen plan, robot üzerinde yürütülür.

çevresel koşulları dikkate alan planlar üretebildiğini [15], [16] göstermiştir.

Bu makale, görsel-dil modellerini kullanarak ters planlar üretirken anlamsal bağlamdan faydalanmayı amaçlayan bir yaklaşımı sunmaktadır. Yaklaşımımız, görev yürütme dizilerini analiz etmek, çevresel bağlamı anlamak ve uygun ters planlar üretmek için görsel-dil modellerinin güçlü akıl yürütme yeteneklerinden yararlanır. Sistemimizin kilit özelliği orijinal eylem dizisini basitçe tersine çevirmekten daha verimli olabilecek alternatif ters planlar üretme yeteneğidir; bu da çeşitli senaryolarda optimal görev tamamlamayı mümkün kılar. Son olarak, çeşitli kapalı kaynak modeller kullanılarak önerilen yöntemin efektifliği ve farklı dil modellerinin planlama kapasiteleri kapsamlı bir şekilde incelenmiştir. Bu çalışmanın ana katkıları şunlardır:

- 1) Üst düzey planları anlama ve oluşturma yeteneğini gösteren, görsel-dil modellerine dayanan ters planlama alanında özgün bir metot.
- 2) Yaklaşımın, farklı büyük dil modelleri üzerindeki, ters planlama görevlerindeki görece güçlü yanlarını ve kısıtlamalarını ortaya koyan kapsamlı bir değerlendirmesi.

II. YÖNTEM

Yöntemimizi açıklamaya ileri ve ters görevleri tanımlayarak başlıyoruz. İleri görev, bir başlangıç durumunun bir dizi eylem gerçekleştirilerek istenen bir sonuca dönüştürüldüğü bir süreci ifade eder. Bu, bir görevin doğal veya amaçlanan ilerleyişini temsil eder ve genellikle belirli bir hedef gözetilerek yürütülür. Ters görev ise ileri görevin tersine çevrilmesidir ve orijinal durumu geri yüklemeyi amaçlar. İleri görevin etkilerini geri almak için gerekli hareketlerin veya manipülasyonların belirlenmesini içerir. Bazı görevler yörüngeyi tersine çevirip yürüterek doğrudan tersine çevrilebilirken, çoğu özel bir planlama gerektirir. Yöntemimiz, Şekil 1'de gösterildiği şekilde, verilen bir robotik veya insan yörünge yürütme dizisinden ileri görevi çıkarsamayı, ters görevi belirlemeyi ve üst düzey bir plan oluşturmayı içerir. Daha sonra bu üst düzey plan, eylem repertuarımızdaki robotik eylemlerle eşlenir ve robot tarafından yürütülür. Bu modülleri görsel-dil modellerini kullanarak gerçekleştiriyoruz. Düşünce zinciri [17] sorgulama

yöntemine benzer şekilde, modellerimizi önce verilen soru üzerinde detaylı bir analiz oluşturmaları ve ardından cevabı sağlamaları için koşullandırıyoruz.

A. Bağlamsal Kavrayış

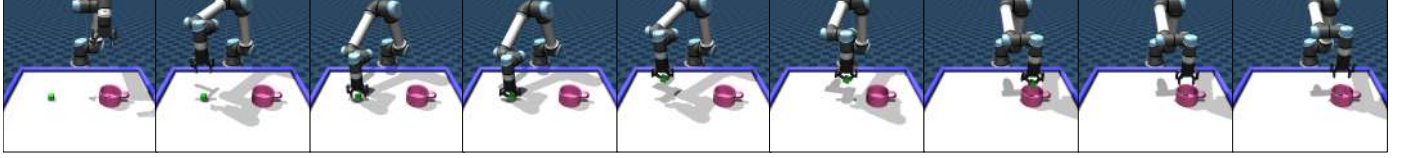
Öncelikle sistemimiz, verilen ileri gösterimi inceleyerek "Robot hangi ortamda?" sorusuna cevap oluşturur. Bu soru için yapılan analiz, ortamın ayrıntılı bir şekilde tanımlanmasını ve önemli nesnelerin bir listesini çıkartmayı içerir. Bu nesne listesi kaydedilir ve daha sonra eylem eşleme için kullanılır. Yörünge ve ortam analizi üzerine koşullandırılarak, modelden ileri görevi tespit etmesi istenir. Buradaki analiz, karelerdeki önemli eylemlerin veya alt görevlerin listelenmesini içerir.

B. Üst Düzey Planlama

Bu modülden, daha önce üretilen bilgileri ve yörünge karelerini bağlam olarak kullanarak bir ters görev tanımlaması istenir. Modelden, başlangıç ve son karelerdeki çevre ve nesne durumlarını analiz ederek bir ters görev tanımlaması istenir. Bu ters görev, üst düzey bir plan oluşturmakla görevlendirilmiş olan bir sonraki modüle beslenir. Etmene mümkün olduğunca detaylı, adım adım bir plan üretmesi için yönlendirme yapılır. Ayrıca, çevre veya nesneler hakkında görüntüden çıkarılabilecek varsayımlar yapmaması açıkça belirtilir. Bu, birçok büyük dil modeli için sorun teşkil eden halüsinasyonları azaltmak ve planın uygulanabilir olmasını sağlamak için önemlidir. Model her adım için gerekçelendirme sağlar ve adımın yürütülmesinden sonraki beklenen durumu tahmin eder. Modeli genel hedefe odaklı tutmak için "gerekçelendirme" ve adımın neden olabileceği fiziksel kısıtlamaların farkında olmasını sağlamak için "beklenen durum bilgisi" istenir. Bu modül genellikle ileri yörüngeyi yaklaşık olarak tam tersini üretmeye eğilimlidir. Ancak birçok durumda, bir ters görevin daha basit bir tamamlanma yolu vardır. Farklı planları keşfetmek için, daha önce oluşturulan planı bağlama ekleyerek üst düzey planlama modülünü yinelenir ve modelden yaratıcı yaklaşım kullanarak adımları en aza indirecek alternatif bir plan oluşturmaları istenir. Bu prosedürün alternatif ve daha yaratıcı çözümler üretmesi beklenir.



Şekil 2: Kule görevi ileri yürütme kareleri.



Şekil 3: Kupa görevi ileri yürütme kareleri.

C. Alt Düzey Yürütme

Bu modülde, modelden önceden üretilen üst düzey planları robotun önceden tanımlanmış eylem repertuvarıyla eşleştirmesi istenir. Sistem, beş temel manipülasyon eylemini desteklemektedir: tutma, yerleştirme, itme, döndürme ve hareket ettirme. Her eylem yapılandırılmış bir formatta tanımlanır ve nesne ve konum bilgilerini içerir. Konum bilgileri iki farklı örnek kümesi kullanılarak belirtilir: tutuş pozisyonu örnekleri (üstten, yandan, arkadan) ve konumsal ilişki örnekleri (üstünde, yanında, içinde, önünde, arkasında). Sistem, her plan adımı için eylem eşleştirmesi üretmeden önce analiz bir gerçekleştirir. Bu analiz, manipülasyonun fiziksel gereksinimlerini ve çevresel kısıtlamaları değerlendirir. Eşleştirme işleminin çıktısı uygulanabilirlik açısından geri bildirim modülü tarafından değerlendirilir. Onaylanan planlar önceden tanımlanmış parametrik beceriler ile eşleştirilerek robota gönderilir. Obje konumunu parametre olarak verildiğinde ters kinematik çözücü ile hesaplanan eylem yörüngesi yürütmeye konulur.

III. DENEYLER

Deneyler iki manipülasyon görevi için UR-5 robot kolu kullanılarak MuJoCo simülöründe gerçekleştirilmiştir. İlk görev olan kule görevi Şekil 2’de gösterilmiştir. İleri görev kırmızı, mavi ve yeşil renkli blokların üst üste dizilerek kule oluşturulması, tersi ise inşa edilen kulenin olarak yıkılmasıdır. İkinci görev olan kupa görevi Şekil 3’te gösterilmektedir. Yeşil bloğun pembe kupanın içerisine yerleştirilmesi istenen bu görevin tersini ise bahsedilen bloğun kupadan çıkarılıp masaya yerleştirilmesi oluşturmaktadır.

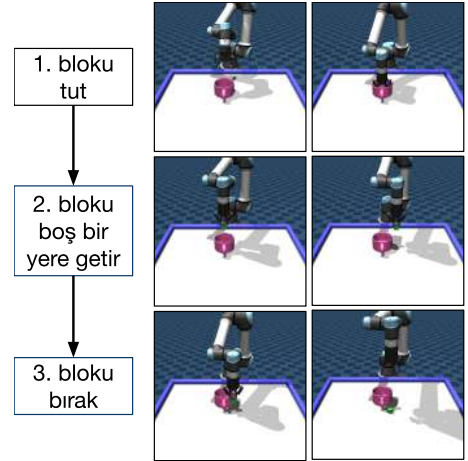
Nesneler her görev için rastgele konumlandırılmış olup modellerin genelgeçerliğinden emin olmak amaçlanarak model ve görev başına onar adet deney yürütülmüştür. İleri görevlerin görsel gösterimleri, modele, 3×3 ’lük kare ızgara şeklinde girdi olarak sunulmuş olup her kare yürütme süresince eşit aralıklarla seçilmiştir. Bu gösterimler 400×400 piksel olacak şekilde küçültülerek hem görsel bilgiyi korumak hem de hesaplama verimliliğini sürdürmek amaçlanmıştır.

Sistemin temel uygulaması Gemini 2.0 Flash kullanılarak yapılmış olup Ters Planlayıcı Gemini (TP_{gemini}) olarak anılacaktır. Karşılaştırmalı ablasyon çalışmaları, sistemin GPT-4o (TP_{gpt}) ve Claude 3.5 Sonnet (TP_{sonnet}) kullanılan varyasyonlarında gerçekleştirilmiştir. Performans değerlendirmesi, anlamsal tutarlılık ve doğruluk baz alınarak, model temellendirilmelerinin ve yanıtlarının el ile değerlendirilmesiyle yapıl-

mıştır. Bir planın başarılı olup olmadığına simülörün son durumunun her görev için beklenen yapılandırmayla eşleşip eşleşmediğine bakılarak karar verilmiştir.



Şekil 4: Kule görevi alternatif geri yürütme kareleri.



Şekil 5: Kupa görevi geri yürütme kareleri.

IV. DENEY SONUÇLARI

Deney sonuçları, Tablo I’de her modülün farklı varyasyon ve görevde başarılı olma miktarını ve eylem eşlemede ortalama deneme sayısını gösterecek şekilde listelenmiştir. Tablo I’da görüldüğü üzere TP_{gemini} hem kule hem de kupa görevlerinde %80 başarı oranıyla öne çıkarken, TP_{gpt} sırasıyla %20 ve %70 başarı elde etmiştir. TP_{sonnet} ise %20 ve %10’luk başarı oranlarıyla daha geride kalmıştır.

Görsel-dil modelleri için görsel kavramanın önemli bir sorun teşkil ettiği gözlenmiştir. Hem obje listeleme hem de görev tespiti aşamalarında farklı başarı oranlarına sahip olan modellerden TP_{gemini} dikkat çekici bir performansla öne

Model	Görev	Çevre Tespiti	Obje Listeleme	Görev Tespiti	Ters Görev Belirleme	Ters Görev Planlama	Eylem Eşleme	Alternatif Plan Üretme	Alternatif Eylem Eşleme	Robotik Yürütme	Alternatif Robotik Yürütme
TP _{gpt}	Kule	10	10	8	5	2	1	0	-	2	0
	Kupa	10	7	7	7	7	1,8	1	3	7	0
TP _{sonnet}	Kule	10	10	7	7	2	1	5	1,2	2	5
	Kupa	10	1	1	1	1	1	0	-	1	0
TP _{gemini}	Kule	10	10	10	10	8	1	7	1,2	8	7
	Kupa	10	10	9	9	9	2	7	2	8	6

TABLO I: Modüllerin model ve görevlere göre başarılı olma miktarları ve eylem eşlemede ortalama deneme sayıları.

çıkılmaktadır. Özellikle karmaşık görsel sahneleri yorumlama ve nesneler arasındaki ilişkileri doğru tespit etme konusunda üstün bir yetenek sergilemiştir. Modelin bu başarısı, sadece nesneleri tanımakla kalmayıp, onların uzamsal ilişkilerini ve fiziksel özelliklerini de doğru bir şekilde kavrayabilmesinden kaynaklanmaktadır. TP_{gpt} ve TP_{sonnet} ise renk ayırt etme ve derinlik algılama gibi temel zorluklar ile karşılaşmıştır. Her iki modelin de kule yapma görevini düzlemsel değerlendirerek masanın üstünde düzenleme olarak algıladığı gözlemlenmektedir.

TP_{sonnet}, kule oluşturma görevinde standart yaklaşımlarda zorlanmasına rağmen, alternatif stratejilerde nesneleri yandan itmek gibi yaratıcı çözümler geliştirerek bazı modellere göre daha fazla ters görev üretebilmiştir. Bu sayede standart yaklaşımda başarı oranı %20 iken yaratıcı yaklaşımda bu oran %50'ye yükselmiştir. Bu yaratıcı yaklaşımlar sayesinde modellerin, görsel anlamlandırmadaki eksiklikleri alternatif çözüm yolları geliştirerek aşabileceği gözlemlenmiştir.

Model	Kule	Kupa	Kule Alternatif	Kupa Alternatif
TP _{gpt}	5	3,3	-	-
TP _{sonnet}	12	4	5	-
TP _{gemini}	5,9	4,9	1,3	4,7

TABLO II: Ortalama plan uzunlukları.

Tablo II'de gösterilen ortalama plan uzunlukları incelendiğinde alternatif yaklaşımın daha az adımda çözüme ulaştığı gözlemlenmektedir. Kule görevinde TP_{gpt}'ün planları 4-6 adım, TP_{sonnet}'in standart çözümleri 12 adım gerektirirken, TP_{gemini}'in alternatif çözümleri görevi sadece 1-2 adımda başarmıştır. Kupa görevinde ise standart ve alternatif planların uzunluğunda keskin bir fark olmasa da alternatif planların gerçek robot deneylerinde yardımcı olacağı gözlemlenmiştir. Örnek olarak robotik tutucu aparatının boyutunun kupanın içine giremeyecek kadar büyük olması gibi durumlarda, alternatif planlar sadece modelin kendi eksikliklerini değil çevresel kısıtlamaların da üstesinden gelmek için faydalı olmaktadır.

V. SONUÇ

Bu çalışmada görsel-dil modelleri ile ileri görev gösterimleri üzerinden ters görev planları yapabilen bir sistem önerilmiştir. Bu sistem kıyas testlerinde yüksek performans gösterirken, farklı dil modelleriyle olan varyasyonları üzerinden kabiliyet ve sınırları keşfedilmiştir. TP_{gemini} hem görsel kavramada hem de ters görev planlamada diğer modellere göre üstün performans sergilemiştir. TP_{gpt} kupa görevinde iyi performans gösterirken, derinlik algılamadaki limitasyonlar nedeniyle kule görevinde zorlanmıştır. TP_{sonnet}, benzer görsel kavrama zorluklarıyla karşılaşmasına rağmen, kule görevinde nesneleri yandan itmek gibi yaratıcı alternatif çözümler üreterek başarı oranını önemli ölçüde artırmıştır. Gelecek çalışmalar kapsamında, elde

edilen sonuçların kullanılan algoritmaların teknik yetenekleri ile ilişkili olarak tartışılması ve geliştirilen yaklaşımın farklı yöntemlerle karşılaştırılması planlanmaktadır.

BİLGİLENDİRME

Bu araştırma, AB'nin finanse ettiği INVERSE projesi (No: 101136067) ve TÜBİTAK ARDEB 1001 programı (No: 124E227) kapsamında desteklenmiştir.

KAYNAKLAR

- [1] R. E. Fikes and N. J. Nilsson, "Strips: A new approach to the application of theorem proving to problem solving," *Artificial intelligence*, vol. 2, no. 3-4, pp. 189-208, 1971.
- [2] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100-107, 1968.
- [3] J. Blythe, *An overview of planning under uncertainty*. Springer, 2001, pp. 85-110.
- [4] C. Bradley, A. Pacheck, G. J. Stein, S. Castro, H. Kress-Gazit, and N. Roy, "Learning and planning for temporally extended tasks in unknown environments," in *ICRA*. IEEE, 2021, pp. 4830-4836.
- [5] A. Ahmetoglu, B. Celik, E. Oztup, and E. Ugur, "Discovering predictive relational object symbols with symbolic attentive layers," *RA-L*, 2024.
- [6] A. Ahmetoglu, E. Oztup, and E. Ugur, "Symbolic manipulation planning with discovered object and relational predicates," *RA-L*, 2025.
- [7] L. Chen, L. Platinsky, S. Speichert, B. Osiński, O. Scheel, Y. Ye, H. Grimmert, L. Del Pero, and P. Ondruska, "What data do we need for training an av motion planner?" in *ICRA*. IEEE, 2021, pp. 1066-1072.
- [8] Y. Li, A. Debnath, G. J. Stein, and J. Košecká, "Learning-augmented model-based planning for visual exploration," in *IROS*. IEEE, 2023, pp. 5165-5171.
- [9] N. Widulle and O. Niggemann, "Using reverse reinforcement learning for assembly tasks," in *PRL Workshop Series - Bridging the Gap Between AI Planning and Reinforcement Learning*, 2023.
- [10] A. D. Edwards, L. Downs, and J. C. Davidson, "Forward-backward reinforcement learning," *arXiv preprint arXiv:1803.10227*, 2018.
- [11] S. Nair, M. Babaeizadeh, C. Finn, S. Levine, and V. Kumar, "Trass: Time reversal as self-supervision," in *ICRA*, 2020, pp. 115-121.
- [12] J. Wang, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang, Y. Yao, X. Liu, B. Ge, and S. Zhang, "Large language models for robotics: Opportunities, challenges, and perspectives," *Journal of Automation and Intelligence*, 2024.
- [13] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *ICLR*, 2023.
- [14] Z. Yang, C. R. Garrett, D. Fox, T. Lozano-Pérez, and L. P. Kaelbling, "Guiding long-horizon task and motion planning with vision language models," in *2nd CoRL Workshop on Learning Effective Abstractions for Planning*, 2024.
- [15] Z. Wu, Z. Wang, X. Xu, J. Lu, and H. Yan, "Embodied task planning with large language models," *arXiv preprint arXiv:2307.01848*, 2023.
- [16] H. Aktas and E. Ugur, "Vq-cnmp: Neuro-symbolic skill learning for bi-level planning," *arXiv preprint arXiv:2410.10045*, 2024.
- [17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *NeurIPS*, vol. 35, pp. 24 824-24 837, 2022.