SpatialBoost: Enhancing Visual Representation through Language-Guided Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite the remarkable performance of large-scale pre-trained image representation models (i.e., vision encoders) across various vision tasks, they often fail to learn spatial relationships within images, constraining their effectiveness in various downstream tasks, e.g., visual spatial reasoning and vision-based robot control, etc. This limitation stems from the scarcity of 3D or multi-view images, making it challenging to inject 3D spatial knowledge into the encoders. To overcome this limitation, we propose a novel learning framework that enhances spatial awareness in existing pre-trained image representation models. The core idea involves converting 3D spatial information into linguistic expressions, which is then used to inject such spatial knowledge into vision encoders through a Large Vision Language Model (LVLM). To further improve spatial awareness, we introduce a multi-turn visual spatial reasoning approach; specifically, we adopt a Chain-of-Thought (CoT) framework to build hierarchical spatial understanding through 10 sequential reasoning turns. The proposed approach enhances pre-trained vision encoders, for example, improving average accuracy on the SpatialRGPT visual language spatial reasoning benchmark from 13.3% to 52.0% simply by replacing the vision encoder in LLaVA-1.5-7B.

1 Introduction

Pre-trained image representation models [23, 14, 7, 31, 3, 16] have shown remarkable success in various downstream tasks, such as image classification [11, 30], semantic segmentation [59, 32], monocular depth prediction [44, 20], and vision-language understanding [2, 26]. The core idea behind these successes is extracting transferrable representation from large-scale image datasets such as ImageNet [13], enabling the model to understand semantic information within images that are significantly useful for various downstream tasks.

Despite their success, these models are predominantly trained on 2D images and hence face a fundamental challenge in acquiring 3D spatial awareness capabilities. Large vision language models struggle to discern 3D spatial relationships between objects in images [33, 19, 47, 8], and demonstrate sub-optimal performance in vision-based robotic control tasks compared to approaches that directly utilize 3D information [54, 28, 56]. Training visual models on multi-view images can encode spatial information [55, 48, 5], however, broader applicability is constrained by the need to use carefully curated data [53] or simulation environments [43]. These challenges indicate the need for methodologies that encode spatial information while leveraging widely available 2D image datasets.

We introduce **SpatialBoost**, a learning framework to enhance the spatial understanding of existing pre-trained vision encoders by using language-guided reasoning (see Figure 1). The key idea is to transform geometric and semantic information within images into language descriptions and then use them to enhance the visual encoder via language supervision. This injection of linguistic knowledge

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

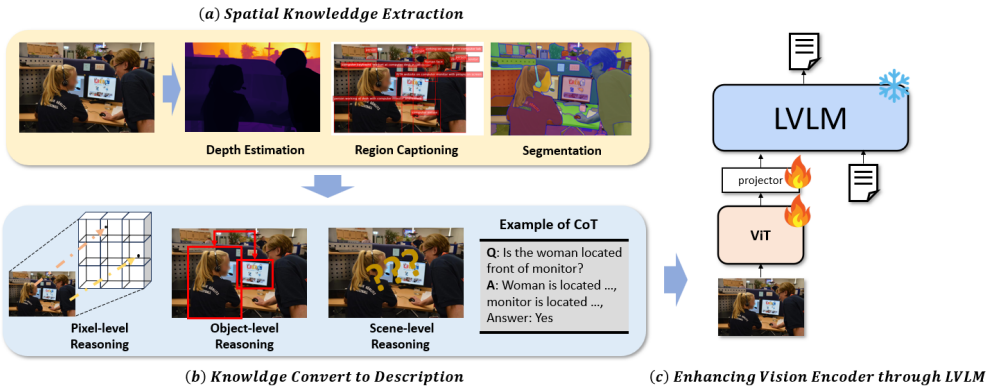

Figure 1: **Overview of the proposed framework**, which enhances spatial and geometric understanding of pre-trained vision encoders by leveraging language-guided spatial reasoning. Our framework, *i.e.*, SpatialBoost, consists of (a) spatial knowledge extraction, (b) converting extracted knowledge into multi-turn spatial reasoning from pixel to scene levels, and (c) building a spatial-aware vision encoder using LVLM.

utilizes an LVLM like LLaVA [33], with all parameters frozen except for a trainable dual-channel attention layer added to the image encoder. We also adopt a multi-turn visual spatial reasoning approach using a Chain-of-Thought (CoT) framework to build hierarchical spatial understanding through 10 sequential reasoning turns.

We apply SpatialBoost to DINOv2 [38] and OpenCLIP [9]. LLaVA-1.5-7B [34] trained with SpatialBoost OpenCLIP and DINOv2 improved from 13.3% to 52.0% and 18.8% to 54.2%, respectively, surpassing GPT-4o [1] (39.7%) and Gemini 2.0-flash [12] (42.5%) simply by changing the encoder. In embodied environments, average score increases by 6.0% for OpenCLIP ViT-L/14 and 7.2% for DINOv2 ViT-L/14. We also show applicability to depth estimation and semantic segmentation, where ViT variants achieve performance comparable to significantly larger models.

2 Method

In this section, we present SpatialBoost, a framework that leverages linguistic expressions of geometric and semantic information within images to enhance pre-trained vision encoders for spatial understanding. We describe how we leverage an LVLM and use dual-channel attention layers to inject linguistic information into image representations, and how we extract 3D spatial information from 2D images and express it in language via a multi-turn visual spatial reasoning dataset (see Figure 1).

2.1 Preliminary: LLaVA

LLaVA [33, 34] is an LVLM designed to generate natural language responses to questions about visual inputs. Given an image $\mathbf{x}$ and QA pairs $(Q_\mathbf{x}, A_\mathbf{x})$, LLaVA extracts visual feature vectors using a pre-trained visual encoder f_V, projects them via g_P to obtain $\mathbf{v}_\mathbf{x} = g_P(f_V(\mathbf{x}))$, and processes textual embeddings and $\mathbf{v}_\mathbf{x}$ through the LLM decoder h_L to predict $A_\mathbf{x}$ auto-regressively.

2.2 Training Strategy for SpatialBoost

We train the model to generate answers containing the knowledge that we aim to inject by taking images and prompts as input to the LVLM. This process employs supervised fine-tuning (SFT) loss while keeping all hyperparameters of the LLM component fixed, allowing only the vision encoder and projector parameters to be trainable. Through this method, the vision encoder learns to generate representations necessary for producing answers. However, this process risks losing useful image representations previously possessed by the vision encoder. To address this challenge, we implement a dual-channel attention mechanism (see Figure 3 and Appendix B).

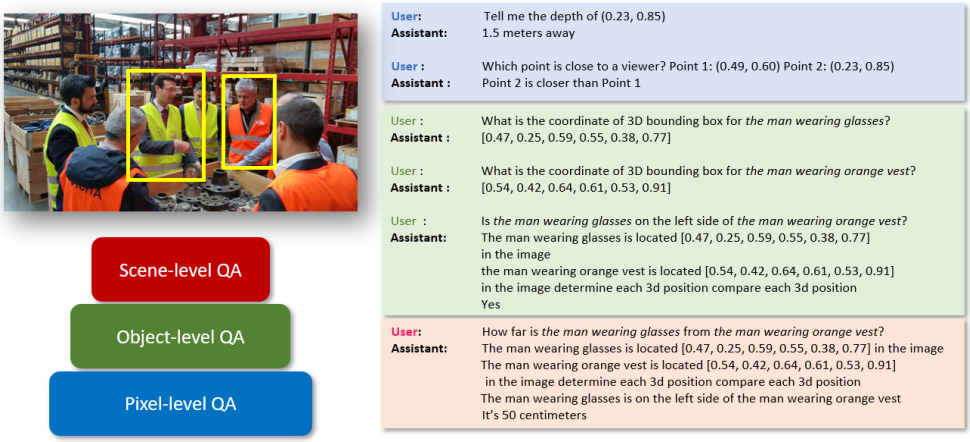

Figure 2: **Illustration of multi-turn visual spatial reasoning dataset**, exhibiting pixel-level, object-level, and scene-level reasoning QAs. At the pixel-level, the QA task queries the 3D positions of points (*e.g.*, via depth estimation). At the object-level, it extracts spatial properties of objects (*e.g.*, by predicting bounding cubes or relative positions). At the scene-level, it determines the exact distances between multiple objects that require the rationales of the previous steps. These are listed in order and constitute 10 multi-turn visual spatial reasoning conservation.

2.3 Enhancing Vision Encoder with Spatial CoT

We now describe our process for extracting 3D spatial information from 2D images and converting it into linguistic expressions. Our key idea is to utilize VQA data rich in spatial relationships that can be effectively processed by LVLMs, enabling us to optimize vision encoders for enhanced spatial understanding. To achieve this, we introduce a multi-turn visual spatial reasoning approach implementing a Chain-of-Thought (CoT) framework [49]. This method builds hierarchical spatial understanding through 10 sequential reasoning turns that progress from narrow to broad view. Each new turn generates reasoning that depends on previous answers, enabling the model to develop progressively deeper spatial comprehension. By fine-tuning vision encoders with this dataset, we effectively transfer spatial knowledge into image representations.

Multi-turn Visual Spatial Reasoning Dataset. We construct a multi-turn visual spatial reasoning dataset, *i.e.*, multi-turn question-answering (QA) dataset specialized in spatial reasoning. Given an image $\mathbf{x}$, we use depth estimation and segmentation models [4, 42] to extract a 3D point cloud and then synthesize QA pairs specialized in spatial reasoning. We construct QA pairs at three levels: **pixel**-level (*e.g.*, depth prediction or comparison), **object**-level (*e.g.*, semantic spatial information via 3D bounding cubes or relative positional relations), and **scene**-level (*e.g.*, exact distances between multiple objects). This hierarchy enables CoT reasoning in order of pixel-, object-, and scene-level.

Multi-turn Fine-tuning. We fine-tune the vision encoders by presenting spatial reasoning as a multi-turn conversation. For each image, we format the 10 QA turns into a single chat template where each turn builds upon previous turns. The complete conversation is then used for supervised fine-tuning, enabling the model to learn the connected nature of spatial reasoning through the entire sequence at once. This approach allows the LVLM to reason at broader view levels based on information obtained from narrower views (see the reasoning process in Figure 2).

3 Experiments

In this section, we design experiments to investigate whether SpatialBoost can effectively enhance visual representations by capturing geometric and semantic information within images. In particular, we evaluate SpatialBoost on VQA tasks that require 3D geometric spatial reasoning, vision-based robot learning tasks, and dense prediction tasks (see Section 3.1). We also provide ablation studies and analysis on our design choices (see Section 3.2). Details of each experiment are described in Appendix C.

Table 1: **Results on visual question answering (VQA) tasks.** We report the accuracy (%) of large vision-language models (LVLM) with various vision encoders on general VQA tasks and spatial reasoning from SpatialRGPT-Bench [8] and BLINK's Relative Depth Benchmark (BLINK-bench). We use ViT-L/14 model for both OpenCLIP [9] and DINOv2 [38].

		Spatial Reasoning		General VQA			
Model	Vision encoder	SpatialRGPT-bench [8]	BLINK-bench [19]	VQAv2 [22]	GQA [26]	MMBench [36]	MME [18]
GPT-4o [1]	-	39.7	64.5	-	-	-	-
Gemini-2.0-flash [12]	-	42.5	68.3	-	-	-	-
LLaVA-1.5-7B [34]	OpenCLIP [9]	13.3	51.6	77.8	61.8	63.9	1510.2
	+ SpatialBoost	**52.0**	**85.1**	**79.0**	**65.6**	**67.7**	**1516.3**
	DINOv2 [38]	18.8	55.2	75.2	61.5	64.0	1506.2
	+ SpatialBoost	**54.2**	**87.5**	**76.8**	**62.5**	**67.4**	**1514.2**

Table 2: **Results on vision-based robot learning.** We report imitation learning agents on 4 domains from CortexBench [37], trained upon frozen representations.

	Robot learning				
Method	Adroit	MetaWorld	DMControl	Trifinger	Avg.
OpenCLIP [9]	50.8 $\pm$ 3.3	75.7 $\pm$ 1.9	58.9 $\pm$ 2.0	64.8 $\pm$ 0.7	62.6
+ SpatialBoost	**53.8** $\pm$ 3.7	**84.0** $\pm$ 2.2	**67.9** $\pm$ 1.6	**68.7** $\pm$ 0.4	**68.6**
DINOv2 [38]	36.8 $\pm$ 3.5	62.6 $\pm$ 1.8	48.4 $\pm$ 1.4	62.3 $\pm$ 0.4	52.5
+ SpatialBoost	**50.1** $\pm$ 3.0	**66.5** $\pm$ 2.1	**55.0** $\pm$ 1.9	**67.2** $\pm$ 1.2	**59.7**

Table 3: **Results on dense prediction tasks.** We report RMSE for monocular depth estimation and mIoU for semantic segmentation. All results are linear probing with frozen representations.

	Depth estimation (↓)		Segmentation (↑)	
Method	NYUd [44]	KITTI [21]	ADE20k [58]	Pascal VOC [17]
OpenCLIP [9]	0.56	3.66	39.1	70.8
+ SpatialBoost	**0.40**	**2.82**	**40.0**	**74.3**
DINOv2 [38]	0.38	2.78	47.7	82.1
+ SpatialBoost	**0.32**	**2.56**	**49.2**	**83.5**

3.1 Results on Downstream Tasks

Visual Question-Answering (VQA) Tasks. SpatialBoost consistently enhances spatial reasoning capabilities while preserving general VQA abilities. In Table 1, LLaVA-1.5-7B with SpatialBoost DINOv2 improves average performance across SpatialRGPT-bench from 18.8% to 54.2%, and with OpenCLIP from 13.3% to 52.0%, surpassing GPT-4o (39.7%) and Gemini 2.0-flash (42.5%) simply by changing the encoder.

Vision-based Robot Learning. Across CortexBench domains, agents using SpatialBoost backbones outperform baselines. In Table 2, SpatialBoost OpenCLIP achieves 67.9% vs. 58.9% on DMControl, with average gains of 6.0%p for OpenCLIP and 7.2%p for DINOv2.

Dense Prediction Tasks. SpatialBoost improves both geometric and semantic spatial understanding. For instance, in Table 3, RMSE on NYUd decreases from 0.56 to 0.40 for OpenCLIP and mIoU on ADE20k rises from 47.7% to 49.2% for DINOv2.

3.2 Ablation Study and Analysis

Effect of Multi-turn Visual Reasoning. We investigate the effect of the construction of a multi-turn visual reasoning dataset. Forward curriculum (*i.e.*, pixel $\rightarrow$ object $\rightarrow$ scene) yields larger gains than single-turn, shuffled, or reversed orders, indicating that the order of reasoning has a greater impact than merely using multi-turn data. We provide details in Appendix D.1.

Effect of Dual-channel Attention Layer. We investigate the effect of the dual-channel attention layer, specifically examining whether it preserves original knowledge. Dual-channel attention preserves pre-trained knowledge and improves classification performance on ImageNet-1K [13] and CIFAR-100 [30], while full fine-tuning or LoRA [25] degrades performance. We provide details in Appendix D.2.

Dataset Scalability. We find that increasing the size of dataset consistently improves the performance of monocular depth estimation and semantic segmentation under matched update budgets, demonstrating robustness to scaling and potential for further gains. We provide details in Appendix D.3.

4 Conclusion

In this paper, we have presented SpatialBoost, a framework to enhance the vision encoders by leveraging linguistic expressions of geometric and semantic information within images. SpatialBoost uses an LVLM and dual-channel attention layers, generates a multi-turn visual spatial reasoning dataset, and leverages it to improve image representations.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*, 2015.

[3] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[4] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

[5] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[6] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[7] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020.

[8] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. 2024.

[9] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[10] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[11] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[12] G. DeepMind. Gemini 2.0 model updates: 2.0 flash, flash-lite, pro experimental. https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/, Feb. 2025.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[14] J. Donahue and K. Simonyan. Large scale adversarial representation learning. 2019.

[15] Z. Dong, K. Xu, Y. Yang, H. Bao, W. Xu, and R. W. Lau. Location-aware single image reflection removal. In *IEEE International Conference on Computer Vision*, 2021.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. 2010.

[18] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

[19] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024.

[20] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[21] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 2013.

[22] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[24] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[26] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[27] S. James and A. J. Davison. Q-attention: Enabling efficient learning for vision-based robotic manipulation. *IEEE Robotics and Automation Letters*, 2022.

[28] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *IEEE International Conference on Computer Vision*, 2023.

[30] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[31] T. Li, H. Chang, S. Mishra, H. Zhang, D. Katabi, and D. Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[33] F. Liu, G. Emerson, and N. Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.

[34] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2024.

[35] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[36] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[37] A. Majumdar, K. Yadav, S. Arnaud, J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, T. Wu, J. Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? 2023.

[38] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[39] D. A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. 1988.

[40] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

[41] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *IEEE International Conference on Computer Vision*, 2021.

[42] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[43] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, et al. Habitat: A platform for embodied ai research. In *IEEE International Conference on Computer Vision*, 2019.

[44] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012.

[45] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[46] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. d. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[47] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, S. Li, and N. Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. 2025.

[48] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[49] 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. 2022.

[50] M. Wüthrich, F. Widmaier, F. Grimminger, J. Akpo, S. Joshi, V. Agrawal, B. Hammoud, M. Khadiv, M. Bogdanovic, V. Berenz, et al. Trifinger: An open-source robot for learning dexterity. *arXiv preprint arXiv:2008.03596*, 2020.

[51] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[52] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.

[53] X. Yu, M. Xu, Y. Zhang, H. Liu, C. Ye, Y. Wu, Z. Yan, C. Zhu, Z. Xiong, T. Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[54] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

[55] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.-H. Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

[56] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[57] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. 2023.

[58] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[59] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. 2019.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims in the introduction and abstract accurately reflect the contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

Justification: We did not discuss limitations in this paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not have theory in this paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in Appendix A and Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The benchmark is already open-sourced, but we do not currently submit code and data when submitting.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: : We provide the details in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted with multiple seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide compute resources we used in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not have any ethical concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work has no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our framework does not introduce risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all papers and datasets in Reference.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: We do not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: We do not have human subject.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.