SpatialBoost: Enhancing Visual Representation through Language-Guided Reasoning

Byungwoo Jeon[1⋆], Dongyoung Kim[1,2⋆], Huiwon Jang[1,2], Insoo Kim[1,3], and Jinwoo Shin[1,2]

[1]KAIST [2]RLWRLD [3]NAVER Cloud
{imbw2024,kingdy2002,jinwoos}@kaist.ac.kr

Abstract. Despite the remarkable success of large-scale pre-trained image representation models (i.e., vision encoders) across various vision tasks, they are predominantly trained on 2D image data and therefore often fail to capture 3D spatial relationships between objects and backgrounds in the real world, constraining their effectiveness in many downstream applications. To address this, we propose SpatialBoost, a scalable framework that enhances the spatial awareness of existing pre-trained vision encoders by injecting 3D spatial knowledge expressed in linguistic descriptions. The core idea involves converting dense 3D spatial information from 2D images into linguistic expressions, which is then used to inject such spatial knowledge into vision encoders through a Large Language Model (LLM). To this end, we adopt a multi-turn Chain-of-Thought (CoT) reasoning process that progressively incorporates dense spatial knowledge and builds hierarchical spatial understanding. To validate effectiveness, we adapt SpatialBoost to state-of-the-art vision encoders such as DINOv3, and evaluate its performance gains on a wide range of benchmarks requiring both 3D perception and general vision abilities. For instance, SpatialBoost improves DINOv3 performance from 55.9 to 59.7 mIoU on ADE20K, achieving state-of-the-art performance with 3.8%p gain over the pre-trained DINOv3. Project page.

Keywords: Multimodal Vision Representation · Spatial Reasoning

1 Introduction

Pre-trained image representation models [4,12,21,22,33,43] have shown remarkable success in various downstream tasks, such as image classification [18, 42], semantic segmentation [44, 87], monocular depth prediction [27, 64], and vision-language understanding [2, 37]. The core idea behind these successes is extracting transferrable representation from large-scale image datasets such as ImageNet [20], enabling the model to understand semantic information within images that is 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. Consequently, large vision language models struggle to discern 3D

⋆ Equal contribution.

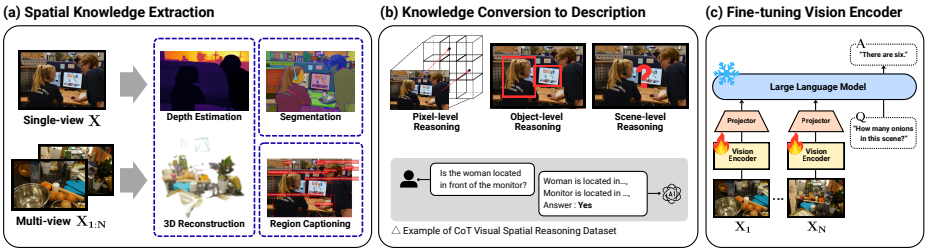


Fig. 1: Overview of SpatialBoost. We enhance spatial and geometric understanding of pre-trained vision encoders by leveraging language-guided spatial reasoning. SpatialBoost consists of (a) spatial knowledge extraction through depth estimation, 3D reconstruction, segmentation, and region captioning, (b) converting spatial knowledge into multi-turn spatial reasoning from pixel to scene levels, and (c) building a spatial-aware vision encoder with LLM using generated data in (b).

spatial relationships between objects in images [16, 25, 45, 71], and demonstrate sub-optimal performance in vision-based robotic control tasks compared to approaches that directly utilize 3D information [40, 81, 85]. To address these limitations, naïve approaches are to train vision models on multi-view images that inherently encode spatial information [10, 72, 83]. While these approaches have shown promise in robot control tasks [61, 62], their broader applicability remains constrained by the need to use carefully curated data [80] or obtain multi-view datasets from simulation environments [60], creating significant limitations for scaling up these approaches. These challenges highlight the need for a novel framework that enables effective learning of 3D information with substantially less data.

To address this problem, we note that vision models specialized for individual tasks are able to infer object positions and point depths from standard 2D images. These extracted cues make it possible to extend spatial information by modeling geometric relationships between objects in a scene. We hypothesize that such spatial information can be systematically converted into explicit representations by leveraging language. Moreover, since language naturally composes information in a sequential and structured form, this property allows the construction of labels that capture dense spatial relationships within a scene.

Importantly, several recent works have shown that language can serve as a scalable supervision signal for learning visual representations [9, 24, 39]. While these approaches primarily focus on semantic understanding or object localization, they highlight the potential of linguistic descriptions as scalable supervision for training vision encoders.

Based on these insights, we introduce SpatialBoost, a training framework that enhances the spatial understanding of pre-trained vision encoders by leveraging language-guided reasoning (see Figure 1). We inject linguistically described spatial knowledge through decoder-based fine-tuning with Large Language Models (LLMs), enabling the framework to process large-scale textual descriptions in a single pass while handling both single-view and multi-view images. In particular, to leverage this knowledge without forgetting the existing knowledge, we

incorporate additional learnable parameters (*i.e.*, dual-channel attention module) into the vision encoder and train only them while freezing the existing parameters. Furthermore, to incorporate dense spatial information in a structured manner, we present a multi-turn visual spatial reasoning approach that builds hierarchical spatial understanding through pixel-level, object-level, and scene-level sub-questions and answers.

To validate the effectiveness of our method, we apply SpatialBoost to state-of-the-art image encoders, including DINOv3 [65] and SigLIPv2 [69], and evaluate them across a diverse set of vision tasks: monocular depth estimation, semantic segmentation, 3D scene understanding, vision-based robotic control, image classification, and image retrieval. Our experiment first shows that SpatialBoost consistently improves performance on tasks requiring 3D spatial knowledge. For example, on the 3D scene understanding task, SpatialBoost improves DINOv3 by 3.5%p (51.4% $\rightarrow$ 54.9%) on the SQA3D task from Lexicon3D Benchmark [50]. In addition, on depth estimation tasks, SpatialBoost improves SigLIPv2 from an RMSE score of 0.51 to 0.39 on NYUd linear probing. Moreover, we show that SpatialBoost even improves the performance of the vision encoders across all benchmarks, notably in image classification: SpatialBoost improves ImageNet linear probing performance of DINOv3 from 88.4% to 90.2%. In addition, DINOv3 with SpatialBoost achieves state-of-the-art performance across all evaluated tasks.

2 Related Work

2.1 Self-supervised Learning for Image Representation

In earlier years, most approaches relied on supervised learning with large-scale labeled datasets to train models [20, 34, 66, 67]. However, the dependence on annotated data introduced scalability challenges due to label expense. To address this, self-supervised learning (SSL) has emerged as a dominant paradigm, leveraging unlabeled data to learn image representations. Contrastive learning methods, including SimCLRv2 [13], MoCov3 [14], DINOv2 [53], and iBOT [88], are trained to distinguish between representations of augmented views of the same image and those of different images. Concurrently, mask prediction approaches such as BEiT [6] and MAE [32], learn representations by reconstructing masked portions of input images. While these methods excel at capturing rich semantic features within 2D images, they lack mechanisms to effectively encode 3D spatial knowledge. On the other hand, we overcome this limitation by enhancing image representations through a novel method that injects 3D spatial knowledge by utilizing language decoding.

2.2 Multi-modal Learning for Image Representation

The increasing prominence of multi-modal tasks has catalyzed the development of vision-language models that jointly represent visual and textual information.

These models typically employ weakly supervised learning by leveraging text caption. Contrastive learning schemes, *e.g.*, CLIP [55], SigLIP [82] and OpenCLIP [17], consist of vision and text encoders and are trained to align their representations in a shared embedding space. Alternative methodologies like M3AE [28], jointly encode image patches and text tokens, employing masked prediction objectives to reconstruct both modalities. More recently, autoregressive formulations such as iGPT [12], have emerged, treating image patches and text tokens as sequential elements for predictive modeling. These approaches successfully enrich visual representations with semantic context derived from natural language descriptions. However, existing models necessitate joint pre-training of both modalities from scratch, imposing significant computational demands and preventing efficient adaptation of existing pre-trained models. Our method eliminates the need for joint text-image representation learning by using LLM, thereby enhancing pre-trained models with relevant linguistic information efficiently.

2.3 Multi-View Learning for Image Representation

Recent advances in vision tasks that require 3D spatial understanding and generation have increased the demand for effective 3D spatial representations [15, 30, 63, 74]. Multi-view images from different camera viewpoints or video sequences serve as input for these tasks. Our focus is specifically on augmenting image representations with useful 3D information. Typically, following approaches similar to single-view image representation learning, multi-view data has been processed by converting images into patches for masked prediction such as MV-MWM [61] or through contrastive learning methods [62]. Additionally, to learn 3D-related information more explicitly, approaches that predict 3D features from image representation [29, 40, 81] have been proposed. These approaches have led to significant performance improvements in vision-based robot control. However, such methods are limited by multi-view data, making it difficult to develop them into pre-trained models for general 3D understanding. Our approach proposes a method to learn 3D spatial representations from both single-view and multi-view images, avoiding these limitations.

3 Method

In this section, we introduce SpatialBoost, a visual representation learning framework designed to improve vision encoders by injecting 3D spatial information expressed in natural language. We first present a multi-modal architecture that incorporates linguistically expressed visual information into the vision encoder through a dual-channel attention layer, ensuring that original visual features are preserved while 3D spatial information is fully exploited (see Section 3.1). On top of this architecture, we design a Visual-Question-Answering (VQA) dataset that hierarchically disentangles 3D spatial relations from both single/multi-view images, enabling the vision encoder to learn spatial information more effectively (see Figure 1 and Section 3.2).

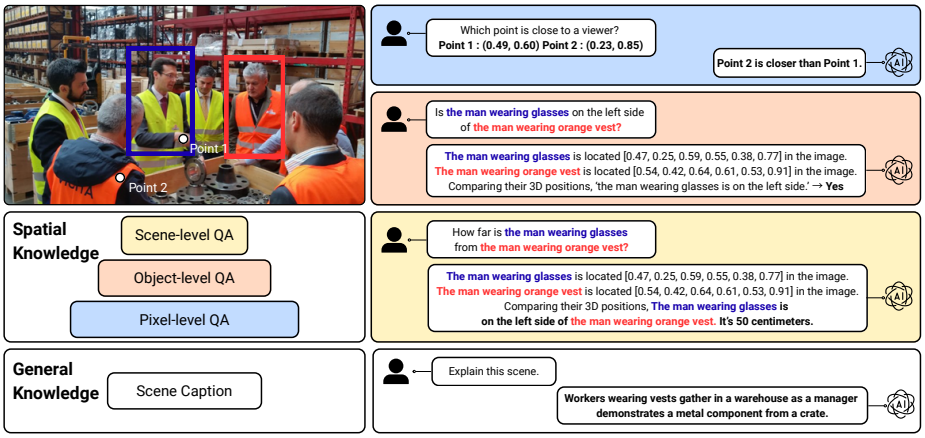


Fig. 2: Illustration of multi-turn visual spatial reasoning dataset, exhibiting pixel-level, object-level, and scene-level reasoning QAs. At the pixel-level, the QA task queries the 3D positions of points (*e.g.*, via depth estimation). At the object-level, it extracts spatial properties of objects (*e.g.*, 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. At last, we add 2-turn for general scene caption. These are listed in order and constitute 12 multi-turn visual spatial reasoning conversation.

3.1 Training Pipeline

To train a vision encoder from rich spatial information encoded in large-scale linguistic expressions, our key idea is to utilize Large-Language Models (LLM) by constructing a multi-modal architecture composed of a vision encoder f_V, a trainable projection module g_P, and the LLM f_L. However, without proper alignment between visual and textual representations, the training signals from the LLM cannot effectively propagate back to the vision encoder, making the learning process ineffective. To fully exploit language supervision, we begin by aligning the visual encoder with the textual embedding space of the LLM. Specifically, we adopt LLaVA [47], a two-stage training for the alignment: feature alignment (Stage 1) and visual instruction tuning (Stage 2). After the alignment, we introduce a training framework that uses a language-guided reasoning dataset to fine-tune the vision encoder (Stage 3). Notably, direct full fine-tuning in this final stage would lead to catastrophic forgetting of the pre-trained knowledge embedded in the vision encoder. To address this challenge, we introduce *dual-channel attention* layers that enable the model to acquire spatial understanding while preserving its original representational capabilities.

Formally, given an input image $\mathbf{x}$ and multi-turn conversation data $(\mathbf{x}_\mathrm{q}^1, \mathbf{x}_\mathrm{a}^1, \cdots, \mathbf{x}_\mathrm{q}^T, \mathbf{x}_\mathrm{a}^T)$ from question-answering (QA) pairs $(Q_\mathbf{x}, A_\mathbf{x})$, we first encode $\mathbf{x}$ to obtain visual features $\mathbf{z}_\mathrm{v} = f_V(\mathbf{x})$, which are mapped into the token embedding space via $g_P(\mathbf{z}_\mathrm{v})$. These visual tokens are then concatenated with text tokens and fed into the LLM. Given the multi-turn conversation data and input image, we optimize the model through cross-entropy loss. Our training pipeline consists

of three stages and all stages are trained with cross-entropy loss. We describe each stage in the following paragraphs.

Stage 1: Feature alignment. In this stage, we train a projector g_P that maps image features into the textual embedding space of the LLM. This projector pre-training contributes to the stable vision-language alignment. Following the training setup in multi-modal large language models [46, 47], we freeze the parameters of both the visual encoder f_V and the language model f_L, and optimize only the projector g_P.

Stage 2: Visual instruction tuning. Following the projector alignment in Stage 1, this stage extends the alignment to the LLM. We freeze the visual encoder f_V and fine-tune the projector g_P and the language model f_L using our multi-view VQA data, combined with the single-view visual instruction data from LLaVA [47]. This step enables f_L and g_P to handle multi-view visual questions. We provide details of proposed multi-view VQA data in Section 3.2.

Stage 3: Vision encoder fine-tuning with dual-channel attention. Finally, we fine-tune the vision encoder f_V to have the capability of spatial understanding. To effectively inject dense spatial knowledge into the vision encoder, we use multi-turn visual spatial reasoning dataset (see Section 3.2), which is carefully designed for hierarchical spatial reasoning. We train the vision encoder f_V and the projection module g_P while keeping the parameters of the LLM f_L frozen, allowing only the vision encoder to benefit from language-driven spatial information. We employ cross-entropy loss, and through this training process, the vision encoder learns to extract meaningful representations necessary for producing answers. However, direct full fine-tuning risks forgetting of the pre-trained knowledge embedded in the vision encoder. To address this challenge, we introduce a dual-channel attention mechanism (see Figure 3). Specifically, for each attention layer $\texttt{Attn}(\cdot)$ in the visual encoder f_V, we introduce an additional attention layer $\texttt{Attn}^{+}(\cdot)$, whose weight parameters are initialized to the same values as those of $\texttt{Attn}(\cdot)$. Given an input $\mathbf{x}$ to each attention layer, we merge the outputs of $\texttt{Attn}(\cdot)$ and $\texttt{Attn}^{+}(\cdot)$ by introducing a trainable mixture factor $\boldsymbol{\alpha} = \text{sigmoid}(\mathbf{a}) \in (0, 1)^d$ with zero-initialized parameter $\mathbf{a} \in \mathbb{R}^d$, where d is the hidden dimension of $\mathbf{x}$, as follows:

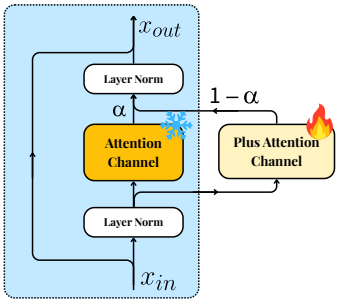


Fig. 3: Illustration of the dual-channel attention layer [35], where an additional attention block is introduced alongside the original attention block and merged via a learnable mixture factor α.

$$\texttt{Attn}^{\text{final}}(\mathbf{x}) = \boldsymbol{\alpha} \cdot \texttt{Attn}(\mathbf{x}) + (1 - \boldsymbol{\alpha}) \cdot \texttt{Attn}^{+}(\mathbf{x}). \tag{1}$$

During fine-tuning, we only update the parameters of $\texttt{Attn}^{+}$ and $\boldsymbol{\alpha}$ while keeping all other parameters frozen. This approach allows the vision encoder to ini-

Table 1: Results on monocular depth estimation from NYUd [64] and KITTI [26] benchmarks. We report the RMSE score between ground truth and predicted depth values. Lower is better. For all results, we freeze the encoder backbone and train a linear head (lin.) or DPT head [57] on top of the image features of the last layer.

	NYUd		KITTI	
Method	lin.	DPT	lin.	DPT
Vision-only trained encoder				
V-JEPAv2 [3]	0.44	0.42	3.07	2.59
Vision-Language trained encoder				
AIMv2 [24]	0.33	0.28	2.56	2.10
`dino.txt` [39]	0.42	0.39	2.98	2.65
TIPS [51]	0.35	0.31	2.58	2.11
PE-Core [9]	0.25	0.23	2.31	2.00
OpenCLIP	0.53	0.41	3.54	2.70
+**SpatialBoost (ours)**	**0.40**	**0.38**	**2.79**	**2.54**
SigLIPv2	0.51	0.40	3.32	2.64
+**SpatialBoost (ours)**	**0.39**	**0.34**	**2.71**	**2.50**
DINOv2	0.37	0.29	2.60	2.11
+**SpatialBoost (ours)**	**0.30**	**0.25**	**2.53**	**2.07**
DINOv3	0.31	0.25	2.33	2.02
+**SpatialBoost (ours)**	**0.25**	**0.21**	**2.20**	**1.84**

Table 2: Results on semantic segmentation from ADE20K [86] and Pascal VOC [23] benchmarks. We report mIoU score. Higher is better. For all results, we freeze the encoder backbone and report results of linear probing (lin.) or multi-scale evaluation (+ms), where the multi-scale approach uses features from the last four layers of the visual encoder to perform segmentation.

	ADE20K		Pascal VOC	
Method	lin.	+ms	lin.	+ms
Vision-only trained encoder				
V-JEPAv2 [3]	51.3	53.8	83.2	86.6
Vision-Language trained encoder				
AIMv2 [24]	31.9	37.9	66.6	72.1
`dino.txt` [39]	50.6	52.8	83.4	86.3
TIPS [51]	49.9	54.1	83.6	86.7
PE-Core [9]	41.5	48.0	73.2	80.5
OpenCLIP	39.5	46.0	71.7	79.3
+**SpatialBoost (ours)**	**40.5**	**47.3**	**75.1**	**80.9**
SigLIPv2	42.8	48.7	72.6	79.1
+**SpatialBoost (ours)**	**45.1**	**50.8**	**79.0**	**82.2**
DINOv2	49.3	53.0	83.0	86.2
+**SpatialBoost (ours)**	**52.0**	**54.9**	**84.5**	**87.6**
DINOv3	55.9	60.3	86.6	89.8
+**SpatialBoost (ours)**	**59.7**	**63.1**	**88.5**	**90.9**

tially rely on pre-trained attention weights and gradually incorporate new attention weights, smoothly enhancing spatial awareness without discarding existing knowledge (see classification result in Figure 6).

3.2 Enhancing Vision Encoder with Spatial CoT

To effectively inject dense spatial information into vision encoders, we address the fundamental limitations of existing spatial datasets. Current spatial VQA data consist of simple single-turn QA pairs with limited information content, insufficient for transferring comprehensive 3D understanding. To overcome this limitation, We introduce Multi-view VQA, which helps align the vision encoder with the LLM to effectively handle multi-view data and a multi-turn Chain-of-Thought (CoT) framework [73] for both single-view and multi-view images that enables the injection of substantially richer spatial information in a single training instance.

Multi-view VQA Dataset. To enhance multi-view VQA capabilities during the visual instruction tuning (Stage 2), we construct multi-view VQA dataset. We first apply LPIPS [84] metric to the 3D or video dataset to obtain a pair of images. Given the pair of images, we employ GPT-4o [1] to generate visual

Table 3: Results on 3D-centric tasks. We evaluate unified probing on diverse 3D-related tasks from ScanNet [19] scenes. We report BLEU-1 score for Vision-Language Reasoning (VLR) on ScanQA [5] and SQA3D [48]. For Visual Grounding (VG), we report accuracy (%) on overall category of ScanRefer [11] dataset. For Geometric Understanding (GU), we report Registration Recall (RR) at 0.05m RMSE threshold and Relative Translation Error (RTE). For 3D Semantic Understanding (3D SU), we report accuracy and mIoU. Lower is better for RTE and higher is better for all other metrics.

	VLR		VG	GU		3D SU	
Method	ScanQA ↑	SQA3D ↑	ScanRefer (%) ↑	RR@0.05m (%) ↑	RTE (m) ↓	Acc ↑	mIoU ↑
Vision-only trained encoder							
V-JEPAv2 [3]	37.7	49.0	55.5	95.4	0.08	69.5	33.2
Vision-Language trained encoder							
AIMv2 [24]	37.4	48.1	50.9	47.8	0.31	53.9	11.8
`dino.txt` [39]	39.8	50.4	52.5	81.7	0.16	85.7	59.4
TIPS [51]	37.4	49.2	52.0	84.0	0.16	71.3	39.7
PE-Core [9]	40.5	51.7	51.7	44.6	0.29	85.0	47.5
OpenCLIP	36.9	48.0	50.1	22.6	0.40	39.8	6.9
+SpatialBoost (ours)	**39.2**	**49.9**	**56.6**	**78.8**	**0.17**	**76.9**	**54.9**
SigLIPv2	38.1	48.5	51.4	47.8	0.28	47.7	9.2
+SpatialBoost (ours)	**40.8**	**50.1**	**56.8**	**86.4**	**0.15**	**81.0**	**55.5**
DINOv2	39.5	49.8	52.7	82.4	0.15	83.0	64.1
+SpatialBoost (ours)	**40.3**	**50.4**	**57.0**	**92.4**	**0.13**	**89.8**	**68.3**
DINOv3	40.6	51.4	56.2	86.9	0.10	91.1	69.1
+SpatialBoost (ours)	**43.3**	**54.9**	**61.1**	**97.5**	**0.06**	**91.9**	**70.6**

questions targeting general multi-view knowledge, which do not require spatial knowledge. We provide more details in Section B.

Multi-turn Visual Spatial Reasoning Dataset. To enhance spatial reasoning capabilities of the vision encoder (Stage 3), we construct multi-turn visual spatial reasoning dataset for single-view and multi-view. Additionally, to enhance general knowledge of the vision encoder, we append GPT-generated scene captions after spatial reasoning turn. For single-view image, we first extract a 3D point cloud from given an image $\mathbf{x}$ by applying diverse vision models (*e.g.*, depth estimation model [8] and image segmentation model [58]). For multi-view images $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, we use 3D reconstruction model [70] to extract a 3D point cloud from given images. Using the point cloud, we synthesize QA pairs specialized in spatial reasoning about $\mathbf{x}$ or $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$.

We then design spatial reasoning QA pairs at three hierarchical levels: pixel, object, and scene, enabling LLM to perform CoT reasoning from narrow to broad view. Specifically, at the pixel-level, the QA task is designed to capture the overall geometry in the image by querying the absolute or relative 3D position of a point, *e.g.*, "What is the depth value at coordinate (x, y)?". At the object-level, the QA task tackles the semantic spatial information of objects inside the image using a bounding cube of the object in 3D space, *e.g.*, "Is [A] on the left side of [B]?", where [A] and [B] is the descriptions about the object in image. We note that this level uses the pixel-level spatial information as a rationale, enabling LLM to reason about the geometry of objects in 3D space. Lastly, at the scene-level, the QA task is designed to predict the exact distance between multiple objects that requires coherent 3D spatial understanding, *e.g.*, "How far is [A] from [B]?".

Table 4: Results on vision-based robot learning. We report the performance of imitation learning agents on 4 domains from CortexBench [49], which are trained upon the image representations. In particular, we report the normalized score for DMControl and success rates (%) for other tasks.

Method	Adroit	MetaWorld	DMControl	Trifinger	Avg.
OpenCLIP	52.6 ± 4.9	83.0 ± 2.7	58.5 ± 1.9	67.7 ± 0.5	65.5
+**SpatialBoost (ours)**	**61.1** ± 3.4	**87.0** ± 3.3	**61.0** ± 1.6	**72.9** ± 0.3	**70.5**
SigLIPv2	56.5 ± 3.0	84.7 ± 2.9	69.4 ± 2.1	68.3 ± 0.8	69.7
+**SpatialBoost (ours)**	**66.5** ± 1.9	**89.1** ± 0.9	**73.5** ± 1.8	**73.9** ± 0.7	**75.8**
DINOv2	55.4 ± 2.7	82.4 ± 4.0	67.9 ± 1.0	66.8 ± 0.2	68.1
+**SpatialBoost (ours)**	**68.1** ± 2.9	**88.5** ± 3.1	**75.0** ± 1.1	**71.4** ± 0.8	**75.8**
DINOv3	63.9 ± 1.5	83.8 ± 1.6	70.8 ± 1.8	72.8 ± 0.5	72.8
+**SpatialBoost (ours)**	**71.8** ± 3.4	**92.0** ± 1.9	**80.4** ± 2.4	**79.0** ± 0.6	**80.8**

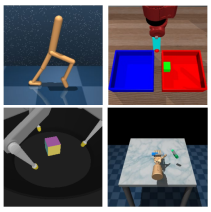


Fig. 4: Examples of visual observations from CortexBench. We train imitation learning agents to learn a mapping from these visual observations to expert actions.

4 Experiments

Through extensive experiments, we validate the performance of SpatialBoost and ablate its key components, focusing on following questions:

- Can SpatialBoost improve spatial knowledge of the vision encoder? (Tables 1 to 4)
- Isn't SpatialBoost overfitted to spatial knowledge? (Table 5)
- Which components contribute to SpatialBoost performance? (Tables 6 to 7 and Figure 6)

4.1 Experimental Setup

VQA Dataset Construction. For single-view image, we use randomly sampled 100K images from the SA1B dataset [41] to construct the single-view VQA dataset specialized in chain-of-thought spatial reasoning. For multi-view images, we use filtered 200K samples from the ego-centric video dataset [31] and 3D dataset [7, 19, 38, 52] to construct multi-view VQA dataset niche in multi-view reasoning or alignment. More details are provided in Section C.

Baselines. For all experiments, we compare our methods with the recent widely-used pre-trained image representation models. To be specific, we first consider OpenCLIP [17] ViT-G/14 and SigLIPv2 [69] ViT-g/16, known for language-aligned vision encoder. We also consider DINOv2 [53] ViT-g/14 and DINOv3 [65] ViT-7B/16, which is a recent state-of-the-art vision encoder. We further include comparable methods, including the vision-only trained encoder like V-JEPAv2 [3], as well as the vision-language trained encoders such as AIMv2 [24], `dino.txt` [39], TIPS [51], and Perception Encoder [9].

Implementation Details. We choose Qwen-2.0-7B [76] as the LLM backbone and 2-layer MLP as the projector, following the architecture of LLaVA-1.5 [46]. Further details are provided in Section A.

Table 5: Results on image classification and retrieval tasks. We report Top-1 accuracy of kNN performance and linear probing (lin.) for image classification on validation set of ImageNet-1K [59]. For image retrieval, we report global average precision (GAP) on Met [78] and mean average precision (mAP) on Oxford-Hard (Oxford-H) [54], Paris-Hard (Paris-H) [54], and AmsterTime dataset [77]. For all results, we freeze the encoder backbone.

	Image classification		Image retrieval			
Method	ImageNet (kNN)	ImageNet (lin.)	Oxford-H	Paris-H	Met (GAP)	AmsterTime
Vision-only trained encoder						
V-JEPAv2 [3]	82.1	84.0	40.6	76.1	38.2	31.9
Vision-Language trained encoder						
AIMv2 [24]	85.7	88.3	60.8	86.4	52.0	50.3
`dino.txt` [39]	80.0	81.4	49.7	78.5	41.8	33.1
TIPS [51]	83.3	86.2	29.0	60.0	20.1	18.5
PE-Core [9]	86.8	89.5	27.9	64.1	15.9	20.0
OpenCLIP	84.0	86.8	23.4	59.7	7.4	24.4
+SpatialBoost (ours)	**86.1**	**87.9**	**32.8**	**69.4**	**19.7**	**30.3**
SigLIPv2	86.3	89.1	25.1	60.9	13.9	15.5
+SpatialBoost (ours)	**87.6**	**90.0**	**36.0**	**69.1**	**24.0**	**27.2**
DINOv2	84.5	87.3	58.2	84.6	44.6	48.9
+SpatialBoost (ours)	**86.4**	**88.6**	**61.3**	**85.2**	**45.1**	**50.8**
DINOv3	85.8	88.4	60.7	87.1	55.4	56.5
+SpatialBoost (ours)	**87.7**	**90.2**	**64.1**	**88.6**	**57.0**	**56.9**

4.2 Dense Prediction Tasks

Setup. We evaluate SpatialBoost on dense prediction tasks requiring geometric and semantic spatial understanding. For geometric understanding, we perform monocular depth estimation on NYUd [64] and KITTI [26] using linear or DPT [57] heads. For semantic understanding, we evaluate on ADE20K [86] and Pascal VOC [23] segmentation benchmarks using linear or multi-scale heads. All experiments freeze the visual backbone during training (see Section A for details).

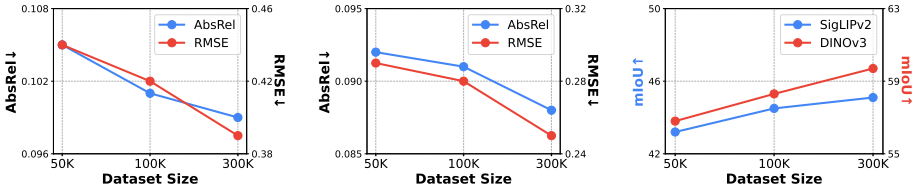
Results. As shown in Table 1 and 2, SpatialBoost consistently improves both geometric and semantic spatial understanding across various encoders. For instance, OpenCLIP's RMSE on NYUd decreases from 0.53 to 0.40 with a linear head, while DINOv3's mIoU on ADE20K increases from 55.9% to 59.7%. These consistent gains demonstrate that language-based spatial knowledge transfer effectively enhances visual encoders' spatial understanding capabilities.

4.3 Complex 3D-centric Tasks

Setup. We evaluate SpatialBoost on Lexicon3D [50], a unified benchmark for 3D scene understanding covering vision-language reasoning, visual grounding, semantic understanding, and geometric understanding. Following Lexicon3D protocols, we freeze visual backbones and train task-specific heads (see Section A for details).

Table 6: Effect of LLM-based fine-tuning. We fine-tune the vision encoder with different headers. We report accuracy (%) for classification (Cls) on ImageNet-1K, mIoU for segmentation (Seg) on ADE20K, RMSE for depth estimation on NYUd, and BLEU-1 score for vision-language reasoning (VLR) on ScanQA. We use ViT-L/14 as the backbone architecture of the encoder.

Method	Cls ↑	Seg ↑	Depth ↓	VLR ↑
DINOv2	86.3	47.7	0.38	39.2
+Linear (depth)	85.7 (-0.70%)	47.9 (+0.42%)	0.35 (-7.89%)	36.9 (-5.87%)
+Linear (seg.)	86.6 (+0.35%)	48.8 (+2.31%)	0.45 (+18.42%)	37.1 (-5.36%)
+SAM decoder	86.3 (+0.0%)	50.1 (+5.03%)	0.42 (+10.53%)	37.6 (-4.08%)
+VGGT decoder	84.8 (-1.74%)	45.6 (-4.40%)	0.35 (-7.89%)	37.3 (-4.85%)
+LLM (ours)	88.3 (+2.32%)	51.5 (+7.97%)	0.32 (-15.79%)	40.0 (+2.04%)



(a) SigLIPv2 depth estimation **(b)** DINOv3 depth estimation **(c)** Semantic segmentation

Fig. 5: Effect of dataset scalability. We investigate the effect of the size of analysis of data scalability effects on (a) depth estimation results (AbsRel, RMSE) on NYUd benchmark for SigLIPv2, (b) depth estimation results (AbsRel, RMSE) on NYUd benchmark for DINOv3, and (c) semantic segmentation results (mIoU) on ADE20K benchmark for SigLIPv2 and DINOv3. The results show scalable performance improvements with increased data size.

Results. As shown in Table 3, SpatialBoost shows comprehensive improvements across diverse 3D tasks. OpenCLIP's BLEU-1 improves from 36.9 to 39.2 on ScanQA [5], while DINOv3 increases from 51.4 to 54.9 on SQA3D [48], demonstrating that SpatialBoost improves spatial understanding without compromising language capabilities. Notably, OpenCLIP's 3D semantic segmentation dramatically improves from 6.9 to 54.9 mIoU, highlighting SpatialBoost can inject robust spatial knowledge into encoders with initially limited spatial awareness.

4.4 Vision-based Robot Learning

Setup. We evaluate SpatialBoost on vision-based robot control using 4 domains from CortexBench [49] spanning locomotion and manipulation tasks [56, 68, 75, 79]. Following CortexBench protocols, we train behavior cloning agents using `[CLS]` representations to predict expert actions from visual observations. We report the mean of best performance across 5 evaluation runs (see Section A for details).

Results. As shown in Table 4, SpatialBoost significantly improves robot task performance across all vision encoders. For example, DINOv2 + SpatialBoost

Table 7: Component-wise analysis. We investigate the effect of multi-turn spatial reasoning data and the effect of single-view and multi-view data. Multi-turn order means the order of three levels (*i.e.*, pixel, object, and scene) in our visual spatial reasoning data. We report accuracy (%) for classification (Cls) on ImageNet-1K, mIoU for segmentation (Seg) on ADE20K, RMSE for depth estimation on NYUd.

Method	Multi-turn order	Single-view data	Multi-view data	Cls ↑	Seg ↑	Depth ↓
DINOv2	✗	-	-	86.3	47.7	0.38
+SpatialBoost	Reverse	+100K	-	87.4	48.4	0.35
	Random	+100K	-	87.4	48.5	0.36
	Forward	+100K	-	87.6	48.9	0.34
	Forward	-	+100K	87.6	48.2	0.36
	Forward	+50K	+50K	87.6	49.2	0.32

achieves 68.1% on Adroit versus 55.4% for DINOv2 alone, demonstrating that enhanced spatial representations directly benefit robot control.

4.5 Image Classification and Retrieval Tasks

Setup. We evaluate SpatialBoost's impact on instance recognition using ImageNet-1K [59] classification and retrieval benchmarks (Oxford, Paris [54], Met [78], AmsterTime [77]). Following DINOv3 protocols, we use linear probing on `[CLS]` representations for classification and similarity-based ranking for retrieval (see Section A for details).

Results. As shown in Table 5, SpatialBoost improves both classification and retrieval despite these tasks not explicitly requiring spatial understanding. DINOv3's ImageNet accuracy increases from 88.4% to 90.2%, while Oxford-Hard mAP improves from 60.7 to 64.1. These results demonstrate that SpatialBoost enhances general vision capabilities without overfitting to spatial features, likely due to our dual-channel attention preserving pre-trained knowledge and the inclusion of general scene captions alongside spatial reasoning.

4.6 Ablation Study and Analysis

Effect of LLM-based Fine-tuning. In Table 6, we investigate whether LLM-based decoders provide superior supervision compared to pixel-level alternatives. We fine-tune the vision encoder with linear layer, SAM [41] decoder, VGGT [70] decoder, and LLM [76]. We then evaluate encoders on ImageNet-1K classification, ADE20K segmentation, and NYUd depth estimation. The results show that LLM consistently outperform pixel-level supervision methods, validating that language provides superior dense information transfer for vision encoders (see Section D for details).

Effect of Multi-turn Visual Reasoning. In Table 7, we investigate how the hierarchical structure of reasoning affects representation learning. We compare dataset construction strategies: (a) shuffled multi-turn, (b) reversed order

Table 8: Comparison with post-training. We fine-tune vision encoders with their original pre-training objectives (simple FT). We report RMSE for monocular depth estimation on NYUd, mIoU for semantic segmentation on ADE20K, BLEU-1 score for vision-language (VL) reasoning on ScanQA, average score for robot learning on CortexBench, and Top-1 accuracy (%) for classification on ImageNet-1K. Lower is better for depth estimation.

Method	Depth Estimation ↓	Segmentation ↑	VL Reasoning ↑	Robot Learning ↑	Classification ↑
OpenCLIP	0.53	39.5	36.9	65.5	84.0
+Simple FT	0.56	39.6	37.7	63.7	84.3
+SpatialBoost (ours)	**0.40**	**40.5**	**39.2**	**72.9**	**86.1**
SigLIPv2	0.51	42.8	38.1	69.7	86.3
+Simple FT	0.53	43.0	38.4	67.9	86.4
+SpatialBoost (ours)	**0.39**	**45.1**	**40.8**	**75.8**	**87.6**
DINOv2	0.37	49.3	39.5	68.1	84.5
+Simple FT	0.36	49.6	39.4	69.4	84.7
+SpatialBoost (ours)	**0.30**	**52.0**	**40.3**	**75.8**	**86.4**
DINOv3	0.31	55.9	40.6	72.8	85.8
+Simple FT	0.31	56.4	40.2	75.5	86.1
+SpatialBoost (ours)	**0.25**	**59.7**	**43.3**	**80.8**	**87.7**

(scene→object→pixel), and (c) forward order (pixel→object→scene). The forward hierarchical ordering shows optimal performance, demonstrating that reasoning order significantly impacts the quality of representation.

Effect of Single-view and Multi-view Data. In Table 7, we investigate the effect of single-view and multi-view reasoning data. With fixed total samples, we compare single-view only, multi-view only, and combined training. While both data types independently improve performance, the combination achieves the highest results, confirming their complementary nature.

Comparison with Naive Post-training. In Table 8, we investigate the effect of post-training. With fixed total samples (*i.e.*, 300K data in multi-turn reasoning data), we compare the naive post-training scheme and SpatialBoost. We evaluate the performance of the vision encoder across five tasks: depth estimation, segmentation, vision-language reasoning, robot learning, and classification. The results show that naive post-training does not yield effective representations for downstream tasks.

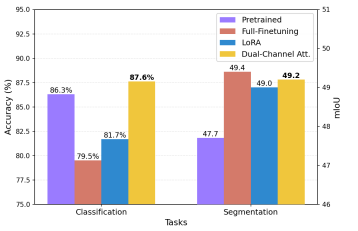


Fig. 6: Effect of dual-channel attention layer. We report the linear evaluation performance of DINOv2-ViT-L/14 across different fine-tuning strategies.

Effect of Dual-channel Attention Layer. In Figure 6, we investigate whether our dual-channel attention mechanism preserves pre-trained knowledge during fine-tuning. We evaluate several approaches for fine-tuning the vision encoder including full fine-tuning, LoRA [36], and dual-channel [35] on ImageNet [59] and ADE20K [86]. Dual-channel attention uniquely preserves and even enhances pre-trained knowledge, while other approaches cause degradation.

Table 9: Application to spatial-aware vision encoder. We apply SpatialBoost on spatial-enhanced vision encoder [9, 51]. We report RMSE for monocular depth estimation on NYUd, mIoU for semantic segmentation on ADE20K, BLEU-1 score for vision-language (VL) reasoning on ScanQA, average score for robot learning on CortexBench, and Top-1 accuracy (%) for classification on ImageNet-1K. Lower is better for depth estimation. Parameters of [9] and [51] are 1.9B and 1.1B, respectively.

Method	Depth Estimation ↓	Segmentation ↑	VL Reasoning ↑	Robot Learning ↑	Classification ↑
TIPS [51]	0.35	49.9	37.4	66.0	83.3
+SpatialBoost (ours)	**0.25**	**53.5**	**41.1**	**76.5**	**85.8**
PE-Core [9]	0.25	41.5	40.5	69.0	86.8
+SpatialBoost (ours)	**0.20**	**50.9**	**44.1**	**81.5**	**88.0**
PE-Spatial [9]	0.26	49.3	40.2	68.7	85.7
+SpatialBoost (ours)	**0.19**	**56.3**	**43.6**	**81.2**	**87.5**

Dataset Scalability. We analyze the impact of dataset sizes on depth estimation results from NYUd [64] benchmark and semantic segmentation results from ADE20K [86] benchmark. As shown in Figure 5, with matched training iterations (*i.e.*, one epoch for 300K data), larger datasets yield consistent improvements, indicating robust scalability potential.

Application to spatial-aware vision encoder. We investigate whether SpatialBoost further improves vision encoders that already possess strong spatial awareness. We evaluate the encoders on depth estimation, segmentation, vision-language reasoning, robot learning, and classification tasks. As shown in Table 9, SpatialBoost consistently improves performance across all benchmarks, demonstrating that our training approach is complementary to existing methods that enhance the spatial awareness of vision encoders.

5 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 LLM and dual-channel attention layers to exploit linguistic information into image representations, generates a multi-turn visual spatial reasoning dataset, and leverages them to improve the image representations. Our experiments show that SpatialBoost consistently enhances the vision encoders on various downstream tasks that require a spatial understanding of images. We hope that our work further facilitates future research on designing and enhancing vision encoders.

Acknowledgments

This work was partly supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST); RS-2022-II220959, Few-shot Learning of Causal Inference in Vision and Language for Decision Making; RS-2025-25442469, Development of Edge AI Server Technology with High-Performance and High-Reliability, 33%), and RLWRLD Inc.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 7
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015) 1
3. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Robert Hogan, F., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025) 7, 8, 9, 10
4. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: CVPR (2023) 1
5. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: CVPR. pp. 19129–19139 (2022) 8, 11
6. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR (2022) 3
7. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR. pp. 5470–5479 (2022) 9
8. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: ICLR (2025) 8
9. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. In: NeurIPS (2025) 2, 7, 8, 9, 10, 14
10. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR (2024) 2
11. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: ECCV. pp. 202–221. Springer (2020) 8
12. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: ICML (2020) 1, 4
13. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS (2020) 3

14. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV (2021) 3
15. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: ECCV (2024) 4
16. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. In: NeurIPS (2024) 2
17. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023) 4, 9
18. Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S.: Large scale fine-grained categorization and domain-specific transfer learning. In: CVPR (2018) 1
19. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017) 8, 9
20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 1, 3
21. Donahue, J., Simonyan, K.: Large scale adversarial representation learning. In: NeurIPS (2019) 1
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 1
23. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010) 7, 10
24. Fini, E., Shukor, M., Li, X., Dufter, P., Klein, M., Haldimann, D., Aitharaju, S., da Costa, V.G.T., Béthune, L., Gan, Z., et al.: Multimodal autoregressive pre-training of large vision encoders. In: CVPR. pp. 9641–9654 (2025) 2, 7, 8, 9, 10
25. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: ECCV (2024) 2
26. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research (2013) 7, 10
27. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012) 1
28. Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., Abbeel, P.: Multimodal masked autoencoders learn transferable representations. In: ICML (2022) 4
29. Gervet, T., Xian, Z., Gkanatsios, N., Fragkiadaki, K.: Act3d: 3d feature field transformers for multi-task robotic manipulation. In: CoRL (2023) 4
30. Goyal, A., Xu, J., Guo, Y., Blukis, V., Chao, Y.W., Fox, D.: Rvt: Robotic view transformer for 3d object manipulation. In: CoRL (2023) 4
31. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022) 9
32. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022) 3
33. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 1
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 3

35. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In: ICLR (2023) 6, 13
36. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022) 13
37. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019) 1
38. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: CVPR. pp. 406–413 (2014) 9
39. Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., Li, D., Szafraniec, M., Ramamonjisoa, M., Oquab, M., et al.: Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In: CVPR. pp. 24905–24916 (2025) 2, 7, 8, 9, 10
40. Ke, T.W., Gkanatsios, N., Fragkiadaki, K.: 3d diffuser actor: Policy diffusion with 3d scene representations. In: CoRL (2024) 2, 4
41. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023) 9, 12
42. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada (2009) 1
43. Li, T., Chang, H., Mishra, S., Zhang, H., Katabi, D., Krishnan, D.: Mage: Masked generative encoder to unify representation learning and image synthesis. In: CVPR (2023) 1
44. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 1
45. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. TACL (2023) 2
46. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR. pp. 26296–26306 (2024) 6, 9
47. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 5, 6
48. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. In: ICLR (2023) 8, 11
49. Majumdar, A., Yadav, K., Arnaud, S., Ma, J., Chen, C., Silwal, S., Jain, A., Berges, V.P., Wu, T., Vakil, J., et al.: Where are we in the search for an artificial visual cortex for embodied intelligence? In: NeurIPS (2023) 9, 11
50. Man, Y., Zheng, S., Bao, Z., Hebert, M., Gui, L., Wang, Y.X.: Lexicon3d: Probing visual foundation models for complex 3d scene understanding. In: NeurIPS (2024) 3, 10
51. Maninis, K.K., Chen, K., Ghosh, S., Karpur, A., Chen, K., Xia, Y., Cao, B., Salz, D., Han, G., Dlabal, J., Gnanapragasam, D., Seyedhosseini, M., Zhou, H., Araujo, A.: TIPS: Text-Image Pretraining with Spatial Awareness. In: ICLR (2025) 7, 8, 9, 10, 14
52. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 9
53. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 3, 9
54. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: CVPR. pp. 5706–5715 (2018) 10, 12

55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 4
56. Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., Levine, S.: Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In: RSS (2018) 11
57. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021) 7, 10
58. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: ICLR (2025) 8
59. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015) 10, 12, 13
60. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: ICCV (2019) 2
61. Seo, Y., Kim, J., James, S., Lee, K., Shin, J., Abbeel, P.: Multi-view masked world models for visual robotic manipulation. In: ICML (2023) 2, 4
62. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: ICRA (2018) 2, 4
63. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: CoRL (2023) 4
64. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 1, 7, 10, 14
65. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 3, 9
66. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 3
67. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015) 3
68. Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D.d.L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al.: Deepmind control suite. arXiv preprint arXiv:1801.00690 (2018) 11
69. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 3, 9
70. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025) 8, 12
71. Wang, J., Ming, Y., Shi, Z., Vineet, V., Wang, X., Li, S., Joshi, N.: Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In: NeurIPS (2025) 2
72. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024) 2

73. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022) 7
74. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: CVPR (2024) 4
75. Wüthrich, M., Widmaier, F., Grimminger, F., Akpo, J., Joshi, S., Agrawal, V., Hammoud, B., Khadiv, M., Bogdanovic, M., Berenz, V., et al.: Trifinger: An open-source robot for learning dexterity. In: CoRL (2020) 11
76. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., Fan, Z.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024), https://arxiv.org/abs/2407.10671 9, 12
77. Yildiz, B., Khademi, S., Siebes, R.M., Van Gemert, J.: Amstertime: A visual place recognition benchmark dataset for severe domain shift. In: ICPR. pp. 2749–2755. IEEE (2022) 10, 12
78. Ypsilantis, N.A., Garcia, N., Han, G., Ibrahimi, S., Van Noord, N., Tolias, G.: The met dataset: Instance-level recognition for artworks. In: NeurIPS (2021) 10, 12
79. Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., Levine, S.: Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: CoRL (2020) 11
80. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., et al.: Mvimgnet: A large-scale dataset of multi-view images. In: CVPR (2023) 2
81. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In: RSS (2024) 2, 4
82. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023) 4
83. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. In: ICLR (2025) 2
84. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018) 7
85. Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C.: 3d-vla: A 3d vision-language-action generative world model. In: ICML (2024) 2
86. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017) 7, 10, 13, 14
87. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2019) 1
88. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: ICLR (2022) 3