VISUAL JIGSAW POST-TRAINING IMPROVES MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning based post-training has recently emerged as a powerful paradigm for enhancing the alignment and reasoning capabilities of multimodal large language models (MLLMs). While *vision-centric* post-training is crucial for enhancing MLLMs' intrinsic understanding of visual signals, current post-training paradigms are predominantly *text-centric*, where dense visual inputs are only leveraged to extract sparse cues for text-based reasoning. There exist a few approaches in this direction, however, they often still rely on text as an intermediate mediator or introduce additional visual generative designs. In this work, we introduce **Visual Jigsaw**, a generic *self-supervised* post-training framework designed to strengthen visual understanding in MLLMs. Visual Jigsaw is formulated as a general ordering task: visual inputs are partitioned, shuffled, and the model must reconstruct the visual information by producing the correct permutation in natural language. This naturally aligns with reinforcement learning from verifiable rewards (RLVR), requires no additional visual generative components, and derives its supervisory signal automatically without any annotations. We instantiate Visual Jigsaw across three visual modalities, including images, videos, and 3D data. Extensive experiments demonstrate substantial improvements in fine-grained perception, temporal reasoning, and 3D spatial understanding. Our findings highlight the potential of self-supervised vision-centric tasks in post-training MLLMs and aim to inspire further research on vision-centric pretext designs.

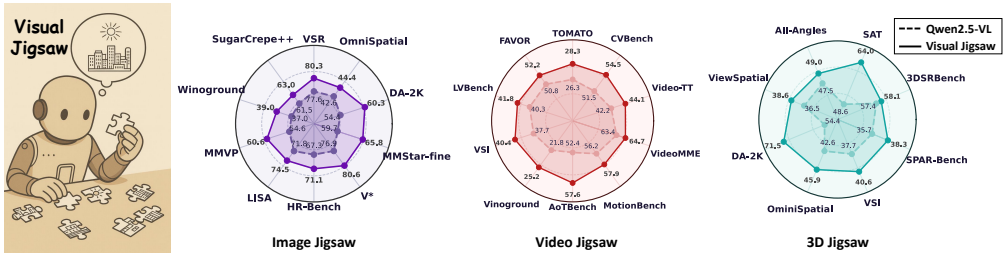


Figure 1: We propose **Visual Jigsaw**, a self-supervised post-training task that enhances visual perception and understanding in MLLMs. Training on visual jigsaw tasks substantially strengthens fine-grained perception, monocular spatial perception, and compositional visual understanding in images; temporal understanding in videos; and geometry-aware understanding in 3D, demonstrating its generality and effectiveness across modalities. For clearer visualization, the value ranges differ across benchmarks in each radar chart.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently demonstrated remarkable progress, achieving strong performance on a wide range of vision–language tasks. Following the success of Reinforcement Learning from Verifiable Reward (RLVR) (Lambert et al., 2024; Guo et al., 2025) in the large language models domain, which has unlocked substantial breakthroughs in complex reasoning abilities, the research community has largely shifted its focus toward replicating this success in the multimodal domain. This has led to a predominant focus on advancing text-based Chain-of-Thought (CoT) multimodal reasoning to enhance multimodal mathematical and scientific reasoning (Huang et al., 2025; Meng et al., 2025; Yuan et al., 2025a; Wang et al., 2025f).

Within this paradigm, dense visual information often serves merely as contextual evidence, from which the model extracts sparse information to support text-based reasoning. Consequently, a deep, fine-grained understanding of the visual signal itself has been considerably undervalued. Some recent studies (Wang et al., 2025b;a) have shown that explicitly incorporating visual reconstruction objectives during the training of MLLMs can improve visual understanding. However, such approaches necessitate the integration of additional visual generation components and learning objectives onto the existing understanding-based MLLM architectures. Furthermore, it remains an open question whether forcing models to achieve pixel-level reconstruction is the optimal strategy for enhancing MLLMs' visual understanding. This raises a pivotal question: Can we enhance an MLLM's visual understanding without altering its architecture or output format?

Delving into the history of self-supervised visual representation learning reveals a rich set of pretext tasks, such as reconstruction-based approaches (He et al., 2022a) and discriminative approaches (He et al., 2020). In parallel, jigsaw-style tasks have emerged as lightweight yet effective paradigms: reordering shuffled image patches (Noroozi & Favaro, 2016), recovering video frame order (Ahsan et al., 2019). While these jigsaw-style approaches provide structural ordering signals, they have generally shown weaker performance compared to more dominant approaches and thus have not become mainstream in vision representation learning. Nevertheless, they demonstrate that the structural ordering jigsaw task, which can be viewed as a simpler version of the reconstruction/generation task, can still offer effective self-supervised signals without requiring pixel-level fidelity.

In this work, we introduce **Visual Jigsaw**, a self-supervised task designed for the RL post-training phase of MLLMs to enhance their visual perception and understanding. The task is formulated as a lightweight ordering problem: visual inputs are partitioned, permuted, and presented to the MLLM, which must then generate the correct permutation order using natural language. Importantly, this formulation requires no additional visual generative designs and is seamlessly compatible with existing MLLMs that produce text-only outputs. Moreover, this task naturally fits in the RLVR framework with deterministic ground-truth and requires no other annotations. We position Visual Jigsaw in the post-training phase, as solving it requires the model to already possess a foundational level of visual understanding. Furthermore, post-training with RL has been shown to offer stronger generalization than Supervised Fine-Tuning (SFT) (Huan et al., 2025; Chu et al., 2025; Chen et al., 2025a), enabling the model to better transfer the vision-centric skills acquired from the jigsaw task to downstream applications.

We implement Visual Jigsaw across three visual modalities: images, videos, and 3D data. Through a post-training phase with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) on visual jigsaw tasks, we substantially improve the ability of MLLMs to perceive and comprehend these visual modalities (shown in Fig 1). In the image domain, we partition the input into patches, shuffle them, and require the model to recover the correct spatial arrangement. We find that this task enhances fine-grained perception, monocular spatial understanding, and compositional visual understanding. For video, we segment the input along the temporal axis, shuffle the clips, and challenge the model to reconstruct the original sequence, leading to marked improvements in temporal understanding. In the 3D domain, we sample points with distinct depth values from an RGB-D image, shuffle and annotate them in the RGB view, and require the model to recover their order from nearest to farthest, thereby augmenting its 3D perceptual capabilities.

Our main contributions are: **1)** We introduce Visual Jigsaw, a lightweight and verifiable self-supervised post-training task that enhances vision-centric perception and understanding capabilities in MLLMs. It requires no additional generative modules and integrates seamlessly with existing text-only models. **2)** We instantiate Visual Jigsaw across three visual modalities—images, videos, and 3D data—and demonstrate consistent improvements in fine-grained perception, temporal understanding, and 3D spatial reasoning, thereby establishing its generality and effectiveness. **3)** We highlight the potential of self-supervised tasks focused explicitly on the visual signal as a promising, complementary direction for enhancing the vision-centric abilities of MLLMs.

2 Related Works

2.1 Self-supervised Learning

Self-supervised learning (SSL), wherein pretext tasks derive supervision directly from input data, has become a cornerstone of visual representation learning. Early approaches included context-

based tasks such as predicting relative patch positions (Doersch et al., 2015) and patch orderings (Noroozi & Favaro, 2016). While these works revealed the potential of such proxy tasks, they were limited in scalability. More recently, SSL has been dominated by two major families: (1) reconstruction-based methods (Zhou et al., 2021; He et al., 2022b; Bao et al., 2021; Assran et al., 2023) and (2) discriminative methods (He et al., 2020; Chen et al., 2020; Caron et al., 2021). These approaches have demonstrated impressive scalability and transferability, establishing strong foundations for large-scale vision pre-training (Oquab et al., 2023; Siméoni et al., 2025).

Parallel to these approaches, jigsaw pretext tasks explicitly formulate visual learning as an ordering problem, requiring the model to recover the spatial or temporal structure of visual inputs. Noroozi & Favaro (2016) pioneered the 3×3 image jigsaw puzzle, which was later extended for iterative refinements (Wei et al., 2019), domain generalization (Carlucci et al., 2019), and fine-grained reasoning (Du et al., 2020). Extensions to video include spatiotemporal jigsaws (Ahsan et al., 2019; Huo et al., 2021; Wang et al., 2022) that jointly exploit appearance and motion cues. Recent jigsaw-style variants (Caron et al., 2024; Wang et al., 2023; Zhai et al., 2022) have shown competitive results with contrastive learning and masked image modeling when properly designed and scaled.

The characteristics of jigsaw-style tasks also make them a good fit for understanding-based MLLMs, which are optimized for visual understanding with textual outputs rather than dense reconstruction. A visual jigsaw task thus provides a lightweight, verifiable objective that requires no additional generative modules. Building on these advantages, our work introduces Visual Jigsaw as a self-supervised post-training stage to enhance vision-centric perception in MLLMs across image, video, and 3D modalities.

2.2 MLLM Visual Understanding

MLLMs (Hurst et al., 2024; Comanici et al., 2025; Bai et al., 2025; Zhu et al., 2025a) have rapidly advanced, achieving strong performance across diverse multimodal tasks. These improvements have largely stemmed from more powerful LLM backbones, better image resolution strategies, improved vision encoders, higher-quality training datasets, and post-training techniques. However, relatively little attention has been devoted to enhancing the *intrinsic visual understanding* of MLLMs. Most existing efforts rely on scaling data to indirectly improve perception-related tasks.

X-Former(Swetha et al., 2024) adds visual reconstruction objective in MLLM training but its target is to better extract and combine vision features from two vision encoders and only supervise the connector with the reconstruction objective. Recent works (Wang et al., 2025b;a) demonstrate that explicitly adding visual reconstruction objectives enhances visual understanding, but such approaches require introducing extra generative modules and objectives, have only been demonstrated in settings where the MLLM is trained jointly with reconstruction from the beginning, and have not been validated on stronger models like Qwen2.5-VL (Bai et al., 2025). Meanwhile, unified multimodal models (UMMs) (Xie et al., 2025b; Chen et al., 2025c; Deng et al., 2025; Chen et al., 2025b) explore combining vision understanding and generation in one model, but it has only been shown that understanding benefits visual generation Xie et al. (2025a) while optimizing generative objectives sometimes harm understanding abilities Pan et al. (2025); Chen et al. (2025b). In contrast, we propose a lightweight, post-training self-supervised task that strengthens visual perception and understanding in MLLMs without altering the architecture.

2.3 MLLM RL Post-training

RL post-training has played a pivotal role in advancing LLMs. Early paradigms such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) focused on improving alignment with human preferences, while recent developments like RLVR (Lambert et al., 2024; Shao et al., 2024) have been shown to substantially enhance reasoning capabilities. Inspired by this success, the MLLM community has begun to apply similar paradigms. Most works concentrate on strengthening multimodal reasoning for mathematical and scientific tasks (Meng et al., 2025; Huang et al., 2025; Yuan et al., 2025a; Wang et al., 2025f). These RL-based approaches have also been extended to video (Feng et al., 2025; Chen et al., 2025d) and 3D domains (Yuan et al., 2025b). Other methods focus on specific vision tasks such as grounding (Liu et al., 2025b) and segmentation (Liu et al., 2025a). More recent efforts (Zheng et al., 2025; Su et al., 2025) also explore teaching MLLMs to use vision

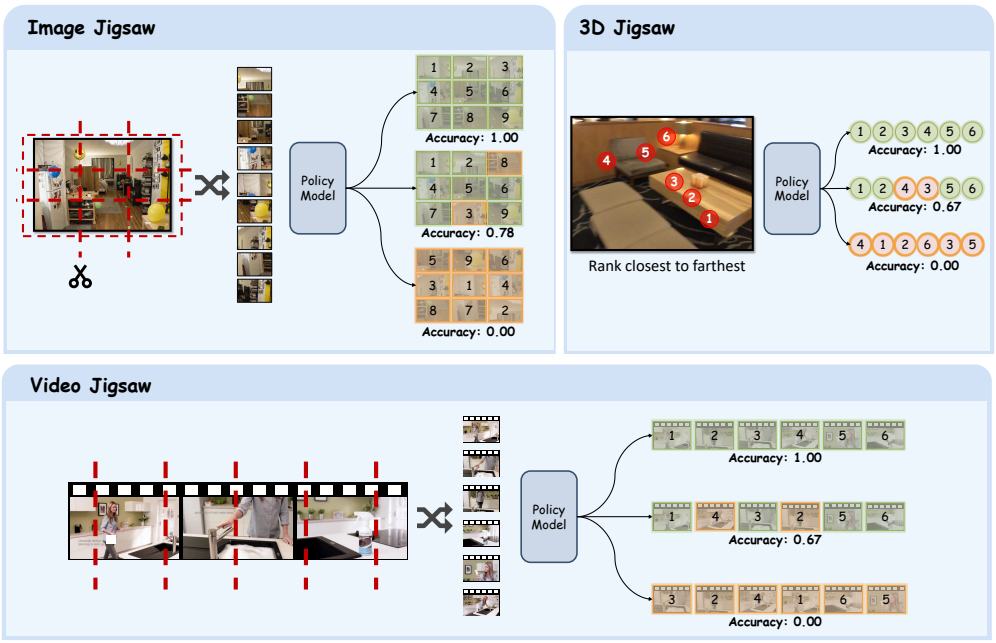


Figure 2: **Illustration of the Visual Jigsaw tasks.** In the Image Jigsaw (top left), an image is partitioned into non-overlapping patches, shuffled into a sequence, and the model is tasked with predicting the correct raster order. In the Video Jigsaw (bottom), a video is segmented into temporal clips, shuffled, and the model predicts their original chronological order. In the 3D Jigsaw (top right), points with distinct depth values are sampled from an RGB-D image, shuffled and annotated in the RGB view, and the model is required to recover the correct depth order from nearest to farthest. Across all tasks, the policy model outputs an ordering that is compared against the ground truth, and a partial accuracy reward is assigned when only some elements are correctly ordered.

tools, enabling models to think with images and better retrieve and perceive the visual input via operations like crop and zoom-in.

However, the majority of these approaches still target text-based reasoning, task-specific objectives, or visual tool-calling, rather than directly improving intrinsic visual perception. Vicrit (Wang et al., 2025e) and LLaVA-Critic-R1 (Wang et al., 2025d) enhance perception and reasoning by detecting errors in captions or judging textual responses, but their training signals are ultimately tied to text–image alignment instead of intrinsic visual signal understanding. The most closely related work, Jigsaw-R1 Wang et al. (2025g), also attempts to introduce a jigsaw task for MLLM post-training. However, its approach struggles even on simple 2×2 image jigsaws, thus focusing mainly on predicting the relative position of a pair of patches. In contrast, our method leverages the standard, more complex visual jigsaw tasks to systematically enhance MLLM perception, and we demonstrate its effectiveness not only on images but also across video and 3D modalities.

3 Method

3.1 Visual Jigsaw

Our proposed Visual Jigsaw framework (illustrated in Fig 2) is formulated as a general visual ordering problem. Given some data from a certain visual modality (image, video, or 3D), we derive a set of K jigsaw elements by applying a modality-specific partitioning rule, such as splitting an image into patches, segmenting a video into clips, or sampling points in a 3D scene. These elements are then shuffled, and the model is tasked with predicting their original structural arrangement. Formally, the model predicts a permutation of size K as a list of indices, which is then compared against the ground-truth permutation. We optimize this task using the GRPO algorithm. The following subsections detail our reward design and describe the specific instantiations of Visual Jigsaw across each of the three visual modalities.

3.1.1 REWARD DESIGN

The ground-truth of the visual jigsaw tasks is a list of indices that is directly verifiable. Instead of assigning only a binary accuracy reward, we design a graded reward function. An output that exactly matches the ground-truth permutation receives an accuracy reward of 1. For a valid but partially correct permutation, the reward is the fraction of correctly placed indices, scaled by a discount factor $\gamma \in (0, 1)$. This discount penalizes incomplete solutions, preventing the model from overestimating partial matches while still providing learning signals. To avoid reward hacking (*e.g.* predicting the same index for all positions), any output that is not a valid permutation of size K is assigned a reward of 0. Formally, the accuracy reward function is given by

$$Reward(o, g) = \begin{cases} 1, & \text{if } o = g \\ \gamma \cdot \dfrac{1}{K} \sum_{i=1}^{K} \mathbf{1}[o_i = g_i], & \text{if ValidPermutation}(o) \wedge o \neq g, \\ 0, & \text{otherwise} \end{cases}$$

where o denotes the model's predicted permutation, g the ground-truth permutation, K the number of jigsaw pieces, γ the discount factor for partial correctness, and ValidPermutation(o) an indicator of whether o is a valid permutation of size K.

Besides the accuracy reward, we also require the model to put its thinking process within `<think></think>` and the final answer within `<answer></answer>`. A format reward of 0.2 will be assigned to outputs following the correct format, while outputs with an incorrect format will receive 0 values for both format and accuracy rewards.

3.1.2 IMAGE JIGSAW

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we first partition it into a grid of $m \times n$ non-overlapping patches, each of size $\frac{H}{m} \times \frac{W}{n}$. This produces $K = m \times n$ patches arranged in raster order (row-major, top-left to bottom-right):

$$\mathcal{P} = [p_1, p_2, \ldots, p_K], \quad p_i \in \mathbb{R}^{\frac{H}{m} \times \frac{W}{n} \times 3}.$$

We then apply a random permutation $\pi : \{1, 2, \ldots, K\} \to \{1, 2, \ldots, K\}$ that maps an original position index i to its shuffled position $\pi(i)$. The shuffled sequence of patches can therefore be written as

$$\mathcal{P}_\pi = [\, p_{\pi^{-1}(1)},\ p_{\pi^{-1}(2)},\ \ldots,\ p_{\pi^{-1}(K)}\,],$$

where the j-th element corresponds to the patch originally at position $\pi^{-1}(j)$. Given $\mathcal{P}_\pi$, the model's objective is to recover the original arrangement by predicting the correct permutation of the input patch indices $[\pi(1), \pi(2), \ldots, \pi(K)]$.

For training, we use 118K images from the COCO dataset Lin et al. (2014). We set $m = n = 3$, yielding 9 patches per image, and we filter out images with side lengths smaller than 84 pixels to avoid overly small patches. The prompt template for this task is provided in Appendix A.6.

3.1.3 VIDEO JIGSAW

Given a video $V \in \mathbb{R}^{T \times H \times W \times 3}$ with T frames, we segment it uniformly along the temporal axis into K non-overlapping clips, each containing $\frac{T}{K}$ consecutive frames:

$$\mathcal{V} = [v_1, v_2, \ldots, v_K], \quad v_i \in \mathbb{R}^{\frac{T}{K} \times H \times W \times 3}.$$

We then apply a random permutation $\pi : \{1, 2, \ldots, K\} \to \{1, 2, \ldots, K\}$, where $\pi(i)$ denotes the shuffled position of the i-th clip in the original chronological order. The shuffled sequence is written as

$$\mathcal{V}_\pi = [\, v_{\pi^{-1}(1)},\ v_{\pi^{-1}(2)},\ \ldots,\ v_{\pi^{-1}(K)}\,].$$

The model's objective is to restore the original chronological order by predicting the correct permutation $[\pi(1), \pi(2), \ldots, \pi(K)]$.

For training, we use 100K videos from the LLaVA-Video dataset (Zhang et al., 2024c). Each video is divided into 6 clips ($K = 6$). To prevent the model from exploiting simple frame-matching cues at clip boundaries, we trim 5% of the frames from both the beginning and end of each clip. The maximum number of frames for each clip is set to 12, and the maximum resolution for each frame is set to $128 \times 28 \times 28$ pixels. We remove videos with lengths smaller than 24 seconds. The prompt template for this task can be found in Appendix A.6.

3.1.4 3D JIGSAW

A canonical 3D jigsaw task would mirror its 2D image and video counterparts, involving the partitioning of 3D space into volumetric primitives (*e.g.* voxels, mesh fragments, or point cloud segments) and tasking the model with recovering the original spatial arrangement. Such formulations would fully leverage geometric information in native 3D representations. However, current general-purpose MLLMs typically process 3D-related tasks via 2D images or videos rather than directly operating on raw 3D data structures.

We therefore design a practical variant of the 3D jigsaw based on RGB-D images. Given an RGB-D image, we randomly select K points with distinct depth values, forming a sequence sorted by depth from nearest to farthest:

$$\mathcal{P} = [p_1, p_2, \ldots, p_K], \quad d_{p_1} < d_{p_2} < \cdots < d_{p_K},$$

where d_{p_i} is the depth of point p_i.

We then apply a random permutation $\pi : \{1, 2, \ldots, K\} \to \{1, 2, \ldots, K\}$ to obtain a shuffled sequence of the points

$$\mathcal{P}_\pi = [\, p_{\pi^{-1}(1)},\ p_{\pi^{-1}(2)},\ \ldots,\ p_{\pi^{-1}(K)} \,].$$

Each point is annotated with its index in $\mathcal{P}_\pi$ on the RGB image. The model is tasked with recovering the correct depth order by predicting the permutation $[\pi(1), \pi(2), \ldots, \pi(K)]$ that restores $\mathcal{P}$.

For this task, we use the RGB-D data from ScanNet (Dai et al., 2017) and generate 300K training samples in total. We construct training samples by randomly selecting 6-point combinations from depth maps, restricting the points to lie within a range of 0.1 m to 10 m. To ensure diversity, any two points in a combination must be separated by at least 40 pixels in the image and differ in depth by more than 0.2 m. The prompt template for this task can be found in Appendix A.6. We also experimented with alternative designs of 3D jigsaw tasks, which are provided in Appendix A.2.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

We adopt Qwen2.5-VL-7B-Instruct as the base MLLM for all experiments. We use the GRPO algorithm and remove both the KL regularization and the entropy loss. The discount factor γ for partially correct predictions is set to 0.2. The training is performed with a global batch size of 256 for image jigsaw and 128 for video & 3D jigsaw, and the learning rate is 1×10^{-6}. For each prompt, we sample 16 responses with a decoding temperature of 1.0. Both image and video jigsaw tasks are trained for 1000 steps, and the 3D jigsaw is trained for 800 steps.

For the image jigsaw task, our default RL training with batch size 256 and 1000 training steps costs 840 H100 GPU-hours. Video jigsaw RL training (batch size 128 and 1000 training steps) and 3D jigsaw RL training (batch size 128 and 800 training steps) consume 1600 GPU-hours and 310 GPU-hours, respectively.

4.2 MAIN RESULTS

This section presents quantitative results. Qualitative examples are provided in Appendix A.4.

4.2.1 IMAGE JIGSAW

We evaluate the model trained with image jigsaw across three categories of vision-centric benchmarks including 1) **Fine-grained perception & understanding**: MMVP (Tong et al., 2024), fine-grained perception subset of MMStar (Chen et al., 2024), MMBench (Liu et al., 2024a), HR-Bench

(Wang et al., 2025c), V* (Wu & Xie, 2024), MME-RealWorld (lite) (Zhang et al., 2025b), LISA-Grounding (Lai et al., 2024), OVD-Eval (Yao et al., 2024); 2) **Monocular spatial understanding**: VSR (Liu et al., 2023), OminiSpatial (Jia et al., 2025), DA-2K (Yang et al., 2024); 3) **Compositional visual understanding**: Winoground (Thrush et al., 2022), SugerCrepe++ (Dumpala et al., 2024).

We include three baselines, which are all post-trained from Qwen2.5-VL-7B. ThinkLite-VL (Wang et al., 2025f) mainly focuses on improving multimodal reasoning. VL-Cogito (Yuan et al., 2025a) is trained on a broader set of tasks, including general image understanding and counting, in addition to mathematical and scientific reasoning. LLaVA-Critic-R1 (Wang et al., 2025d) is trained with the critic task and shows improvement in image perception and understanding. As these vision-centric benchmarks mainly focus on direct visual perception and understanding, we directly evaluate the model to give the short answer without the thinking/reasoning process for fair comparison. This protocol is further motivated by our finding that enabling chain-of-thought reasoning can actually degrade the performance of some reasoning models on some specific benchmarks (*e.g.* 35.78 $\rightarrow$ 31.44 on OVD-Eval for ThinkLite-VL).

Table 1: **Evaluation results on image benchmarks.** Image Jigsaw achieves consistent improvements across fine-grained perception, spatial understanding, and compositional understanding tasks.

	Fine-grained Perception & Understanding								Spatial Und (Mono)			Compositional Und	
Model	**MMVP**	**MMStar (fine-grained)**	**MMBench**	**HR-Bench-8K**	**V***	**MME-RealWorld**	**LISA-Grounding**	**OVD-Eval**	**VSR**	**OmniSpatial**	**DA-2K**	**Winoground**	**SugarCrepe++**
	test	fine	en_dev	test	test	lite	test	test	test	test	val	g-acc	test
ThinkLite-VL	55.33	59.95	84.19	68.12	76.96	46.17	73.70	35.78	78.09	42.60	58.46	35.25	61.49
VL-Cogito	55.33	56.64	82.98	69.62	79.58	47.63	72.26	35.78	79.82	44.29	56.43	38.25	63.59
LLaVA-Critic-R1	53.33	57.80	83.16	67.50	78.01	45.18	68.52	35.28	78.50	42.73	53.82	34.75	61.93
Qwen2.5-VL-7B	54.66	59.75	83.33	67.38	76.96	43.41	71.89	35.07	77.68	42.66	54.45	37.00	61.59
Image Jigsaw (SFT)	56.00	60.94	83.67	69.75	80.10	43.88	66.59	34.35	80.68	43.55	61.46	38.75	62.03
Image Jigsaw	60.66	65.81	84.45	71.13	80.63	45.96	74.54	36.49	80.36	44.49	60.35	39.00	63.02
(Gain)	+6.00	+6.06	+1.12	+3.75	+3.66	+2.55	+2.65	+1.42	+2.68	+1.83	+5.90	+2.00	+1.43

Tab 1 shows that our method consistently improves the vision-centric capabilities on the three types of benchmarks. These results confirm that incorporating image jigsaw post-training significantly enhances MLLMs' perceptual grounding and fine-grained vision understanding beyond reasoning-centric post-training strategies. We attribute these improvements to the fact that solving image jigsaw requires the model to attend to local patch details, infer global spatial layouts, and reason about inter-patch relations, which directly benefits fine-grained, spatial, and compositional understanding.

4.2.2 VIDEO JIGSAW

For video jigsaw, we evaluate on a comprehensive suite of video benchmarks: AoTBench (Xue et al., 2025), Vinoground (Zhang et al., 2024a), TOMATO (Shangguan et al., 2024), FAVOR-Bench (Tu et al., 2025), TUNA-Bench (Kong et al., 2025), Video-MME (Fu et al., 2025), TempCompass (Liu et al., 2024b), TVBench (Cores et al., 2024), MotionBench (Hong et al., 2025), LVBench (Wang et al., 2024b), VSI-Bench (Yang et al., 2025), Video-TT (Zhang et al., 2025c), CVBench (Zhu et al., 2025b).

We include the Video-R1 (Feng et al., 2025) baseline for comparison, which is trained with cold-start SFT followed by RL for video understanding and reasoning. We enable the thinking process when evaluating Video-R1, as we find its performance is generally better than direct answering. For all models, we set the maximum number of pixels to $256 \times 28 \times 28$ and evaluate under three different frame settings (16, 32, 64).

From the results shown in Tab 2, we observe that Video Jigsaw brings consistent improvements across all video understanding benchmarks and frame settings. While our method enhances gen-

Table 2: **Evaluation results on video benchmarks.** Video Jigsaw consistently improves over the baseline across all benchmarks and frame settings.

Model	Frames	AoTBench	Vinoground	TOMATO	FAVOR-Bench	TUNA-Bench	VideoMME	TempCompass	TVBench	MotionBench	LVBench	VSI-Bench	Video-TT	CVBench
		vqa	group	test	test	test	wo subs	mc	test	val	test	test	mcq	test
Video-R1	16	45.06	9.40	27.29	49.47	53.00	56.62	70.19	51.80	55.82	34.53	34.34	42.95	47.50
Video-R1	32	47.53	10.20	27.29	49.90	54.26	59.88	71.77	53.54	56.12	38.61	35.11	42.63	48.10
Video-R1	64	48.68	10.60	27.36	50.51	54.33	60.85	72.59	53.43	56.09	38.80	36.61	42.74	48.69
Qwen2.5-VL-7B	16	45.52	12.60	25.87	48.54	53.14	57.44	71.77	49.94	55.56	33.51	32.79	38.39	47.70
Qwen2.5-VL-7B	32	49.48	18.20	26.34	49.34	54.88	60.70	72.59	51.96	56.47	39.19	35.34	41.57	49.60
Qwen2.5-VL-7B	64	52.41	21.80	26.35	50.86	55.79	63.44	72.84	53.74	56.29	40.35	37.74	42.25	51.50
Video Jigsaw	16	51.67	15.20	27.56	49.69	55.10	58.07	73.10	51.33	56.87	36.41	35.39	40.19	49.80
(Gain)		+6.15	+2.60	+1.69	+1.15	+1.96	+0.63	+1.33	+1.39	+1.31	+2.90	+2.60	+1.80	+2.10
Video Jigsaw	32	55.00	21.40	28.03	50.56	56.49	62.37	73.60	53.31	57.99	39.70	38.47	43.27	51.60
(Gain)		+5.52	+3.20	+1.69	+1.22	+1.61	+1.67	+1.01	+1.35	+1.52	+0.51	+3.13	+1.70	+2.00
Video Jigsaw	64	57.64	25.20	28.30	52.27	56.63	64.74	73.60	54.18	57.91	41.83	40.40	44.11	54.50
(Gain)		+5.23	+3.40	+1.95	+1.41	+0.84	+1.30	+0.76	+0.44	+1.62	+1.48	+2.66	+1.86	+3.00

eral video perception and comprehension, the gains are particularly pronounced on tasks requiring temporal-centric understanding and reasoning about temporal directionality (*e.g.* AoTBench). Furthermore, the strong gains on CVBench demonstrate improved cross-video understanding and reasoning. These results confirm that solving video jigsaw tasks encourages the model to better capture temporal continuity, understand relationships across videos, reason about directional consistency, and generalize to holistic and generalizable video understanding scenarios.

4.2.3 3D JIGSAW

For the 3D modality, we evaluate the model on a diverse set of benchmarks that span various aspects of 3D understanding: SAT-Real (Ray et al., 2024), 3DSRBench (Ma et al., 2024), ViewSpatial (Li et al., 2025), All-Angles (Yeh et al., 2025), OminiSpatial (Jia et al., 2025), VSI-Bench (Yang et al., 2025), SPARBench (tiny) (Zhang et al., 2025a), and DA-2K (Yang et al., 2024).

Table 3: **Evaluation results on 3D benchmarks.** 3D Jigsaw consistently enhances performance on both directly related depth comparison tasks (DA-2K) and broader 3D perception tasks spanning single-view, multi-view, and egocentric video inputs.

Model	SAT-Real	3DSRBench	ViewSpatial	All-Angles	OmniSpatial	VSI-Bench	SPARBench	DA-2K
	test	test	test	test	test	test	tiny	test
Qwen2.5-VL-7B	48.66	57.42	36.52	47.56	42.66	37.74	35.75	54.45
3D Jigsaw	64.00	58.13	38.62	49.06	45.99	40.64	38.31	71.56
(Gain)	+15.34	+0.71	+2.10	+1.50	+3.33	+2.90	+2.56	+17.11

As shown in Tab 3, 3D Jigsaw achieves significant improvements across all benchmarks. Unsurprisingly, the largest gain is on DA-2K, a depth estimation benchmark that is directly related to our depth-ordering pre-training task. More importantly, we observe consistent improvements on a wide range of other tasks, including those with single-view (*e.g.* 3DSRBench, OminiSpatial), multi-view (*e.g.* ViewSpatial, All-Angles), and egocentric video inputs (*e.g.* VSI-Bench). These results demonstrate that our approach not only teaches the specific skill of depth ordering but also effectively strengthens the model's general ability to perceive and reason about 3D spatial structure.

4.3 Ablation Studies and Discussions

SFT vs. RL. We investigate the difference between using SFT and RL to train the visual jigsaw task, focusing on the image jigsaw setting. As shown in the Image Jigsaw (SFT) entry of Tab 1, SFT leads to moderate improvements on some benchmarks, but the gains are notably smaller than those achieved with RL. Moreover, on certain benchmarks (*e.g.* LISA-Grounding and OVD-Eval), SFT causes a significant performance degradation, suggesting that the model overfits to the jigsaw task and fails to transfer the learned skills. This observation is consistent with recent findings that SFT tends to memorization, while RL is better at promoting generalization (Huan et al., 2025; Chu et al., 2025). Our results confirm that RL enables the model to more effectively generalize the vision-centric capabilities acquired from visual jigsaw training to related downstream tasks. For comparison, the SFT training cost is 58 H100 GPU-hours which is much smaller than RL (840 GPU-hours), as the on-policy training process requires generating multiple rollouts for each sample to estimate rewards and update the policy. However, the additional computational investment is justified as it brings robust generalization and substantially improved performance.

How does the difficulty of the visual jigsaw tasks affect the performance?

We conducted ablation experiments to investigate how the difficulty of the jigsaw task affects model performance. We varied the complexity of both the image and video jigsaw tasks: for the image task, we reduced the grid size from 3×3 to 2×2; for the video task, we reduced the number of clips from six to four. We then measured the average performance across all corresponding benchmarks (using 16 frames for the video evaluation). The results in Fig 3 show that while easier jigsaw tasks still yield performance improvements over the baseline, the gains are substantially smaller than those from the standard, more difficult tasks. This indicates that a higher degree of difficulty provides a stronger supervisory signal for enhancing fine-grained perception and temporal reasoning. Critically, we also found that on challenging setups (*e.g.* 3×3 for images), the design of the partial accuracy reward becomes crucial. Without this design, the model fails to learn the task, as sparse binary feedback is insufficient to bootstrap learning in the early stages of training.

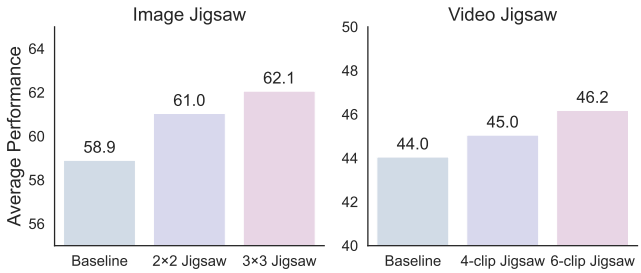


Figure 3: Performance with different jigsaw difficulties on image and video tasks.

Visual Jigsaw on other base models. To validate the general effectivenss of our Visual Jigsaw framework, we apply it on a stronger base model MiMo-VL-7B-SFT-2508 (Xiaomi, 2025) and show the results in Table 4, where Visual Jigsaw also yields consistent gains over the baseline.

Model	Image Avg	Video Avg	3D Avg
MiMo-VL-7B-SFT	63.77	51.84	50.67
+ Visual Jigsaw	65.14	54.47	52.91

Table 4: Results of Visual Jigsaw on MiMo-VL.

Visual Jigsaw on Reasoning MLLMs.

Harder Jigsaw Configurations. We also conduct experiments on harder jigsaw configurations with 4×4 for image jigsaw and 8-clip for video jigsaw. For 4×4 image jigsaw, We find that the model failed to learn this task effectively, with jigsaw accuracy remaining low. This happens even when we do cross-configuration adaptation by first training on 3×3 jigsaw task. We attribute this to three key factors: (1) Information Scarcity: At the typical resolution of our COCO training data ($\approx$640$\times$480), 4×4 tiles are often too small and fragmented to contain sufficient semantic information for re-assembly. (2) Semantic Ambiguity: Small tiles from uniform areas (e.g., sky, walls) become perceptually indistinguishable, making the ordering task ill-posed. (3) Combinatorial Explosion. The permutation space expands from 9! to 16!, a significant increase in complexity for the model. This suggests that scaling to harder configurations is not merely a matter of increasing grid size, but will likely require a combination of higher-resolution training data, tile-diversity constraints, and curriculum learning strategies. As for the 8-clip video jigsaw, the average performance

is 46.24 with 16-frame setting which does not bring notable improvement over 6-clip. This is mainly because many LLaVA-Video samples are around 30 s, so splitting into 8 clips makes segments short and sometimes ambiguous. We leave the study 8-clip on longer videos for future works.

Table 5: Performance of Visual Jigsaw on a reasoning-oriented MLLM (ThinkLite-VL), showing improved visual perception while preserving reasoning ability.

Model	Vision-Centric Avg	MathVista testmini	MathVision testmini	MathVerse testmini	MMMU val	EMMA mini
ThinkLite-VL	59.69	75.20	30.92	50.76	55.11	26.75
+ Image Jigsaw	61.60	75.10	35.20	50.50	54.22	29.00

We further explore whether Visual Jigsaw can also benefit reasoning MLLMs that have already undergone reasoning-intensive RL post-training. To this end, we select ThinkLite-VL as the base model and apply the image jigsaw training. We enable the KL constraint to better preserve the reasoning capability during training. We evaluate the resulting model on both vision-centric benchmarks and multimodal reasoning benchmarks, including MathVista (Lu et al., 2023), MathVision (Wang et al., 2024a), MathVerse (Zhang et al., 2024b), MMMU (Yue et al., 2024), and EMMA(Hao et al., 2025). As shown in Tab 5, the reasoning MLLM trained with Visual Jigsaw achieves clear improvements in visual perception and understanding, while maintaining its strong reasoning ability.

How does model solve jigsaw problems? Here we provide an example think trace for image jigsaw training in Fig 4. We can see that the model identifies semantic objects (trees, sky, a person) in different image tiles, infers spatial context (horizon line, pathway edges), and uses logical deduction to place the tiles into a coherent global scene. More examples are provided in Appendix A.1.

<think>
I am mentally reassembling the image based on the spatial relationships and features of the tiles. Tile 2 is positioned at the upper right corner, showing a dense cluster of snow-covered trees with a dark, shadowed area to its right, suggesting the edge of a forested area or a hillside. Tile 5, located at the top-left portion of the image, appears to be part of a snowy, overcast sky with tree branches extending from the left side, indicating a continuation of the snowy landscape with a clear horizon line. Tile 1, situated at the bottom-left and lower middle-left areas, shows a snowy slope with a darker, shaded section that transitions into a lighter, snow-covered ground, consistent with a path or trail leading into a wooded area. Tile 7, occupying the central and right-middle parts of the image, depicts a person wearing winter gear, standing on a snowy path with ski poles, likely engaged in skiing or walking through a snowy environment. Tiles 1 and 8 appear to form the sides of this snowy pathway, with tile 1 on the left and tile 8 on the right, framing the central figure and the trail. This arrangement suggests a continuous snowy outdoor scene with a focus on a person navigating a snowy terrain.
</think>
<answer> 6, 5, 2, 1, 7, 8, 4, 9, 3</answer>

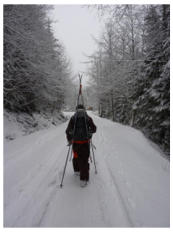


Figure 4: Example thinking process for solving image jigsaw.

5 Conclusion

In this work, we introduced Visual Jigsaw, a verifiable self-supervised post-training framework that enhances vision-centric understanding in MLLMs. By formulating visual understanding as an ordering problem and optimizing it with RLVR, visual jigsaw avoids the need for dense visual reconstruction and integrates seamlessly into text-only MLLMs. Our experiments demonstrate the generality of this approach, yielding consistent improvements across images, videos, and 3D data in fine-grained perception, temporal reasoning, and 3D spatial understanding. Ultimately, our work highlights the potential of perception-focused self- and weakly-supervised tasks as a powerful and complementary path toward developing more capable and robust multimodal models.

Ethics Statement

This work uses only publicly available datasets that follow established licenses and guidelines. Our method focuses on improving vision-centric perception and understanding in MLLMs without introducing additional risks beyond existing models. As with other MLLMs, potential misuse or biases may arise if training data are not carefully curated. We emphasize responsible usage, transparency, and alignment with human intentions to maximize benefits while mitigating risks.

Reproducibility statement

All datasets used in the experiments are publicly available, and we provide detailed descriptions of the used datasets and preprocessing information in Sec 3. The evaluation benchmarks and details are provided for each modality in Sec 4. The training setup and hyperparameters are described in Sec 4.1. Our implementation is based on the open-source `verl` (Sheng et al., 2024) library, with the main modifications including the construction of visual jigsaw data and the reward calculation. The corresponding code is provided in the supplementary materials. To further facilitate reproducibility, we will release the code, data, and models to reproduce all main experiments, along with instructions for running ablation studies and evaluations.

References

Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *WACV*, 2019.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Mathilde Caron, Neil Houlsby, and Cordelia Schmid. Location-aware self-supervised transformers for semantic segmentation. In *WACV*, 2024.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025a. URL `https://arxiv.org/abs/2504.11468`.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025c.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint arXiv:2503.24376*, 2025d.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *ICML*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Tvbench: Redesigning video-language evaluation, 2024.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *ECCV*, 2020.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *NeurIPS*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. In *ICML*, 2025.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022a.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022b.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *CVPR*, 2025.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Yuqi Huo, Mingyu Ding, Haoyu Lu, Zhiwu Lu, Tao Xiang, Ji-Rong Wen, Ziyuan Huang, Jianwen Jiang, Shiwei Zhang, Mingqian Tang, et al. Self-supervised video representation learning with constrained spatiotemporal jigsaw. 2021.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Fanheng Kong, Jingyuan Zhang, Hongzhi Zhang, Shi Feng, Daling Wang, Linhao Yu, Xingguang Ji, Yu Tian, Fuzheng Zhang, et al. Tuna: Comprehensive fine-grained temporal understanding evaluation on dense dynamic videos. *arXiv preprint arXiv:2505.20124*, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *TACL*, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024b.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

Sirnam Swetha, Jinyu Yang, Tal Neiman, Mamshad Nayeem Rizve, Son Tran, Benjamin Yao, Trishul Chilimbi, and Mubarak Shah. X-former: Unifying contrastive and reconstruction learning for mllms. In *ECCV*, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024.

Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding. *arXiv preprint arXiv:2503.14935*, 2025.

Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi Wang. Autoregressive semantic visual reconstruction helps vlms understand better. *arXiv preprint arXiv:2506.09040*, 2025a.

Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *ECCV*, 2022.

Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tong Wang, and ZHAO-XIANG ZHANG. Droppos: Pre-training vision transformers by reconstructing dropped positions. *NeurIPS*, 2023.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Ge Zheng, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *ICLR*, 2025b.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *AAAI*, 2025c.

Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. Llava-critic-r1: Your critic model is secretly a strong policy model. *arXiv preprint arXiv:2509.00676*, 2025d.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Yongyuan Liang, Yuhang Zhou, Xiaoyu Liu, Ziyi Zang, Ming Li, Chung-Ching Lin, Kevin Lin, et al. Vicrit: A verifiable reinforcement learning proxy task for visual perception in vlms. *arXiv preprint arXiv:2506.10128*, 2025e.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025f.

Zifu Wang, Junyi Zhu, Bo Tang, Zhiyu Li, Feiyu Xiong, Jiaqian Yu, and Matthew B Blaschko. Jigsaw-r1: A study of rule-based visual reinforcement learning with jigsaw puzzles. *arXiv preprint arXiv:2505.23590*, 2025g.

Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *CVPR*, 2019.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024.

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025a.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025b.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024.

Yiyang Yao, Peng Liu, Tiancheng Zhao, Qianqian Zhang, Jiajia Liao, Chunxin Fang, Kyusong Lee, and Qing Wang. How to evaluate the generalization of detection? a benchmark for comprehensive open-vocabulary detection. In *AAAI*, 2024.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*, 2025a.

Zhihao Yuan, Shuyi Jiang, Chun-Mei Feng, Yaolun Zhang, Shuguang Cui, Zhen Li, and Na Zhao. Scene-r1: Video-grounded large language models for 3d scene reasoning without 3d annotations. *arXiv preprint arXiv:2506.17545*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Shuangfei Zhai, Navdeep Jaitly, Jason Ramapuram, Dan Busbridge, Tatiana Likhomanenko, Joseph Yitan Cheng, Walter Talbott, Chen Huang, Hanlin Goh, and Joshua Susskind. Position prediction as an effective pretraining strategy. In *ICML*, 2022.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025a.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024a.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*, 2024b.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024c. URL https://arxiv.org/abs/2410.02713.

Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *ICCV*, 2025c.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. *arXiv preprint arXiv:2508.19542*, 2025b.

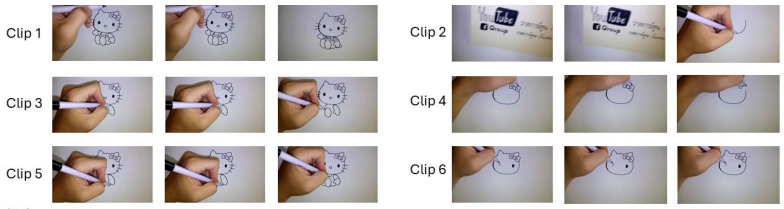
A APPENDIX

USE OF LARGE LANGUAGE MODELS

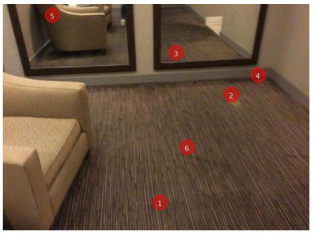
In this work, we have used large language models to assist in polishing the writing of the paper. Specifically, the LLM was employed to check grammar correctness and to provide alternative phrasings or stylistic suggestions for certain sentences. All suggestions were carefully reviewed by the authors, and only adopted after manual verification and modification when appropriate.

A.1 QUALITATIVE EXAMPLES OF THINKING PROCESS IN VISUAL JIGSAW

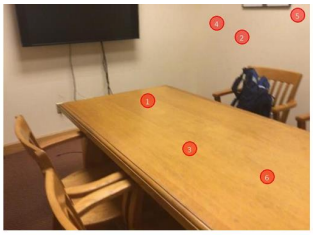
<think>
I am mentally reassembling the image based on the spatial relationships and context of the visible elements in each tile. Tile 8 is positioned at the top-left corner of the image, showing part of a building with a partial view of a street or urban environment, likely indicating the upper-left side of a scene with buildings and possibly residential or commercial structures. Tile 1 occupies the top-right portion of the image, showing a continuation of the urban environment with trees and parts of a building facade, consistent with a street or plaza area. Tile 7 appears to be situated on the left side, showing people from behind, interacting near what seems like a market or street vendor setup, with boxes and goods visible, suggesting this could be a marketplace or public gathering area. Tile 6, located centrally towards the right, shows a continuation of the market activity with a focus on goods being sold, including produce and signage, consistent with a busy street market environment. Tile 3, positioned on the lower right, shows a continuation of the street scene with a vehicle and pavement, providing a sense of depth and perspective to the lower right-hand side of the image, completing the street-level view. By integrating these observations, I deduce the spatial organization of the image.
</think>
<answer> 8, 9, 1, 7, 6, 5, 4, 3, 2</answer>



<think>
To determine the correct order, we analyze the logical progression of drawing Hello Kitty: \n\n1. **Clip 2** begins with a hand holding a pen, then starts drawing the initial U - shape for the head. \n2. **Clip 4** builds on Clip 2 by adding the bow to the head. \n3. **Clip 6** continues by drawing the eyes, nose, and whiskers on the head. \n4. **Clip 3** moves to the body, drawing the arms and starting the lower part of the body. \n5. **Clip 5** further develops the body, adding more details to the limbs. \n6. **Clip 1** completes the body with additional details like stripes.
</think>
<answer> 2, 4, 6, 3, 5, 1</answer>



<think>
Let's analyze this step-by-step:\n\n1. **Point 1**: It's right in the foreground, so it has to be very close.\n2. **Point 6**: It's on the carpet directly in front of Point 1.\n3. **Point 2**: It's further back but closer to the wall than Point 3.\n4. **Point 4**: It's on the wall to the right.\n5. **Point 5**: It's in the mirror on the left.\n6. **Point 3**: It's in the mirror on the right.
</think>
<answer> 1, 6, 2, 4, 5, 3</answer>



<think>
Let's analyze this step-by-step:\n\n1. **Point 6**: It's right in the foreground on the table.\n2. **Point 3**: It's a bit further back on the table.\n3. **Point 1**: It's on the table but a little further back than Point 3.\n4. **Point 2**: It's on the wall in the background.\n5. **Point 5**: It's also on the wall but a bit higher up.\n6. **Point 4**: It's on the wall to the left of Point 2.
</think>
<answer> 6, 3, 1, 5, 2, 4</answer>

Figure 5: Examples of the thinking process solving visual jigsaw tasks.

In this section, we provide an example thinking process of our model solving the visual jigsaw task in Fig 5. For the image jigsaw task, the model needs to identify the content for each image patch,

imagine the overall scene, and reason over the relationships between the image patches to solve the jigsaw puzzle. For the video jigsaw, the model identifies the content for each clip and reasons about the correct chronological order. As for the 3D jigsaw, the model grounds the point marker on the actual object, infers the distance between points and the camera, and compares the distances among the points.

A.2 Additional Experiments on 3D Jigsaw

Besides the depth ordering task for the 3D modality, we have also explored two variants of 3D jigsaw designs, which we detail below.

(1) View–Motion Matching. Given a scene captured from multiple camera poses, we randomly select one view as the anchor view, and sample several candidate views that differ from the anchor while also being diverse from one another. For each candidate, we construct a natural language description of the ego-motion from the anchor to that candidate (*e.g.* "move forward 2.0 meters and rotate left by 15°"). The model is provided with the anchor image, the shuffled candidate images, and the corresponding ego-motion descriptions, and is tasked with correctly matching each candidate view to its ego-motion description.

(2) BEV–Pose Matching. We render a bird's-eye-view (BEV) image of the scene and randomly select a set of candidate views with different camera poses. These camera poses are annotated on the BEV image with numerical identifiers. The model is then given the annotated BEV image and a shuffled set of candidate view images, and must correctly match each camera pose to its corresponding candidate view.

Table 6: **Evaluation of 3D Jigsaw Variants.** Comparison of depth ordering, view–motion matching, and BEV–pose matching tasks on 3D benchmarks.

Model	SAT-Real	3DSRBench	ViewSpatial	All-Angles	OmniSpatial	VSI-Bench	SPARBench	DA-2K
	test	test	test	test	test	test	tiny	test
Qwen2.5-VL-7B	48.66	57.42	36.52	47.56	42.66	37.74	35.75	54.45
Depth Ordering	64.00	58.13	38.62	49.06	45.99	40.64	38.31	71.56
View–Motion Matching	64.67	55.89	37.94	48.22	44.55	38.97	36.53	60.15
BEV–Pose Matching	62.00	57.17	36.69	48.22	44.22	38.78	34.31	58.99

Analysis and Results. Both tasks are intuitively reasonable, as they explicitly encourage the model to connect 2D visual observations with underlying 3D spatial configurations. However, our preliminary experiments (shown in Tab 6) show that these variants do not lead to significant improvements on downstream benchmarks, and overall underperform the depth ordering formulation. We hypothesize that this may be due to the relatively weak 3D perception and reasoning capability of current base MLLMs, which limits their ability to transfer the learned skills from these complex jigsaw formulations to downstream tasks. Exploring ways to strengthen this foundation and better exploit such 3D-aware self-supervised tasks remains an interesting direction for future work.

A.3 Visual Jigsaw Examples

The visual jigsaw task examples for the three modalities are provided in Fig. 6, Fig. 7, and Fig. 8.

A.4 Qualitative Examples

Some qualitative examples of the model trained with image, video, and 3D jigsaw tasks are shown in Fig. 9, Fig. 10, and Fig. 11.

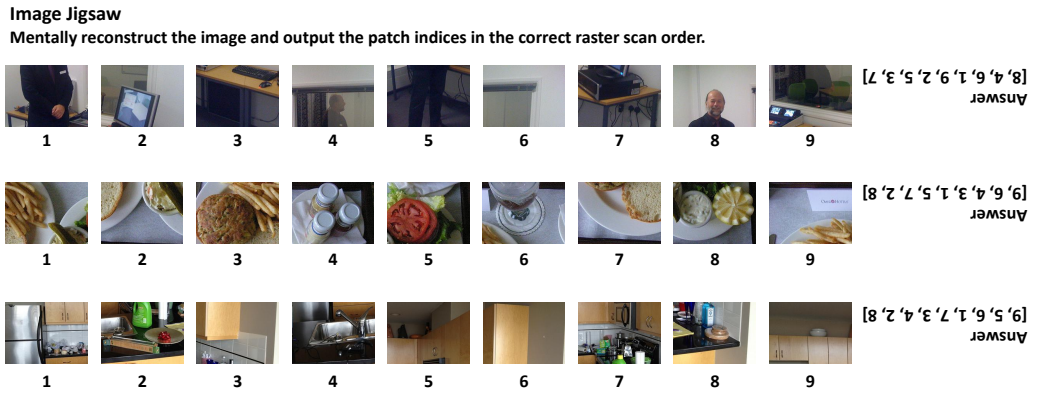


Figure 6: Examples of the image jigsaw task. Each row shows a shuffled set of patches from an image, where the model is required to reconstruct the correct raster scan order. The ground-truth answers are displayed on the right.

A.5 Limitations and Future Works

While our study demonstrates the effectiveness of Visual Jigsaw across images, videos, and 3D modalities, several limitations remain. First, for both image and video jigsaw, we adopt a standard and relatively simple formulation. Future work could explore more complex or hybrid jigsaw configurations, such as jointly partitioning video along spatial and temporal dimensions, or using heterogeneous piece sizes to introduce richer structural constraints. Second, due to computational constraints, we have not scaled the training data and model size extensively; investigating the scalability of this self-supervised task remains a promising direction. Third, some of our 3D jigsaw variants did not yield the expected improvements. We believe that applying these tasks to base models with stronger 3D reasoning capabilities and richer 3D priors could unlock further potential. Finally, beyond jigsaw, it is worth exploring a broader range of self- and weakly-supervised vision-centric tasks to enhance the perceptual and reasoning abilities of multimodal large language models.

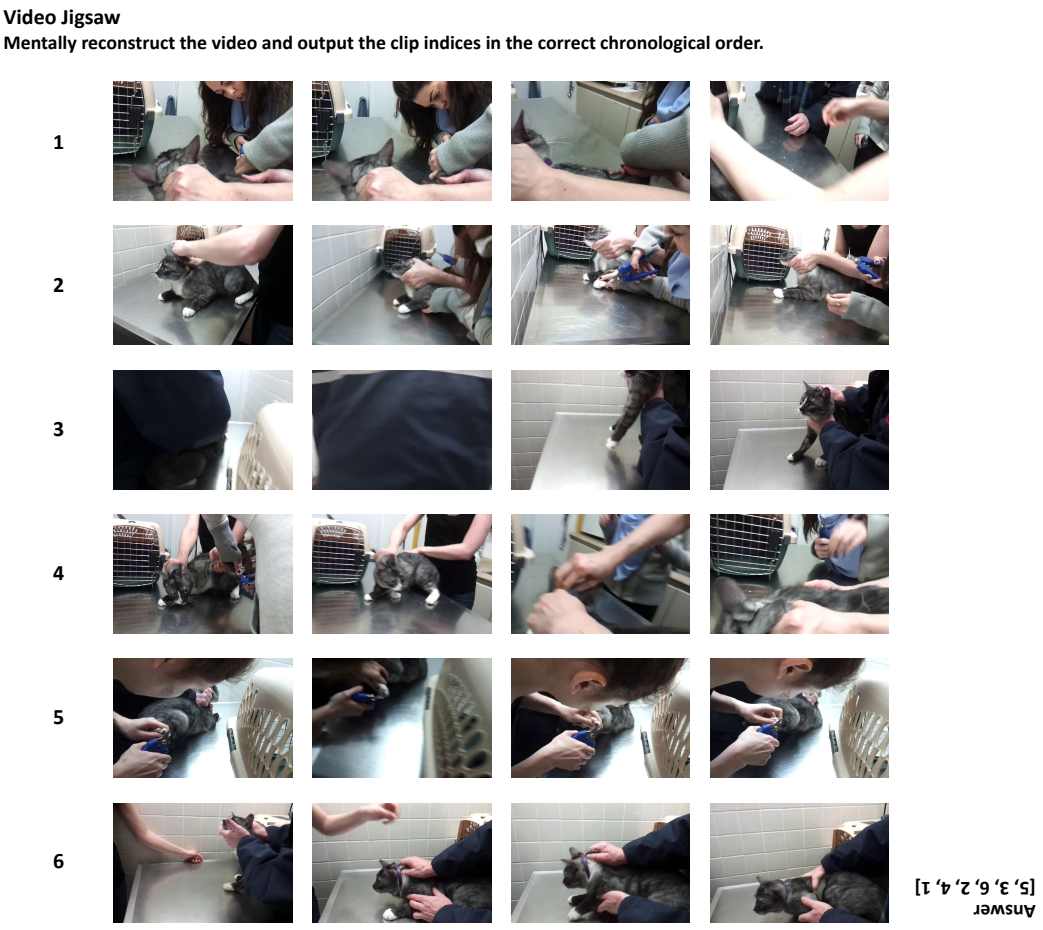


Figure 7: Example of the video jigsaw task. Each row shows a clip from the original video image, and the 6 clips are shuffled. The model is required to reconstruct the correct chronological order. The ground-truth answers are displayed on the right.

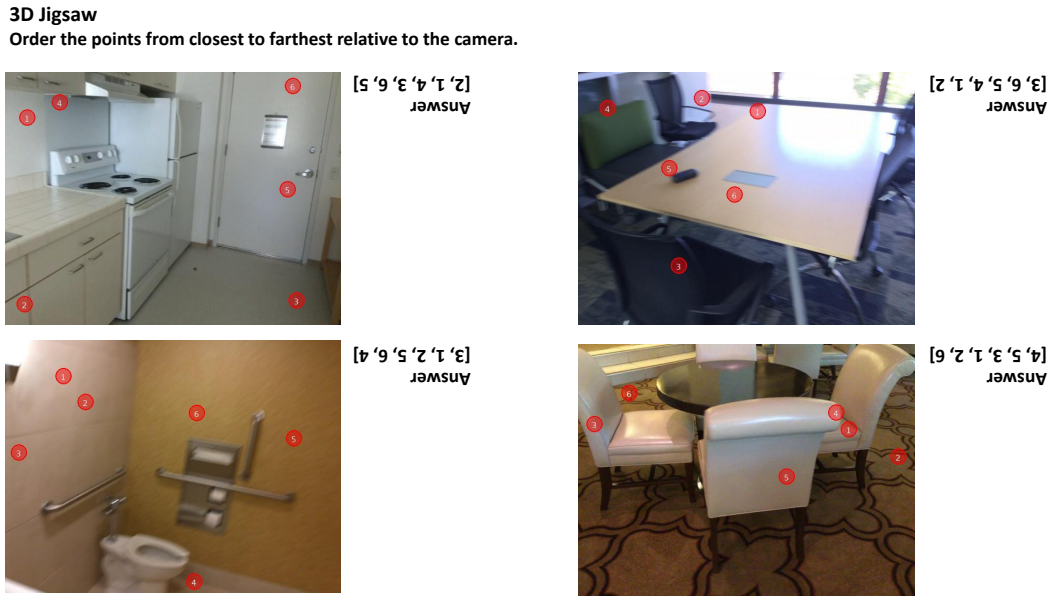


Figure 8: Examples of the 3D jigsaw task. The model is required to order the points in each image from closest to farthest relative to the camera. The ground-truth answers are displayed on the right.

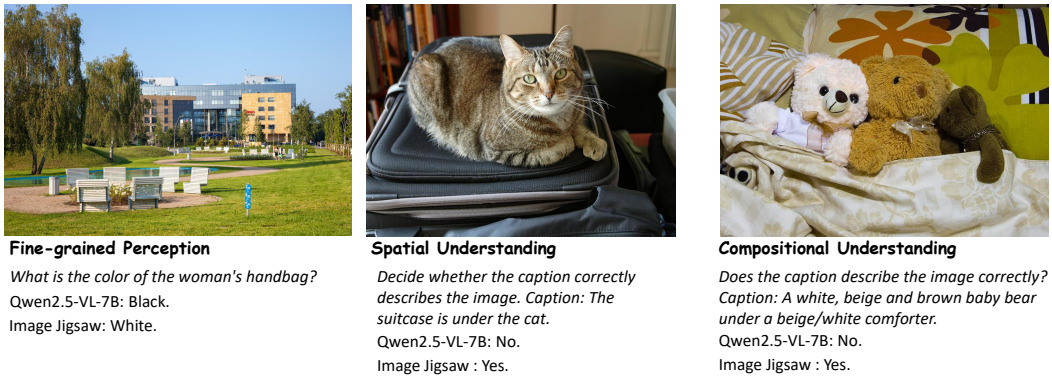


Figure 9: Qualitative examples on image tasks.

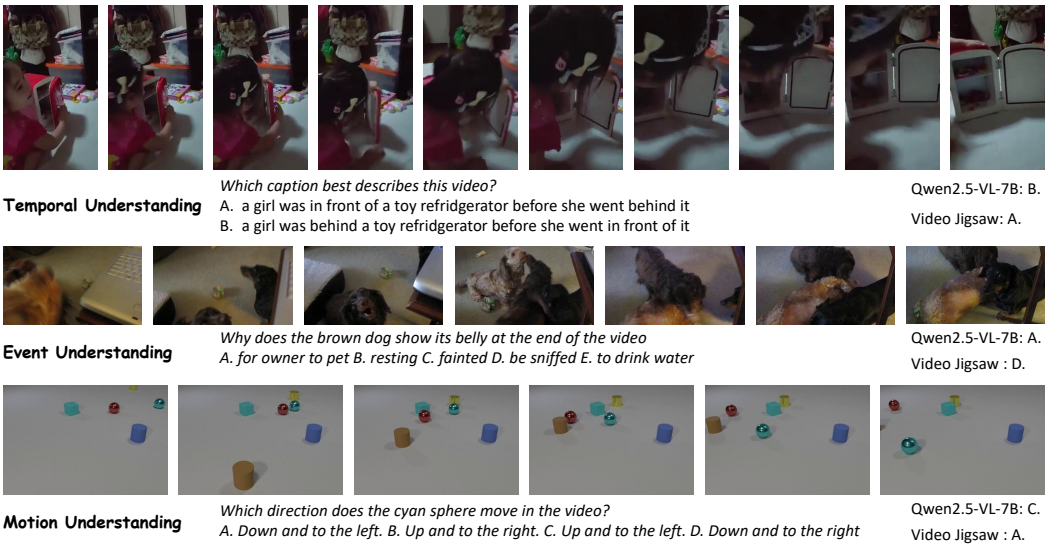


Figure 10: Qualitative examples on video tasks.

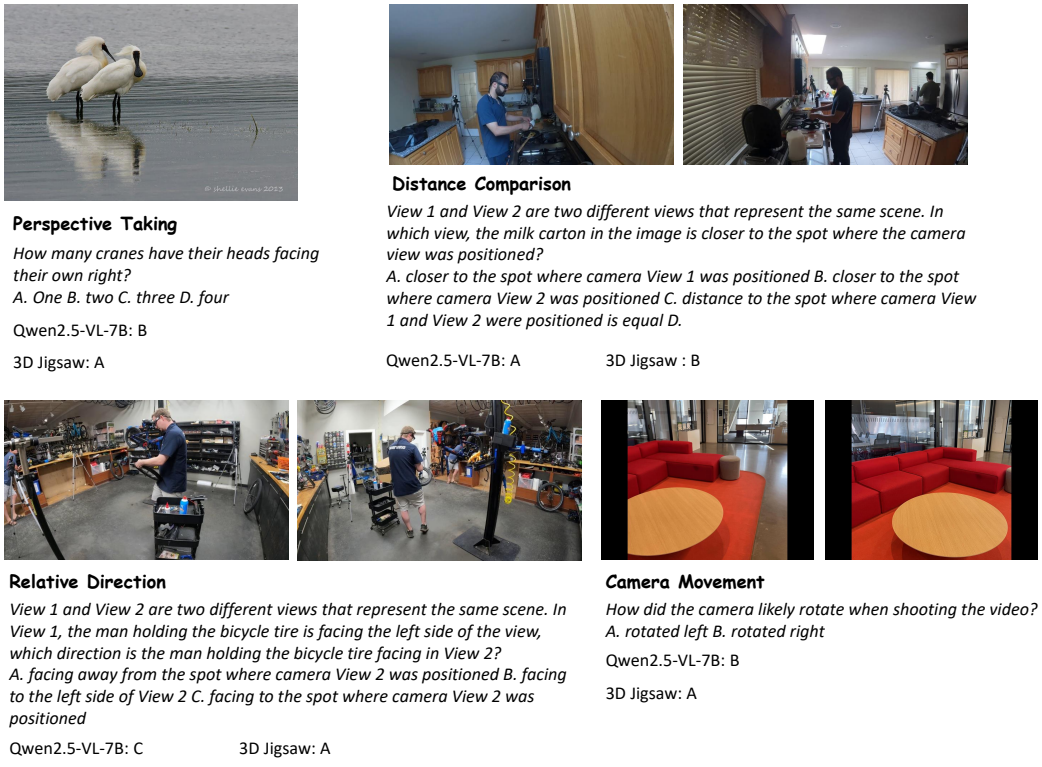


Figure 11: Qualitative examples on 3D tasks.

A.6 TASK PROMPTS

Prompt for Image Jigsaw

You are given nine shuffled image tiles that were created by slicing one image into a 3*3 grid.

Here are the tiles, each tagged with an index reflecting the current (shuffled) order in which they are shown:

Tile 1: <image>
Tile 2: <image>
Tile 3: <image>
Tile 4: <image>
Tile 5: <image>
Tile 6: <image>
Tile 7: <image>
Tile 8: <image>
Tile 9: <image>

Task:
Mentally reassemble the original image, arranging the tiles into the correct 3*3 layout and provide the tile indices in raster-scan order (left-to-right, top-to-bottom), separated by commas.

Answer format example:
5, 1, 3, 7, 9, 2, 4, 8, 6
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE put within <answer></answer> tags.

Prompt for Video Jigsaw

You are given four **shuffled** video clips that were created by slicing one original video into 6 equal-length temporal segments.

Here are the clips, each tagged with an index reflecting the current (shuffled) order in which they are shown:

Clip 1: <video>
Clip 2: <video>
Clip 3: <video>
Clip 4: <video>
Clip 5: <video>
Clip 6: <video>

Task:
1. Mentally reassemble the original video by arranging the clips in their correct chronological order (earliest segment first, latest segment last).
2. Output the clip indices in that order, separated by commas.

Answer format example:
2, 3, 1, 4, 6, 5
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE put within <answer></answer> tags.

Prompt for 3D Jigsaw

<image>
You are given an indoor RGB image. Six points are marked on the image with red circular labels (1, 2, 3, . . .).

Your task is to order the points from closest to farthest relative to the camera, judging the distance based on the center of the red circular marker.

Answer with the ordered sequence of point numbers.
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE put within <answer></answer> tags.