POMA-3D: The Point Map Way to 3D Scene Understanding

Supplementary Material

A. ScenePoint Language Generation Details

The proposed ScenePoint dataset provides both scene-level and view-level textual annotations. The scene-level captions are directly inherited from the original SceneVerse [24] dataset. For view-level captions, we generate descriptions using InternVL3-14B [58], following the prompting strategy illustrated in Fig. 6. Each view is annotated with 15 diverse, raw captions that comprehensively describe the visible objects, their attributes, and spatial relations. The prompt also includes four in-context exemplars to guide consistent caption generation. As shown in Fig. 8, the generated captions exhibit high levels of detail, accuracy, and completeness.

View-level Caption Generation

You are given an image of an indoor scene. Your task is to generate captions that are 100% accurate to the image, with no hallucination.

Instructions:
- Generate exactly 15 different CLIP-style captions for the image.
- Captions must clearly describe the visible objects, their attributes, and their spatial relationships.
- Use spatial prepositions such as: on, under, next to, beside, behind, in front of, between, above, below.
- Focus only on what is visible in the image. Do not speculate or add details that are not present.
- Be precise and factual. Avoid opinions, emotions, or subjective language.

Examples:
1. Gray laptop centered on desk, coffee mug to the side, rolling chair positioned behind.
2. Light gray couch, wooden table placed in front, standing lamp on the left.
3. White bed with pillows, small wooden nightstand beside, bedside lamp on top.
4. Rectangular table with two chairs around, bright window in background, curtain partly covering.

Now, generate 15 captions following these rules. Output must be a numbered list (1. ... 2. ... up to 15.).

Figure 6. View-level caption generation prompt used by InternVL3.

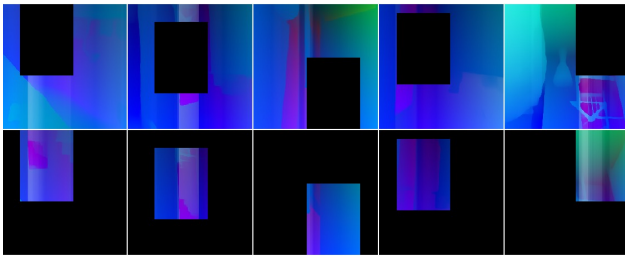

Figure 7. Visualization of the POMA-JEPA masking strategy. The first row shows the context blocks that remain visible for the context encoder. The second row depicts the target blocks that are masked out and whose embeddings need to be reconstructed.

B. More Implementation Details

B.1. Pretraining Details

In the view-to-scene vision–language alignment stage, the point-map context encoder is trained to align with the FG-CLIP [50] text encoder. Because scene-level captions are substantially longer than view-level descriptions, we set the maximum token length to 77 for view-level captions and 248 for scene-level captions. All trainable components of POMA-3D use a learning rate of 1×10^{-4} with cosine annealing and a minimum learning-rate ratio of 0.1. For POMA-JEPA, we adopt a linear EMA schedule that increases the momentum from 0.996 to 1.0 over all training steps. The warmup stage requires approximately 1.5 days on $8 \times$ A100 GPUs, and the main stage completes in an additional 1 day.

B.2. Downstream Finetuning Details

3D QA & Embodied Navigation. Following the SceneVerse [24] protocol, POMA-3D$_{\text{spec}}$ is finetuned on each 3D QA dataset for 100 epochs with a learning rate of 1×10^{-4}. The QA head architecture is the same as SceneVerse. For POMA-3D$_{\text{llm}}$, we modify LLaVA-OV-7B [27] by replacing its visual encoder with POMA-3D, allowing the model to take multi-view point maps as input. Both the LLM and the projector are finetuned on the 3D QA datasets. The entire finetuning procedure runs for one epoch using LoRA with rank 512, $\alpha = 1024$, and a learning rate of 1×10^{-4}.

Scene Retrieval & Coarse Localization. All evaluated methods for scene retrieval and coarse localization are assessed in a zero-shot manner using their aligned 3D encoder and language encoder. For the scene retrieval task, the number of candidate scenes is 397 for ScanRefer [8], 641 for Nr3D, and 586 for Sr3D [1], respectively.

B.3. POMA-JEPA Masking Details

Fig. 7 illustrates the masking strategy used in POMA-JEPA. The first row shows the context blocks sampled from five views of point-map inputs for a single scene, which are fed into the context encoder. The second row displays the corresponding target blocks, whose embeddings the predictor network is tasked with reconstructing.

C. More Experimental Results

C.1. Effect of Chamfer Distance

In our method section, we describe $\mathcal{L}_{\text{pjepa}}$ as a Chamfer Distance objective, due to the order-agnostic nature of point

Figure 8. Examples of image views alongside their generated view-level captions.

Figure 9. Visualization of the top-6 coarse localization results for the query "I am handling some paperwork." The top 1–4 views focus on office desk regions, while the top 5–6 views shift attention toward dining table areas that could serve as alternative workspaces.

maps in world space. The results in Tab. 7 further quantitatively validate the effectiveness of using Chamfer distance over MSE loss, observing consistent improvements across all three QA benchmarks.

C.2. Coarse Localization

Fig. 9 presents additional top-6 coarse localization results of POMA-3D for the situational text query "I am handling some paperwork." Remarkably, the top-4 retrieved point-map views focus on the office-desk region, which is the most appropriate location for such an activity. The 5th–6th views correspond to the dining-table areas, which could reasonably serve as alternative workspaces. These results demonstrate that our model can effectively prioritize and localize regions most relevant to the query, even when it contains no explicit object references.

Table 7. Evaluation of POMA-JEPA under two reconstruction losses (MSE vs. Chamfer Distance) on ScanQA [4], SQA3D [33], and Hypo3D [36].

POMA-JEPA	ScanQA	SQA3D	Hypo3D
MSE	21.8	50.8	32.8
Chamfer Distance	**22.3**	**51.1**	**33.4**

Table 8. Ablation study of POMA-3D$_{\text{llm}}$ on SQA3D [33] dataset.

Warmup	Main			EM@1
$\mathcal{L}_{\text{view}}$	$\mathcal{L}_{\text{view}}$	$\mathcal{L}_{\text{scene}}$	$\mathcal{L}_{\text{pjepa}}$	
✗	✗	✗	✗	48.5
✓	✗	✗	✗	49.5
✓	✓	✓	✗	51.0
✓	✓	✓	✓	**51.6**

Table 9. Ablation Study on Scene Retrieval Task.

	ScanRefer		Sr3D	
Setting	R@1-5	R@5-5	R@1-5	R@5-5
w/o Warmup	26.3	57.1	15.2	41.2
w/o view-align	25.6	58.0	13.7	41.3
w/o scene-align	24.8	56.0	13.0	38.7
w/o POMA-JEPA	25.8	57.2	13.5	36.1
Full	27.9	59.4	15.7	42.2

C.3. Additional Ablations on Scene Retrieval Task

Tab. 9 reports ablation results on scene retrieval, showing that components for global modeling (scene-level alignment and POMA-JEPA) yield the largest gains.

C.4. Additional Ablations in Generalist 3D Model

In the main manuscript, we present ablation studies of POMA-3D in the context of building a specialist model POMA-3D$_{\text{spec}}$. Here, we provide additional results for constructing the generalist 3D model POMA-3D$_{\text{llm}}$, as shown in Tab. 8. On SQA3D, all proposed modules contribute to performance gains. Notably, the warmup stage yields a 1% improvement, underscoring the importance of incorporating 2D image scenes during pretraining.