KAIST AI

Imagine3D-LLM: Teaching MLLMs to Imagine 3D Scenes Before Answering

Jaewoo Jung[1,2,*], **Hyeonseo Yu**[1], **Honggyu An**[1], **Jisang Han**[1], **Mungyeom Kim**[1], **Minkyeong Jeon**[1], **Heeseong Shin**[1], **WonJun Moon**[1], **Federico Tombari**[3,4], **Daniel Barath**[2], **Marc Pollefeys**[2,6,†], **Seungryong Kim**[1,†] **and Sunghwan Hong**[2,5,†]

[1]KAIST AI, [2]ETH Zürich, [3]Google, [4]TUM, [5]ETH AI Center, [6]Microsoft
[*]Work done during a visiting researcher period at ETH Zürich.

https://cvlab-kaist.github.io/Imagine3D-LLM

Reasoning about the 3D world from multi-view images remains a fundamental challenge for Multimodal Large Language Models (MLLMs). While modern MLLMs handle single-image inputs effectively, they struggle to integrate evidence across viewpoints into a coherent 3D understanding. A growing body of work attempts to close this gap by injecting 3D awareness into MLLMs, either by boosting fine-grained pixel-level cross-view correspondence or by fusing features from 3D geometry foundation models, yet a substantial gap to human reasoning persists. In this work, we revisit human spatial reasoning, which suggests that rather than relying on fine-grained geometry cues, humans roughly identify common objects across views, infer the relative geometry between viewpoints, and assemble a coarse 3D layout of the scene. Inspired by this process, we introduce **Imagine3D-LLM**, an MLLM that learns to assemble a similar compact 3D representation of the scene and conditions its answer on this representation. Concretely, we append a small set of learnable summary tokens after the image tokens, decode them into a compact 3D Gaussian Splatting representation supervised by a photometric reconstruction loss, and train jointly with the standard next-token prediction objective. Notably, although only the summary tokens receive direct reconstruction supervision, this objective also induces stronger cross-frame correspondence within the LLM's underlying image features, suggesting that learning to reconstruct propagates 3D-aware signals throughout the model. As a result, Imagine3D-LLM consistently outperforms prior approaches across multiple spatial reasoning and 3D understanding benchmarks, suggesting that imagining the scene can be more effective than being told its pixel-wise geometry.

1. Introduction

Reasoning about 3D structure from multi-view images is a fundamental problem in computer vision, underpinning a wide range of downstream applications such as robotics (Shen et al., 2023), embodied AI (Zhao et al., 2023), AR/VR (Brachmann et al., 2023), and scene understanding (An et al., 2026). Driven by rapid progress in Multimodal Large Language Models (MLLMs) (Yang et al., 2025; Bai et al., 2025; Hurst et al., 2024; Team et al., 2024), machines can now interpret a single image or a continuous video stream at a level that often rivals, and in some cases surpasses, human performance (Yue et al., 2024; Lu et al., 2024). However, when the input shifts from a single viewpoint to a set of multi-view images that demand genuine 3D reasoning, even the largest and most capable models fall short of human-level competence (Yang et al., 2024b).

A growing body of work has attempted to close this gap by injecting strong 3D priors into MLLMs. One line of research implicitly or explicitly boosts pixel-level correspondences across views by either embedding scene point clouds' 3D coordinates into patch-level features (Zheng et al., 2025b; Zhu et al., 2024; Wang et al., 2025a), augmenting images with explicit visual markers (Qi et al., 2025), distilling features with accurate correspondences (Huang et al., 2025), or designing pretext tasks that encourage the model to align overlapping content across frames (Wang et al., 2025a). A complementary line (Zheng et al., 2025a; Fan et al., 2025; Hu et al., 2025) fuses MLLM image features with ones from recently developed 3D reconstruction foundation models such as CUT3R (Wang et al., 2025c) and VGGT (Wang et al., 2025b). While both approaches attempt to improve 3D awareness of MLLMs with *fine-grained pixel-level* 3D signals, these approaches yield only incremental gains (Yang et al., 2024b; Zhang et al., 2025b).

In this work, we take a step back and ask whether such *fine-grained* 3D signals are truly the most direct route to improved 3D reasoning. To answer this, we revisit how humans actually perceive 3D structure from multi-view

observations. Cognitive studies suggest that humans do not maintain pixel-perfect correspondences or dense depth maps in their heads (Burgess, 2006; Shepard and Metzler, 1971). When presented with several views of a scene, humans instead identify common objects across views, use them as anchors to infer *coarse and abstract* spatial relationships between viewpoints, and mentally assemble a compact layout of the scene in 3D. Therefore, we hypothesize that the representation that supports their downstream reasoning is rather object-level and approximate, not pixel-accurate.

Inspired by this human thinking process, we present **Imagine3D-LLM**, a framework that enables an MLLM to build an analogous *compact 3D reconstruction* of the scene before answering. To enable this, we introduce a *compact* set of learnable **Gaussian summary tokens**, which are appended to the sequence of image tokens fed into the LLM. From these tokens, we decode the parameters of 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) primitives, so that each Gaussian summary token corresponds to a small set of 3D Gaussians that together explain a portion of the scene.

To efficiently and effectively enable MLLMs to build an abstract and compact reconstruction of the scene, our Gaussian estimation pipeline incorporates two deliberate design choices. **(1)** To prevent the model from simply copy-pasting 2D image content into image-aligned Gaussians rather than recovering the underlying geometry, we predict the Gaussians from the dedicated Gaussian summary tokens instead of predicting them directly from the image features. **(2)** We impose an information bottleneck by using fewer Gaussian summary tokens than image tokens, which compels overlapping content across views to be merged into a shared set of tokens, encouraging the model to reason about which objects recur across views and how they fit together within a unified 3D layout.

We jointly train Imagine3D-LLM with a photometric reconstruction loss on the rendered Gaussians and the standard next-token prediction loss for language modeling. Surprisingly, although only the Gaussian summary tokens receive direct reconstruction supervision, our analysis shows that the LLM's image features themselves become more 3D-aware: corresponding image features show more sharp attentions, where PCA visualizations further reveal that image features become more semantically organized, with the same object encoded consistently across views. These analyses validate that requiring the model to infer the underlying structure of the scene propagates 3D-aware signals throughout the LLM, reshaping its internal representations toward a coherent, object-centric 3D understanding. As a result, Imagine3D-LLM achieves significant performance gains across seven diverse spatial reasoning and 3D scene understanding benchmarks, suggesting that abstract scene reconstruction is a powerful inductive bias for building 3D-aware MLLMs, and that imagining the scene before answering can be more effective than being told its pixel-wise geometry.

2. Related Works

3D Scene Understanding. Grounding natural language in 3D environments is a long-standing goal of the 3D scene understanding community, supporting tasks such as 3D visual grounding (Chen et al., 2020; Zhang et al., 2023), 3D dense captioning (Chen et al., 2021, 2023, 2024b), and 3D question answering (Azuma et al., 2022; Ma et al., 2022, 2026; Yang et al., 2024b). A prominent line of work distills features from 2D foundation models (most commonly CLIP (Radford et al., 2021) or LSeg (Li et al., 2022)) into a reconstructed scene representation, including point clouds (Yang et al., 2024a; Zhou et al., 2025a; Jatavallabhula et al., 2023), NeRFs (Kerr et al., 2023; Engelmann et al., 2024), and more recently 3D Gaussian Splatting (Jun-Seong et al., 2025; Qin et al., 2024; Shi et al., 2024), enabling open-vocabulary querying by matching distilled embeddings to a text prompt. These pipelines, however, typically rely on task-specific architectures and per-scene optimization, and their language understanding is bounded by the compositional capacity of the underlying 2D embedding model. We instead extend the language and visual reasoning of pretrained MLLMs to multi-view inputs, allowing a single generalist model to address these tasks through natural language without per-task or per-scene optimization.

3D Multimodal Large Language Models (MLLMs). Motivated by the success of 2D MLLMs (Bai et al., 2025; Yang et al., 2025; Hurst et al., 2024; Li et al., 2024; Lin et al., 2024; Liu et al., 2023; Team et al., 2024; Tong et al., 2024; Wang et al., 2024a; Yoon et al., 2025), a growing body of work extends them to 3D understanding. Early efforts feed explicit 3D inputs to the LLM, either by lifting 2D features into 3D (Zhu et al., 2024; Hong et al., 2023) or by training dedicated point-cloud encoders as an additional modality (Chen et al., 2024a; Wang et al., 2023; Huang et al., 2024b; Xu et al., 2024); the scarcity of paired 3D–language data, however,

has limited their performance on language-heavy tasks. More recent approaches retain the multi-view 2D input format and inject 3D awareness through training objectives or auxiliary modules. One family supplies pixel-level 3D signals to encourage cross-view correspondence: Video-3D-LLM (Zheng et al., 2025b) and LLaVA-3D (Zhu et al., 2024) attach 3D positional embeddings derived from input point clouds, GPT4Scene (Qi et al., 2025) renders bird's-eye-view markers, Ross3D (Wang et al., 2025a) adds a cross-view reconstruction pretext task, and 3DRS (Huang et al., 2025) distills features with explicit correspondence supervision. A second family fuses MLLM image features with representations from 3D foundation models (Wang et al., 2025c,b) to expose the LLM to dense, geometry-aware features, as in VLM3R (Fan et al., 2025), G^{2}VLM (Hu et al., 2025), and the concurrent 3DThinker (Chen et al., 2025). Despite steady progress, a substantial gap to human-level performance on multi-view 3D reasoning benchmarks remains (Yang et al., 2024b).

Compact Scene Representations. Cognitive studies suggest that humans do not maintain pixel-accurate reconstructions of every surface; rather, they rely on coarse abstractions that are nonetheless sufficient to support rich spatial reasoning (Burgess, 2006; Shepard and Metzler, 1971; Lee et al., 2025). A line of work in 3D vision has correspondingly explored decomposing scenes into compact geometric primitives such as meshes, polygons, or superquadrics (Tulsiani et al., 2017; Paschalidou et al., 2019, 2020; Fedele et al., 2025; Monnier et al., 2023), but these methods either struggle to scale beyond a handful of primitives (Monnier et al., 2023) or rely on off-the-shelf 3D instance segmentation (Fedele et al., 2025). More recently, C3G (An et al., 2026) introduced a feed-forward pipeline that represents a scene compactly with 3D Gaussian Splatting (3DGS) parameters, decoding 3D Gaussians from a small set of learnable query tokens supervised purely through photometric rendering. Inspired by this line of work, we bring token-level Gaussian decoding inside an MLLM, where a small set of *Gaussian summary tokens* learns to assemble a compact 3D representation of the scene that the LLM conditions on when reasoning about it.

3. Methodology

3.1. Preliminaries: Multi-image MLLMs

Following prior works (Wang et al., 2025a; Zheng et al., 2025b; Zhu et al., 2024), we build our model on top of MLLMs capable of taking multiple images as input. Here, we briefly explain the input formulation of such MLLMs.

A multi-image MLLM is composed of a vision encoder $V_\psi(\cdot)$, a vision–language projector $P_\phi(\cdot)$, and a pre-trained large language model $LM_\theta(\cdot)$, where ψ, ϕ, and θ denote the corresponding parameters. Given a set of K input frames $\mathcal{I} = \{I_k\}_{k=1}^{K}$ with $I_k \in \mathbb{R}^{H\times W\times 3}$, the vision encoder independently extracts patch-level features from each frame:

$$\mathbf{z}_k = V_\psi(I_k) \in \mathbb{R}^{N\times D_\mathbf{z}}, \quad k = 1, \ldots, K, \tag{1}$$

where N is the number of patches per frame and $D_\mathbf{z}$ is the visual feature dimension. Each frame's features are then projected into the language model's embedding space of dimension D via $P_\phi(\cdot)$, yielding per-frame visual tokens $\mathbf{e}_k^{\text{img}} = P_\phi(\mathbf{z}_k) \in \mathbb{R}^{N'\times D}$, where N' may differ from N due to spatial token-reduction operations such as bilinear pooling and per-row newline insertion that preserve the 2D spatial layout (Zhang et al., 2024). Concatenating all K frames produces the full visual token sequence:

$$\mathbf{e}^{\text{img}} = [\mathbf{e}_1^{\text{img}}; \mathbf{e}_2^{\text{img}}; \ldots; \mathbf{e}_K^{\text{img}}] \in \mathbb{R}^{KN'\times D}. \tag{2}$$

The input text (e.g., the user instruction) is tokenized and embedded into the same space through the language model's embedding layer, producing textual embeddings $\mathbf{e}^{\text{text}} \in \mathbb{R}^{L\times D}$, where L denotes the number of text tokens. The language model then takes the concatenated multimodal sequence $[\mathbf{e}^{\text{img}}; \mathbf{e}^{\text{text}}] \in \mathbb{R}^{(KN'+L)\times D}$ as input and models the causal distribution over the text tokens as:

$$p_{\theta,\phi}(\mathbf{e}_{1:L}^{\text{text}} \mid \mathbf{e}^{\text{img}}) = \prod_{i=1}^{L} p_{\theta,\phi}\big(\mathbf{e}_i^{\text{text}} \,\big|\, \mathbf{e}_{<i}^{\text{text}}, \mathbf{e}^{\text{img}}\big). \tag{3}$$

At inference, the language model autoregressively generates text tokens conditioned on the visual tokens and the previously generated text tokens.

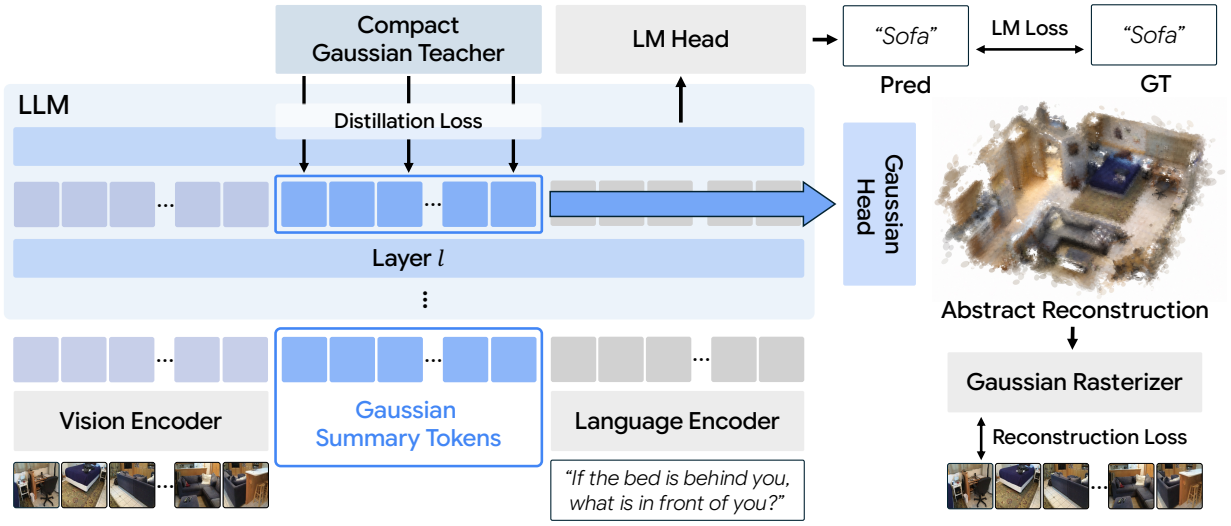

Figure 1: **Overall architecture of Imagine3D-LLM.** Given multi-view images of a scene and a text instruction, our model appends a compact set of learnable *Gaussian summary tokens* between the image and text tokens. At an intermediate layer ℓ, the hidden states at the Gaussian summary token positions are decoded into a compact 3D Gaussian Splatting representation that abstractly reconstructs the scene. The model is trained jointly with a standard language modeling loss on the text output, a photometric reconstruction loss obtained by rendering the predicted Gaussians at the input viewpoints, and a distillation loss from a pretrained compact Gaussian teacher (Veicht et al., 2026).

3.2. Imagine3D-LLM

We now describe how we equip a multi-image MLLM with the ability to perform mental 3D reconstruction of the input scene for improved 3D reasoning. At a high level, we introduce a compact set of learnable *Gaussian summary tokens* that are appended to the image tokens fed into the LLM (§ 3.2.1). These tokens learn to summarize the multi-view observations into a compact set of 3D Gaussians, jointly trained with the reconstruction and standard language modeling objective. Since pure photometric supervision is known to require lengthy training schedules even for learning Gaussians alone (Chen et al., 2024d; Charatan et al., 2024; Ye et al., 2024; Hong et al., 2024; Jiang et al., 2025a; An et al., 2026), we further introduce a distillation objective from a pretrained Gaussian teacher model (An et al., 2026), to substantially accelerate convergence (§ 3.2.2). Finally, we analyze how this joint training improves the MLLM's internal representations (§ 3.3). An overview of our method is shown in Fig. 1.

3.2.1. Mental Reconstruction via Gaussian Summary Tokens

Gaussian summary tokens. Given the visual token sequence $\mathbf{e}^{\text{img}} \in \mathbb{R}^{KN' \times D}$ produced by the vision encoder and projector, we introduce a compact set of M learnable Gaussian summary tokens $\mathbf{e}^{\text{gs}} \in \mathbb{R}^{M \times D}$, where $M \ll KN'$. These tokens are inserted between the visual tokens and the textual tokens, yielding the augmented multimodal sequence

$$[\mathbf{e}^{\text{img}}; \mathbf{e}^{\text{gs}}; \mathbf{e}^{\text{text}}] \in \mathbb{R}^{(KN'+M+L) \times D}, \tag{4}$$

which is processed by the LLM via standard causal self-attention. By placing the Gaussian summary tokens *after* the image tokens, each token can attend to all visual evidence and selectively retrieve the information needed to faithfully describe a portion of the 3D scene, rather than being passively bound to local image content.

Bottleneck-induced object grouping. A central design choice is that the number of Gaussian summary tokens is substantially smaller than the number of image tokens, $M \ll KN'$. This bottleneck prevents the model from naively allocating one token per pixel or per patch, and instead forces overlapping content observed across multiple views to be merged into a shared set of tokens. Under this constraint, the most efficient way for the model to faithfully reconstruct the scene with limited Gaussians is to assign each token to a coherent 3D region or object that recurs across views. As we show in § 3.3, this bottleneck gives rise to an emergent object-centric grouping, mirroring the way humans build object-level mental abstractions of a scene before

reasoning about it. Since our main objective is to improve 3D reasoning rather than novel view synthesis, this design also offers a favorable trade-off between reconstruction quality and computational cost, preventing the input sequence length to increase substantially.

Decoding 3D Gaussians. Given the LLM forward pass over the augmented sequence $[\mathbf{e}^{\text{img}}; \mathbf{e}^{\text{gs}}; \mathbf{e}^{\text{text}}]$, a natural question is from which layer the Gaussian summary tokens should be decoded. Recent analyses (Kaduri et al., 2025; Zhang et al., 2025c; Jiang et al., 2025b; Kang et al., 2025; Yoon et al., 2025) consistently identify that the *middle layers* of the LLM most actively show information flow between visual and text modalities. Since our Gaussian summary tokens also require effectively absorbing visual information from the image tokens, we extract the hidden states at the Gaussian summary token positions from the middle layer ℓ, denoted $\mathbf{h}^{\text{gs}} \in \mathbb{R}^{M\times D}$. Decoding at a middle layer also leaves enough subsequent layers for the text tokens to attend to the Gaussian summary tokens, enabling the text tokens to benefit from the scene-summarizing information encoded by the Gaussian tokens. The effectiveness of choosing ℓ as a middle layer is further validated in Tab. 5.

The extracted hidden states $\mathbf{h}^{\text{gs}}$ are further decoded into 3D Gaussian primitives through a lightweight MLP-based Gaussian head $\mathcal{H}(\cdot)$. Rather than mapping each token to a single Gaussian, we let each token decode G Gaussians simultaneously, providing additional representational capacity per token without expanding the LLM's sequence length:

$$\{\mathbf{G}_{i,j}\}_{j=1}^{G} = \mathcal{H}(\mathbf{h}_i^{\text{gs}}), \quad i = 1, \ldots, M, \tag{5}$$

yielding a total of $M \times G$ Gaussians per scene. Each Gaussian $\mathbf{G}_{i,j} = \{\mu_{i,j}, \sigma_{i,j}, \Sigma_{i,j}, c_{i,j}\}$ is parameterized by its 3D center $\mu \in \mathbb{R}^3$, opacity $\sigma \in [0, 1)$, covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$, and spherical harmonics coefficients $c \in \mathbb{R}^{3(L_{\text{SH}}+1)^2}$ that encode view-dependent color with L_{SH} degrees. Together, the resulting set $\{\mathbf{G}_{i,j}\}$ forms a compact 3D representation of the scene that can be rendered into novel views via differentiable rasterization (Kerbl et al., 2023).

Training objective. We jointly train the model with two losses. The standard *language modeling loss* $\mathcal{L}_{\text{LM}}$ supervises the autoregressive prediction of text tokens conditioned on the visual and Gaussian summary tokens, while the *reconstruction loss* $\mathcal{L}_{\text{recon}}$ supervises the decoded Gaussians by rendering them at the input viewpoints $\{\pi_k\}_{k=1}^{K}$ via differentiable rasterization and comparing the resulting images to the ground-truth frames I_k:

$$\mathcal{L}_{\text{LM}} = -\sum_{i=1}^{L} \log p_{\theta,\phi}\big(\mathbf{e}_i^{\text{text}} \,\big|\, \mathbf{e}_{<i}^{\text{text}}, \mathbf{e}^{\text{img}}, \mathbf{e}^{\text{gs}}\big), \tag{6}$$

$$\mathcal{L}_{\text{recon}} = \sum_{k=1}^{K} \Big[\lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}}(\hat{I}_k, I_k) + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(\hat{I}_k, I_k)\Big], \tag{7}$$

where $\hat{I}_k$ is the image rendered from $\{\mathbf{G}_i\}_{i=1}^{M}$ at viewpoint π_k. The combined objective is $\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}}$.

3.2.2. Accelerating Convergence via Distillation from a Compact Gaussian Teacher

Although the model can in principle be trained end-to-end with the objective above, we find that pure photometric supervision through the LLM converges slowly, requiring substantially more iterations than is practical given the cost of MLLM finetuning. This is consistent with prior observations in the feed-forward 3D Gaussian Splatting literature, where dedicated estimators are typically trained for hundreds of thousands of iterations with large batch sizes to obtain reliable Gaussian predictions (Chen et al., 2024d; Charatan et al., 2024; Ye et al., 2024; Hong et al., 2024; Jiang et al., 2025a; An et al., 2026). To circumvent this bottleneck, we leverage a pretrained compact Gaussian estimator as a teacher model and distill its scene representation into the LLM, providing a strong inductive signal for what the Gaussian summary tokens should encode.

Compact Gaussian teacher. We adopt ZipSplat (Veicht et al., 2026), a recent feed-forward 3DGS framework that estimates a compact set of 3D Gaussians from multi-view images, as our teacher. Given the same multi-view inputs $\mathcal{I}$, ZipSplat first encodes them with a geometry-grounded visual backbone (Wang et al., 2025b), and then refines a small set of M learnable query tokens $\mathbf{Q}^{\text{T}} \in \mathbb{R}^{M\times d_{\text{T}}}$ through a transformer that jointly attends over the queries and the multi-view features, where d_{T} denotes the teacher model's hidden dimension. The refined

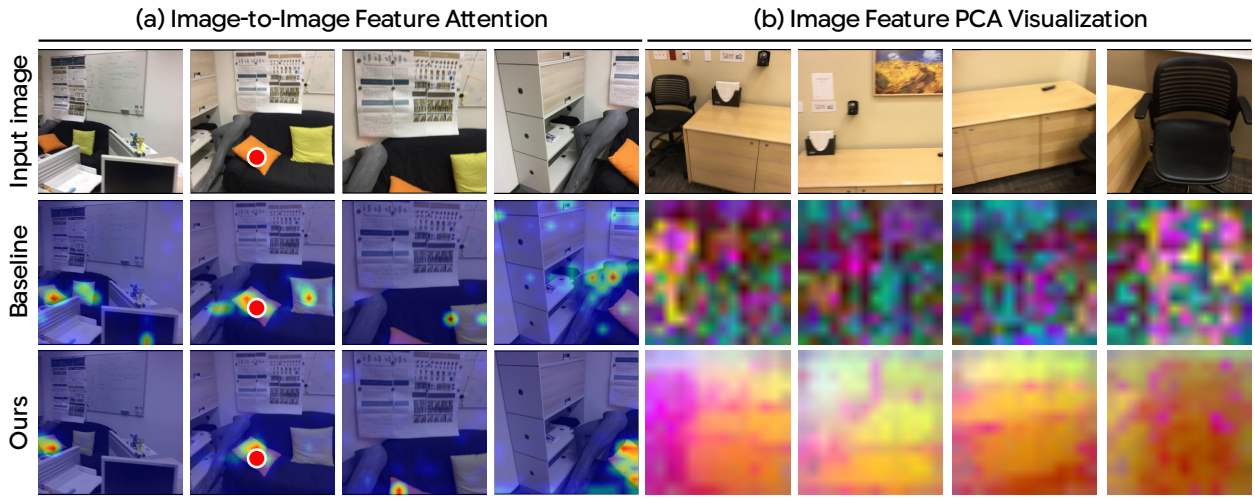

Figure 2: Effects of joint reconstruction training on the LLM's image features. (a) Image-to-image attention maps between a query point (red dot) and other input views. Our model exhibits substantially stronger cross-view correspondences than the baseline, indicating that the Gaussian summary tokens implicitly drive the image features to align across views. **(b)** PCA visualization of the LLM's image features (first three components shown as RGB). Our features are noticeably cleaner and more semantically structured, with corresponding objects (e.g., chairs, desks) encoded with consistent colors across views, reflecting the emergence of object-level abstractions.

tokens $\bar{\mathbf{Q}}^{\mathrm{T}} \in \mathbb{R}^{M \times d_{\mathrm{T}}}$ are then decoded into 3D Gaussians $\{\mathbf{G}^{\mathrm{T}}_{i,j}\}_{j=1}^{G}$ for $i = 1, \ldots, M$ through a lightweight Gaussian head. Crucially, we match the number of teacher tokens M in ZipSplat's inference time with the number of Gaussian summary tokens in our LLM, allowing direct token-level alignment between teacher and student.

Distillation loss. We distill from the teacher at two complementary levels. First, we align the LLM's hidden states at the Gaussian summary token positions, $\mathbf{h}^{\mathrm{gs}} \in \mathbb{R}^{M \times D}$, with the teacher's refined query tokens $\bar{\mathbf{Q}}^{\mathrm{T}}$ via a normalized cosine similarity loss with a lightweight projection layer $g(\cdot)$ that maps to the teacher's dimension d_{T}, $\mathcal{L}_{\mathrm{token}} = \frac{1}{M}\sum_{i=1}^{M}\left[1 - \cos\left(\frac{g(\mathbf{h}_i^{\mathrm{gs}})}{\|g(\mathbf{h}_i^{\mathrm{gs}})\|}, \frac{\bar{\mathbf{Q}}_i^{\mathrm{T}}}{\|\bar{\mathbf{Q}}_i^{\mathrm{T}}\|}\right)\right]$, where $\cos(\cdot,\cdot)$ denotes the cosine similarity operation and $\|\cdot\|$ is the L2-norm. This token-level supervision provides a dense, per-token target that guides the LLM toward representations the teacher's Gaussian head can already decode, bypassing the slow convergence of learning Gaussian-decodable features from photometric loss alone. Second, we additionally supervise the student's decoded Gaussian parameters against the teacher's predictions with an L2 loss over all Gaussian attributes $\mathcal{L}_{\mathrm{param}} = \frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{G}_i - \mathbf{G}_i^{\mathrm{T}}\right\|_2^2$, where $\mathbf{G}_i$ and $\mathbf{G}_i^{\mathrm{T}}$ are taken as the concatenation of all predicted Gaussian attributes from the student and teacher, respectively. The total distillation loss is $\mathcal{L}_{\mathrm{distill}} = \lambda_{\mathrm{token}}\mathcal{L}_{\mathrm{token}} + \lambda_{\mathrm{param}}\mathcal{L}_{\mathrm{param}}$.

Final objective. Combining the language modeling, reconstruction, and distillation losses, the full training objective of Imagine3D-LLM is:

$$\mathcal{L}_{\mathrm{full}} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{recon}}\mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{distill}}\mathcal{L}_{\mathrm{distill}}. \tag{8}$$

The teacher is kept frozen throughout training and is only used at training time.

3.3. Analysis

In this section, we further analyze the characteristics of the trained model to understand how jointly training the model to estimate coarse reconstructions of the scene leads to improved 3D reasoning.

Joint training improves image feature representations. While our reconstruction loss is applied only at the Gaussian summary tokens and does not directly supervise the image features, we observe that the LLM's

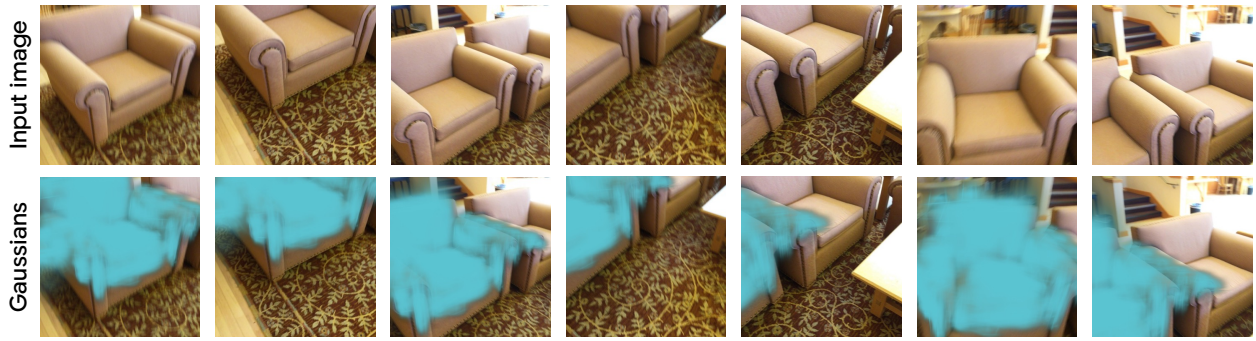

Figure 3: **Visualization of clustered Gaussian from summary tokens.** We visualize the Gaussians clustered from a set of Gaussian summary tokens via K-means over the summary tokens (cyan indicates the Gaussians belonging to one cluster). Similar summary tokens are mapped to the same object without any explicit clustering supervision.

image features themselves become substantially more 3D-aware after joint training. Fig. 2-(a) visualizes the attention maps between a query point in one view (red dot) and the image features of the other input views. Compared to the baseline, our model produces noticeably sharper and more spatially focused attention on the corresponding region across views, indicating that the same object is now encoded with consistent representations regardless of viewpoint. Beyond this qualitative observation, we additionally evaluate cross-view correspondence accuracy of our model's image features throughout training, and observe a steady improvement as training progresses (Appendix B). Fig. 2-(b) makes the underlying structure more explicit by visualizing the first three principal components of the image-token hidden states at the ℓ-th LLM layer in RGB. While the baseline features appear noisy and largely unstructured, our features are markedly cleaner and more semantically organized, with corresponding regions encoded with consistent colors across viewpoints. Together, these two effects indicate that the reconstruction objective propagates 3D-aware structure beyond the Gaussian summary tokens into the LLM's image features themselves, encouraging the same kind of compact, view-consistent representation that the summary tokens are trained to produce.

Summary tokens emergently group by object. A central design choice of Imagine3D-LLM is the bottleneck $M \ll KN'$, which compels overlapping content across views to be merged into a shared set of summary tokens (§3.2.1). To examine whether this bottleneck actually induces object-level grouping, we cluster the trained summary tokens via K-means and visualize, for each cluster, the 3D Gaussians decoded from its member tokens. As shown in Fig. 3, a single cluster's Gaussians reliably localize one object (e.g., the chair) across all views, despite the model receiving no clustering supervision. This emergent grouping effectively validates that learning to reconstruct the scene through a small set of summary tokens naturally guides the model toward the abstract, object-level understanding our design aims to achieve, without requiring any explicit object-level supervision.

4. Experiments

Implementation Details. Following prior work (Zheng et al., 2025b; Wang et al., 2025a), we build our Imagine3D-LLM based on LLaVA-Video-7B (Zhang et al., 2024). LLaVA-Video encodes each image as 210 tokens, resulting in $N' = 210$ and $KN' = 6720$ tokens for most settings where we use 32 images as input. For 3DGS estimation, we choose $\ell = 14$, as it is the middle layer of the base LLM. For the number of Gaussian summary tokens, we set $M = 2592$, which corresponds to 81 tokens per input image and is much fewer than the total number of image tokens $KN' = 6720$. The number of Gaussians decoded per token is set as $G = 32$. The Gaussian decoder head $\mathcal{H}(\cdot)$ is a three-layer MLP. We fine-tune the entire model on the combination of ScanNet-based datasets (SQA3D (Ma et al., 2022), ScanQA (Azuma et al., 2022), Scan2Cap (Chen et al., 2021), ScanRefer (Chen et al., 2020), and Multi3DRefer (Zhang et al., 2023)). Motivated by recent analysis (Zhang et al., 2025a; Zhou et al., 2025b; Zheng et al., 2025a) where training only with ScanNet-based datasets show largely overfitted results to ScanNet rather than unlocking 3D reasoning, we also sample and add 100K subset from SPAR-7M (Zhang et al., 2025b). The training is conducted with 16 GH200 GPUs, with an effective batch size of 64 for 1 epoch. More detailed implementation details can be found in Appendix A.

Table 1: Quantitative results on 3D question-answering over SQA3D (Ma et al., 2022), Real-3DQA (Ma et al., 2026), and ScanQA (Azuma et al., 2022). "Specialists" refer to models tailored to individual tasks via task-specific decoders. General-purpose 2D LMMs (Team, 2024; Wang et al., 2024a; Zhang et al., 2024) are reported under a zero-shot protocol. We report values available from prior works; "–" marks entries unavailable to us.

Method	$SQA3D_{test}$		$Real\text{-}3DQA_{test}$		$ScanQA_{val}$				
	EM	EM-R	EM	EM-R	CIDEr	BLEU-4	METEOR	ROUGE	EM
Specialists									
SQA3D (Ma et al., 2022)	46.6	–	–	–	–	–	–	–	–
ScanQA (Azuma et al., 2022)	–	–	–	–	64.9	10.1	13.1	33.3	21.1
2D LMMs									
InternVL2-8B (Team, 2024)	33.0	45.3	–	–	62.5	3.3	14.5	34.3	–
Qwen2-VL-7B (Wang et al., 2024a)	40.7	46.7	–	–	53.9	3.0	11.4	29.3	–
LLaVA-Video-7B (Zhang et al., 2024)	48.5	–	–	–	88.7	3.1	17.7	44.6	–
3D LMMs									
Scene-LLM (Fu et al., 2025)	53.6	–	–	–	80.0	11.7	15.8	35.9	27.2
LL3DA (Chen et al., 2024a)	–	–	–	–	76.8	–	15.9	37.3	–
LEO (Huang et al., 2024b)	50.0	52.4	–	–	80.0	11.5	16.2	39.3	21.5
ChatScene (Huang et al., 2024a)	54.6	57.5	–	–	87.7	14.3	18.0	41.6	21.6
Grounded 3D-LLM (Chen et al., 2024c)	–	–	–	–	72.7	13.4	–	–	–
LLaVA-3D (Zhu et al., 2024)	55.6	57.6	–	–	91.7	14.5	20.7	50.1	27.0
Video-3D-LLM (Zheng et al., 2025b)	58.6	–	31.2	36.4	102.1	16.4	20.0	49.3	30.1
3DRS (Huang et al., 2025)	60.6	–	-	-	104.8	-	-	-	30.3
Ross3D (Wang et al., 2025a)	63.0	65.7	36.6	41.5	107.0	17.9	20.9	**50.7**	**30.8**
Imagine3D-LLM (Ours)	**63.8**	**66.4**	**39.2**	**44.3**	**109.3**	**20.5**	**21.3**	49.7	29.9

Datasets and metrics. Following prior work (Zheng et al., 2025b; Wang et al., 2025a; Zheng et al., 2025a; Hu et al., 2025), we evaluate on seven benchmarks. Six are from ScanNet (Dai et al., 2017): SQA3D (Ma et al., 2022), Real-3DQA (Ma et al., 2026), and ScanQA (Azuma et al., 2022) for situated and spatial reasoning, Scan2Cap (Chen et al., 2021) for 3D grounded captioning, and ScanRefer (Chen et al., 2020) and Multi3DRefer (Zhang et al., 2023) for 3D object detection. We additionally evaluate on SPAR-Bench (Zhang et al., 2025b), which assesses genuine 3D perception and reasoning across three cognitive levels: low-level (e.g., depth estimation), medium-level (e.g., view-change inference), and high-level (e.g., spatial imagination). Following prior work (Zhu et al., 2024; Zheng et al., 2025b; Wang et al., 2025a; Zheng et al., 2025a; Fan et al., 2025; Hu et al., 2025), we use exact / refined-exact match (EM / EM-R) (Huang et al., 2024b; Zhu et al., 2024; Ma et al., 2026) on SQA3D and Real-3DQA, EM together with BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015) on ScanQA, the same captioning metrics under IoU@0.5 on Scan2Cap, Acc@0.25/0.5 on ScanRefer and the corresponding F1 on Multi3DRefer, and accuracy per cognitive level with the overall mean on SPAR-Bench.

4.1. Quantitative Comparison

ScanNet-based datasets. On the ScanNet-based benchmarks (SQA3D (Ma et al., 2022), Real-3DQA (Ma et al., 2026), ScanQA (Azuma et al., 2022), Scan2Cap (Chen et al., 2021), ScanRefer (Chen et al., 2020), and Multi3DRefer (Zhang et al., 2023)), we compare Imagine3D-LLM against task-specific specialists, general-purpose 2D LMMs, and recent 3D LMMs that differ in how they encode the 3D scene. As shown in Tab. 1, and Tab. 2, Imagine3D-LLM significantly improves performance across benchmarks, outperforming the strongest prior 3D LMM (Ross3D) on almost all benchmarks. Especially, state-of-the-art performance in Real-3DQA is notable, which is a benchmark explicitly designed to validate 3D reasoning capabilities and penalize language-shortcut overfitting (Ma et al., 2026), indicating that our improvements stem from genuine 3D reasoning.

SPAR-Bench. On SPAR-Bench (Zhang et al., 2025b), we compare Imagine3D-LLM against three categories of models: the *proprietary* group (Hurst et al., 2024; Anthropic, 2025), the *general-purpose 2D LMMs* group (Zhang et al., 2024; Wang et al., 2024a; Team, 2024; Bai et al., 2025), and the *spatially-aware* group of recent models specifically designed for spatial reasoning (Fan et al., 2025; Hu et al., 2025; Chen et al., 2025). As shown in Tab. 3, Imagine3D-LLM achieves state-of-the-art performance on SPAR-Bench, outperforming all competing methods

Table 2: Evaluation of 3D dense captioning and visual grounding on Scan2Cap (Chen et al., 2021), ScanRefer (Chen et al., 2020), and Multi3DRefer (Zhang et al., 2023).

Method	$\text{Scan2Cap}_{\text{val}}$ (IoU@0.5)				$\text{ScanRefer}_{\text{val}}$		$\text{Multi3DRefer}_{\text{val}}$	
	ROUGE	BLEU-4	METEOR	CIDEr	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
Specialists								
Scan2Cap (Chen et al., 2021)	44.5	23.3	22.0	35.2	–	–	–	–
3DJCG (Cai et al., 2022)	50.8	31.0	24.2	49.5	49.6	37.3	–	26.6
ScanRefer (Chen et al., 2020)	–	–	–	–	37.3	24.3	–	–
M3DRef-CLIP (Zhang et al., 2023)	–	–	–	–	51.9	44.7	42.8	38.4
3D LMMs								
LL3DA (Chen et al., 2024a)	55.1	36.8	26.0	65.2	–	–	–	–
Ground 3D-LLM (Chen et al., 2024c)	–	–	–	–	47.9	44.1	45.2	40.6
LEO (Huang et al., 2024b)	58.1	38.2	27.9	72.4	–	–	–	–
ChatScene (Huang et al., 2024a)	58.1	36.3	–	77.1	55.5	50.2	57.1	52.4
LLaVA-3D (Zhu et al., 2024)	63.4	41.1	30.2	79.2	54.1	42.4	–	–
Video-3D-LLM (Zheng et al., 2025b)	62.3	42.4	28.9	83.8	58.1	51.7	58.0	52.7
VG-LLM (Zheng et al., 2025a)	62.6	41.5	28.9	80.0	-	-	-	-
3DRS (Huang et al., 2025)	-	41.6	-	86.1	-	-	-	-
Ross3D (Wang et al., 2025a)	66.9	43.4	30.3	81.3	61.1	54.4	59.6	54.3
Imagine3D-LLM (Ours)	**67.6**	**48.3**	**32.6**	**99.2**	**62.8**	**56.3**	**60.2**	**55.0**

Table 3: Evaluation on SPAR-Bench (Zhang et al., 2025b).

Model	SPAR-Bench			
	Avg.	Low	Med.	High
Proprietary				
GPT-4o (Hurst et al., 2024)	36.4	29.3	24.9	45.1
Claude-3.7-Sonnet (Anthropic, 2025)	21.8	25.4	7.3	23.3
General-purpose 2D LMMs				
LLaVA-Video-7B (Zhang et al., 2024)	32.3	23.6	24.8	42.6
Qwen2-VL-7B (Wang et al., 2024a)	30.7	27.5	20.4	37.0
InternVL2.5-8B (Chen et al., 2024e)	36.3	29.5	31.9	43.8
Qwen2.5-VL-72B (Bai et al., 2025)	39.4	35.4	23.1	48.4
Spatially-aware				
VLM3R-7B (Fan et al., 2025)	43.2	39.8	28.4	51.2
G^2VLM-SR-2B (Hu et al., 2025)	54.9	60.0	36.3	56.5
3DThinker-7B (Chen et al., 2025)	63.3	-	-	-
Imagine3D-LLM-7B (Ours)	**68.5**	**60.5**	**67.0**	**76.0**

across every cognitive level by substantial margins. Compared to general-purpose 2D LMMs, Imagine3D-LLM surpasses even the largest 72B-scale model (Bai et al., 2025) by over 29 points despite using a 7B backbone. More notably, Imagine3D-LLM outperforms the strongest spatially-aware baseline, 3DThinker-7B (Chen et al., 2025), by 5.2 points overall, with even larger margins over G^2VLM-SR-2B (Hu et al., 2025) and VLM3R-7B (Fan et al., 2025). The improvement is especially pronounced on the medium- and high-level splits, which most directly require the object-level cross-view reasoning our reconstruction objective targets. These results validate our central claim: equipping MLLMs to assemble a compact 3D representation of the scene before answering substantially improves 3D reasoning.

4.2. Ablation Studies

Controlled comparison with baseline. To isolate the contribution of our mental-reconstruction objective from the effect of the training data and recipe, we compare Imagine3D-LLM against a baseline that shares the identical backbone, training data, and schedule but omits the Gaussian summary tokens together with the reconstruction and distillation objectives, i.e., standard visual instruction tuning. Here, we report one representative metric per benchmark EM score for SQA3D / ScanQA, ROUGE for Scan2Cap, Acc@25 for

Table 4: Controlled comparison against the base MLLM. Both models share the identical backbone, training data, and schedule; the baseline is trained with standard visual instruction tuning, without the Gaussian summary tokens or the reconstruction and distillation objectives.

Method	$\text{SQA3D}_{\text{test}}$	$\text{ScanQA}_{\text{val}}$	$\text{Scan2Cap}_{\text{val}}$	$\text{ScanRefer}_{\text{val}}$	$\text{Multi3DRefer}_{\text{val}}$	$\text{SPAR}_{\text{Avg.}}$
Baseline	56.5	26.2	63.1	58.3	57.4	60.9
Imagine3D-LLM (Ours)	**63.8**	**29.9**	**67.6**	**62.8**	**60.2**	**68.5**

ScanRefer, F1@0.25 for Multi3DRefer, and average performance SPAR. As shown in Tab. 4, Imagine3D-LLM improves over this controlled baseline on every benchmark, confirming that the gains arise from our objective rather than from the additional training data alone. The improvement is especially pronounced on the medium-level cross-view split of SPAR-Bench (+15.5, from 51.5 to 67.0; see Tab. 3), which most directly exercises the object-level cross-view reasoning our reconstruction objective targets.

Layer selection for Gaussian decoding. We further ablate the choice of the LLM layer ℓ from which the Gaussian summary tokens are extracted and decoded into 3D Gaussians. As discussed in § 3.2.1, we hypothesize that the middle layer of the LLM offers the best balance between two opposing requirements: the Gaussian tokens must accumulate enough visual information to faithfully describe the scene, yet the remaining layers must be sufficient to propagate this 3D-aware information into the text tokens for reasoning. We empirically verify this by extracting from layer $\ell \in \{7, 14, 21\}$ of the 28-layer LLM. As shown in Tab. 5, the middle layer ($\ell = 14$) consistently performs best, validating our design choice and aligning with prior findings on information flow in MLLMs (Kaduri et al., 2025; Jiang et al., 2025b; Yoon et al., 2025).

Table 5: Effect of the layer ℓ from which Gaussian tokens are decoded.

Layer ℓ	$\text{SQA3D}_{\text{test}}$	SPAR-Bench
7 (early)	60.7	60.2
14 (middle, Ours)	**63.8**	**68.5**
21 (late)	61.7	64.0

Effectiveness of distillation. We compare our full model against a variant trained without the distillation losses ($\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{param}}$). As shown in Tab. 6, removing distillation (see Recon. only 1 epoch) causes a substantial drop on both benchmarks. We empirically observe that without distillation, $\mathcal{L}_{\text{recon}}$ remains large in scale even after 1 epoch, so a large fraction of the gradient signal is allocated to reducing $\mathcal{L}_{\text{recon}}$, leaving the language modeling objective under-optimized. Consistent with this, the variant improves steadily with longer training ($1 \rightarrow 2 \rightarrow 4$ epochs) and at 4 epochs nearly matches our full 1-epoch model. In contrast, the $\mathcal{L}_{\text{LM}}$-only baseline shows only modest improvements with slightly longer iterations ($1 \rightarrow 2$ epochs), and performance decreases after 4 epochs due to the model overfitting to the training data and losing general reasoning performance when only trained with $\mathcal{L}_{\text{LM}}$. Together, these results indicate that the improved 3D reasoning comes primarily from jointly learning reconstruction and understanding, with distillation serving mainly to accelerate convergence and make training practical.

Table 6: Effect of distillation from the compact Gaussian teacher.

Method	$\text{SQA3D}_{\text{test}}$	SPAR-Bench
Baseline (1 ep.)	56.5	60.9
Baseline (2 ep.)	57.2	61.5
Baseline (4 ep.)	52.3	55.2
Recon. only (1 ep.)	57.7	57.6
Recon. only (2 ep.)	61.9	65.1
Recon. only (4 ep.)	63.7	67.9
Imagine3D-LLM (Full, 1 ep.)	**63.8**	**68.5**

Isolating the source of the gains. Having established that reconstruction, not longer training, drives the gains, we next ask whether the improvement truly comes from learning to reconstruct the scene (mental reconstruction), or merely from access to the teacher's representation, the distillation signal, or the added token capacity. We disentangle these factors in Tab. 7. First, feeding the teacher's query tokens directly to the LLM as additional input (+ *Teacher tokens*), without asking the model to reconstruct, performs on par with the baseline. Second, supervising the summary tokens with the distillation loss but *without* the reconstruction loss (+ *Distill. only*) likewise fails to help. We attribute both negative results to a representational mismatch between the teacher's query tokens and the LLM's own embedding space: simply injecting or distilling these tokens does not

Table 7: Isolating the source of the gains. + *Teacher tokens*: the teacher's query tokens are fed directly to the LLM as input. + *Distill. only*: summary tokens supervised by $\mathcal{L}_{\text{distill}}$ but without $\mathcal{L}_{\text{recon}}$.

Variant	$\text{SQA3D}_{\text{test}}$	SPAR-Bench
Baseline	56.5	60.9
+ Teacher tokens	56.4	60.8
+ Distill. only	56.6	60.5
+ Recon. (4 ep.)	63.7	67.9
Full (Ours, 1 ep.)	**63.8**	**68.5**

teach the LLM how to interpret or attend to them, and without a dedicated mechanism to compensate for this misalignment or to explicitly force the language model to attend to the injected tokens, as in feature-fusion approaches such as VLM3R (Fan et al., 2025) the added tokens are largely ignored, leaving performance at the baseline level. In contrast, the reconstruction loss compels the Gaussian summary tokens to actively attend to the image tokens and extract the visual evidence required to estimate their 3D Gaussian parameters. This active aggregation reshapes the attention and representations *within* the LLM such as sharpening cross-view correspondence and inducing object-level grouping (§3.3), which we believe iss what benefits downstream reasoning. Together, these results indicate that the improvement arises primarily from jointly learning mental reconstruction.

Number of Gaussian summary tokens. An important design choice of Imagine3D-LLM is that the Gaussian summary tokens act as a representational bottleneck, encouraging multi-view content corresponding to the same 3D region to be merged into shared tokens (§ 3.2.1). To validate this design, we ablate the number of summary tokens M while keeping all other components fixed. As reported in Tab. 8, we compare introducing roughly 40, 80, 160 tokens per image, which makes the total number of summary tokens 1296, 2592, 5184, given 32 images. The evaluation shows that too few tokens (1296) provide insufficient capacity, while too many tokens (5184) weaken the bottleneck that drives object-level grouping, leading both extremes to underperform our chosen setting.

Table 8: Number of Gaussian summary tokens M.

M	$\text{SQA3D}_{\text{test}}$	SPAR-Bench
1296	61.9	64.6
2592 (Ours)	**63.8**	**68.5**
5184	60.3	58.4

5. Conclusion

We presented Imagine3D-LLM, a framework that equips a multi-image MLLM with the ability to perform an explicit *mental reconstruction* of a 3D scene before answering. We introduce a compact set of learnable *Gaussian summary tokens* that are appended to the image tokens and decoded into a compact 3DGS representation, jointly trained with the standard language modeling objective. Across challenging spatial reasoning benchmarks, Imagine3D-LLM significantly outperforms prior approaches that rely on explicit pixel-level geometry or geometry-foundation features, and our analyses further reveal that the reconstruction objective propagates beyond the Gaussian summary tokens, reshaping the image features into more spatially aligned and semantically structured representations. We believe that enabling machines to understand how the scene is structured before answering offers a promising step toward MLLMs that perceive 3D space in a more human-like manner.

References

Honggyu An, Jaewoo Jung, Mungyeom Kim, Sunghwan Hong, Chaehyun Kim, Kazumi Fukuda, Minkyeong Jeon, Jisang Han, Takuya Narihira, Hyuna Ko, et al. C3g: Learning compact 3d representations with 2k gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Anthropic. Claude 3.7 sonnet system card. https://www.anthropic.com/news/claude-3-7-sonnet, 2025. Hybrid reasoning multimodal model released November 2025.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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:2502.13923*, 2025.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5044–5053, 2023.

Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10(12):551–557, 2006.

Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, 2022.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19457–19467, 2024.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in RGB-D scans using natural language. In *ECCV*, 2020.

Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in RGB-D scans. In *CVPR*, 2021.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133, 2023.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *CVPR*, 2024a.

Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, YU Gang, Taihao Li, and Tao Chen. Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv:2405.10370*, 2024c.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024d.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024e.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *Third Workshop on Visual Concepts*, 2026.

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

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. OpenNeRF: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Rendered Novel Views. In *International Conference on Learning Representations*, 2024.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Elisabetta Fedele, Boyang Sun, Leonidas Guibas, Marc Pollefeys, and Francis Engelmann. Superdec: 3d scene decomposition with superquadrics primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24625–24635, October 2025.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. In *WACV*, 2025.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2410.22128*, 2024.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2VLM: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. *arXiv preprint arXiv:2511.21688*, 2025.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *NeurIPS*, 2024a.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *ICML*, 2024b.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025.

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:2410.21276*, 2024.

Roni Itkin, Noam Issachar, Yehonatan Keypur, Anpei Chen, and Sagie Benaim. Globalsplat: Efficient feed-forward 3d gaussian splatting via global scene tokens. *arXiv preprint arXiv:2604.15284*, 2026.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025a.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025b.

Kim Jun-Seong, GeonU Kim, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14137–14146, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14549–14558, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9339–9350, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv:2408.03326*, 2024.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *CVPR*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3d large language models really understand 3d spatial relationships? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.23523.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. Spatiallm: Training large language models for structured indoor modeling. *arXiv preprint arXiv:2506.07491*, 2025.

Tom Monnier, Jake Austin, Angjoo Kanazawa, Alexei Efros, and Mathieu Aubry. Differentiable blocks world: Qualitative 3d decomposition by rendering primitives. *Advances in Neural Information Processing Systems*, 36:5791–5807, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10344–10353, 2019.

Despoina Paschalidou, Luc Van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1060–1070, 2020.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv:2501.01428*, 2025.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Jiawei Ren, Michal Jan Tyszkiewicz, Jiahui Huang, and Zan Gojcic. Tokengs: Decoupling 3d gaussian prediction from pixels with learnable tokens. *arXiv preprint arXiv:2604.15239*, 2026.

William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

OpenGVLab Team. InternVL2: Better than the Best—Expanding Performance Boundaries of Open-Source Multimodal Models with the Progressive Scaling Strategy, 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

Anh Thai, Songyou Peng, Kyle Genova, Leonidas Guibas, and Thomas Funkhouser. Splattalk: 3d vqa with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4712–4721, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2635–2643, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

Alexander Veicht, Sunghwan Hong, Dániel Baráth, and Marc Pollefeys. Zipsplat: Fewer gaussians, better splats. *arXiv preprint arXiv:2606.05102*, 2026.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. *arXiv:2503.11651*, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025c.

Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. *Advances in Neural Information Processing Systems*, 37:107326–107349, 2024b.

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19823–19832, 2024a.

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. *arXiv:2412.14171*, 2024b.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, Chanho Eom, Sunghwan Hong, and Seungryong Kim. Visual representation alignment for multimodal large language models, 2025. URL https://arxiv.org/abs/2509.07979.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

Huanyu Zhang, Chengzu Li, Wenshan Wu, Shaoguang Mao, Yifan Zhang, Haochen Tian, Ivan Vulić, Zhang Zhang, Liang Wang, Tieniu Tan, et al. Scaling and beyond: Advancing spatial reasoning in mllms requires new recipes. *arXiv preprint arXiv:2504.15037*, 2025a.

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:2503.22976*, 2025b.

Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19781–19791, 2025c.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d LLM: learning position-aware video representation for 3d scene understanding. In *CVPR*, 2025b.

Mingquan Zhou, Chen He, Ruiping Wang, and Xilin Chen. Ov3d-cg: Open-vocabulary 3d instance segmentation with contextual guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5305–5314, 2025a.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Eric Xin Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8600–8612, 2025b.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv:2409.18125*, 2024.

Xueyan Zou, Yuchen Song, Ri-Zhao Qiu, Xuanbin Peng, Jianglong Ye, Sifei Liu, and Xiaolong Wang. 3d-spatial multimodal memory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision*, 133(2):611–627, 2025.

This appendix provides additional implementation details (§ A), correspondence analysis (§ B), further analyses (§ C), additional related works (§ D), and limitations (§ E) that complement the main paper.

A. Additional Implementation Details

Datasets and input. We use an input resolution of 384×384, matching the native resolution of LLaVA-Video-7B (Zhang et al., 2024). For the ScanNet-based datasets (SQA3D (Ma et al., 2022), ScanQA (Azuma et al., 2022), Scan2Cap (Chen et al., 2021), ScanRefer (Chen et al., 2020), and Multi3DRefer (Zhang et al., 2023)), we uniformly sample 32 frames per scene, following prior work (Zhu et al., 2024; Zheng et al., 2025b; Wang et al., 2025a; Fan et al., 2025; Hu et al., 2025). For SPAR (Zhang et al., 2025b), we use the 104K-sample multi-view subset released by VG-LLM (Zheng et al., 2025a), in which each scene contains either 2, 3, or 32 views, and we feed all available views per scene to the model. We mix SPAR-7M into the ScanNet-based training set to mitigate the language-pattern overfitting tendency previously observed when training on ScanNet alone (Ma et al., 2026; Zheng et al., 2025a).

Compact Gaussian teacher. We build our teacher on top of ZipSplat's (Veicht et al., 2026) open-source codebase, where we control the number of compact tokens to 2592 to make an adequate bottleneck structure as mentioned in § 4.2. The teacher can be readily swapped with other token-level feed-forward 3DGS frameworks (An et al., 2026; Itkin et al., 2026; Ren et al., 2026), and is kept frozen and used only at training time.

Training and architecture details. We train with the AdamW optimizer using a global batch size of 64 with 4 gradient accumulation steps. The peak learning rate after warmup is $1\mathrm{e}{-5}$ for the LLM and $2\mathrm{e}{-6}$ for the vision encoder, and we clip the gradient norm to 1.0. For our full training objective, we set $\lambda_{\text{recon}} = 3.0$ and $\lambda_{\text{distill}} = 1.0$, with $\lambda_{\text{MSE}} = \lambda_{\text{param}} = \lambda_{\text{token}} = 1.0$ and $\lambda_{\text{LPIPS}} = 0.05$. The teacher-alignment projector $g(\cdot)$ in §3.2.2 is a two-layer MLP with a GeLU activation (Hendrycks and Gimpel, 2016) that maps the LLM's hidden dimension $3584 \rightarrow 1536$ to match the teacher's query dimension d_T. The Gaussian decoder head $\mathcal{H}(\cdot)$ is a three-layer MLP with SiLU activations (Elfwing et al., 2018) mapping $1536 \rightarrow 736$, which is the dimension required to specify the parameters of $G = 32$ Gaussians per summary token. We use spherical harmonics of degree 2 for view-dependent color.

The 3D LMMs we compare against in Tab. 1, Tab. 2, and Tab. 3 differ in how they encode the 3D scene: Scene-LLM (Fu et al., 2025) and LL3DA (Chen et al., 2024a) consume point-cloud features; LEO (Huang et al., 2024b), ChatScene (Huang et al., 2024a), and Grounded 3D-LLM (Chen et al., 2024c) use object-centric representations; LLaVA-3D (Zhu et al., 2024) lifts 2D features into a 3D voxel grid; and Video-3D-LLM (Zheng et al., 2025b), GPT4Scene (Qi et al., 2025), and Ross3D (Wang et al., 2025a) treat multi-view images as video sequences.

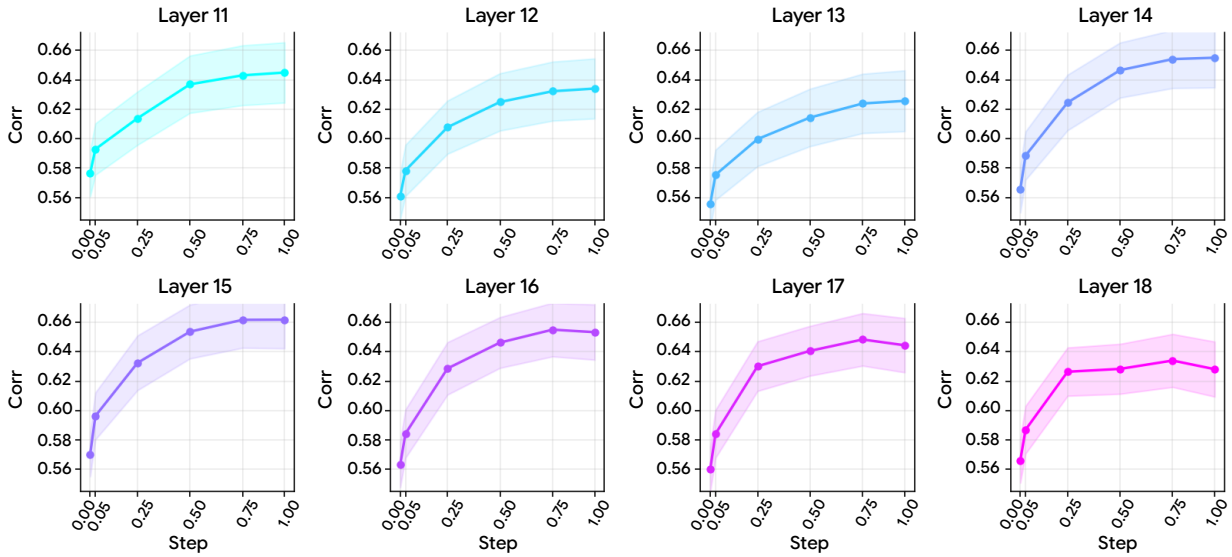

Figure 4: **Layer-wise cross-view correspondence during training.** We report cross-view correspondence scores for image-token representations extracted from middle LLM layers $\ell = 11 \sim 18$ on SQA3D (Ma et al., 2022) test samples. Scores increase consistently throughout training, with pronounced gains near the Gaussian decoding layer ($\ell = 14$), indicating that joint reconstruction training improves the view-consistency of intermediate image features. Shaded regions indicate variance across SQA3D (Ma et al., 2022) test samples.

B. Additional Correspondence Analysis

Following 3DRS (Huang et al., 2025), we evaluate cross-view feature correspondence as a quantitative probe of how 3D-aware the image-token representations have become. We voxelize the 3D coordinates associated with image tokens with the given 3D point clouds for each scene of ScanNet (Dai et al., 2017), construct cross-view pairs from tokens in different views that fall into the same voxel, and report the correspondence score as the average cosine similarity between the ℓ-th layer hidden states of these pairs. A higher score indicates stronger feature consistency for image tokens describing the same 3D region across viewpoints. As shown in Fig. 4, correspondence scores in the middle LLM layers, particularly the layers $\ell = 11 \sim 18$ that surround the Gaussian decoding layer, improve substantially throughout training. This finding plays a dual role. As a corollary, it confirms that reconstruction supervision propagates 3D-aware structure beyond the summary tokens themselves into the image-token representations. We further attribute this propagation to our compact-bottleneck design ($M \ll KN'$): with far fewer summary tokens than image tokens, each summary token is forced to aggregate the same 3D region across multiple views, which in turn pressures the upstream image features to encode that region consistently across viewpoints. More importantly, the result shows that joint reconstruction *delivers* the kind of cross-view correspondence that prior methods such as 3DRS supply through explicit supervision: improved correspondence emerges naturally from learning to reconstruct, without ever being directly optimized.

C. Additional Analysis

C.1. Additional visualization of attention between images

We further visualize the attention maps across multi-view images using the same protocol as Fig. 2. As illustrated in Fig. 5, image tokens from our model reliably focus on matching regions of the same object across different views, demonstrating their robustness across diverse scenes.

C.2. Additional visualization of clustered Gaussian

We additionally visualize the attention maps between Gaussian summary tokens and multi-view images, following the same protocol as Fig. 3. As shown in Fig. 6, common objects are clustered into similar Gaussian

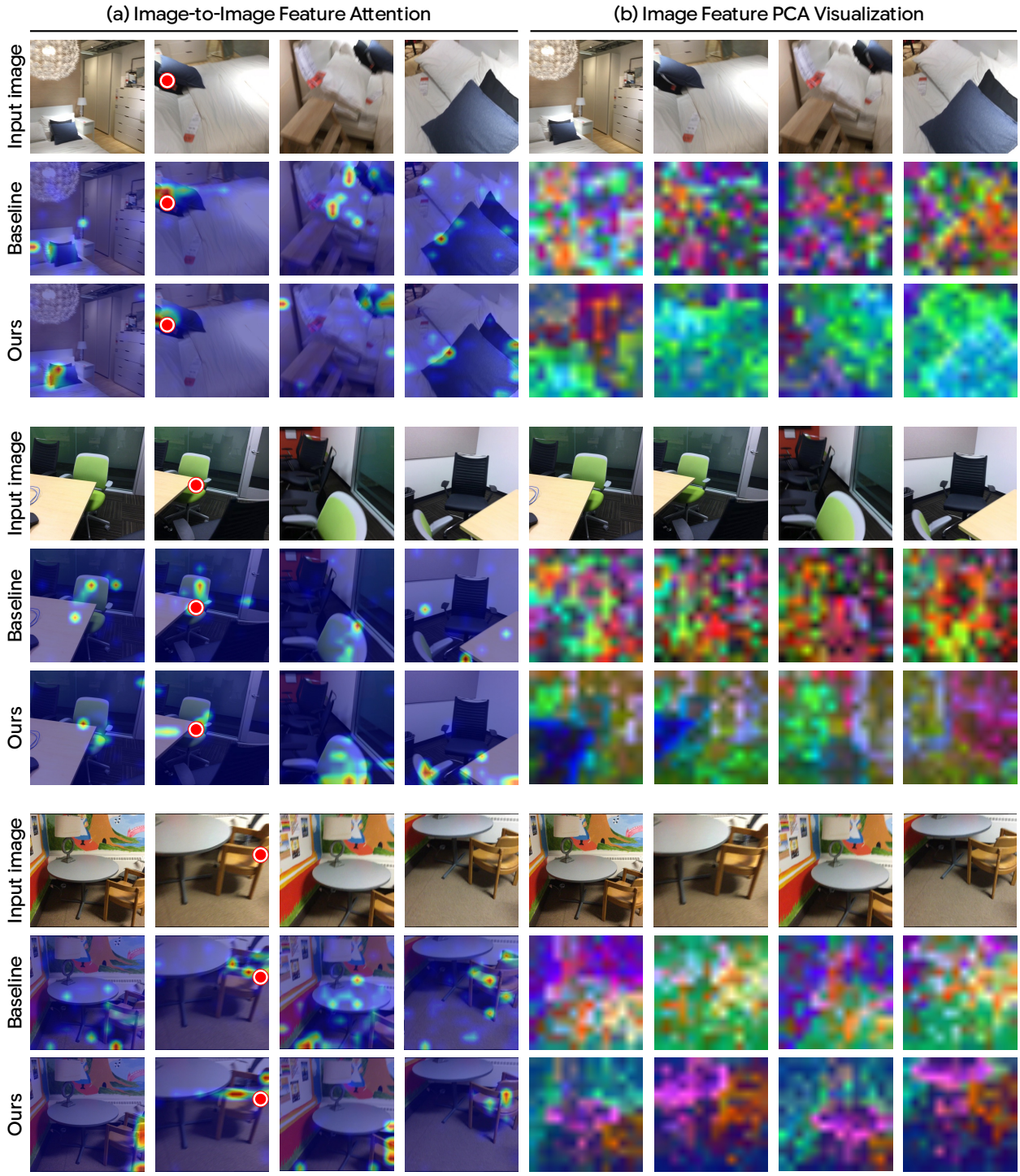

Figure 5: **Visualization of attention maps between multi-view images and PCA visualizations of the image features.** **(a)** Image-to-image attention maps between a query point (red dot) and other input views. Our model exhibits substantially stronger cross-view correspondences than the baseline, indicating that the Gaussian summary tokens implicitly drive the image features to align across views. **(b)** PCA visualization of the LLM's image features (first three components shown as RGB). Our features are noticeably cleaner and more semantically structured, with corresponding objects (e.g., chairs, desks) encoded with consistent colors across views, reflecting the emergence of object-level abstractions.

summary tokens without any explicit supervision.

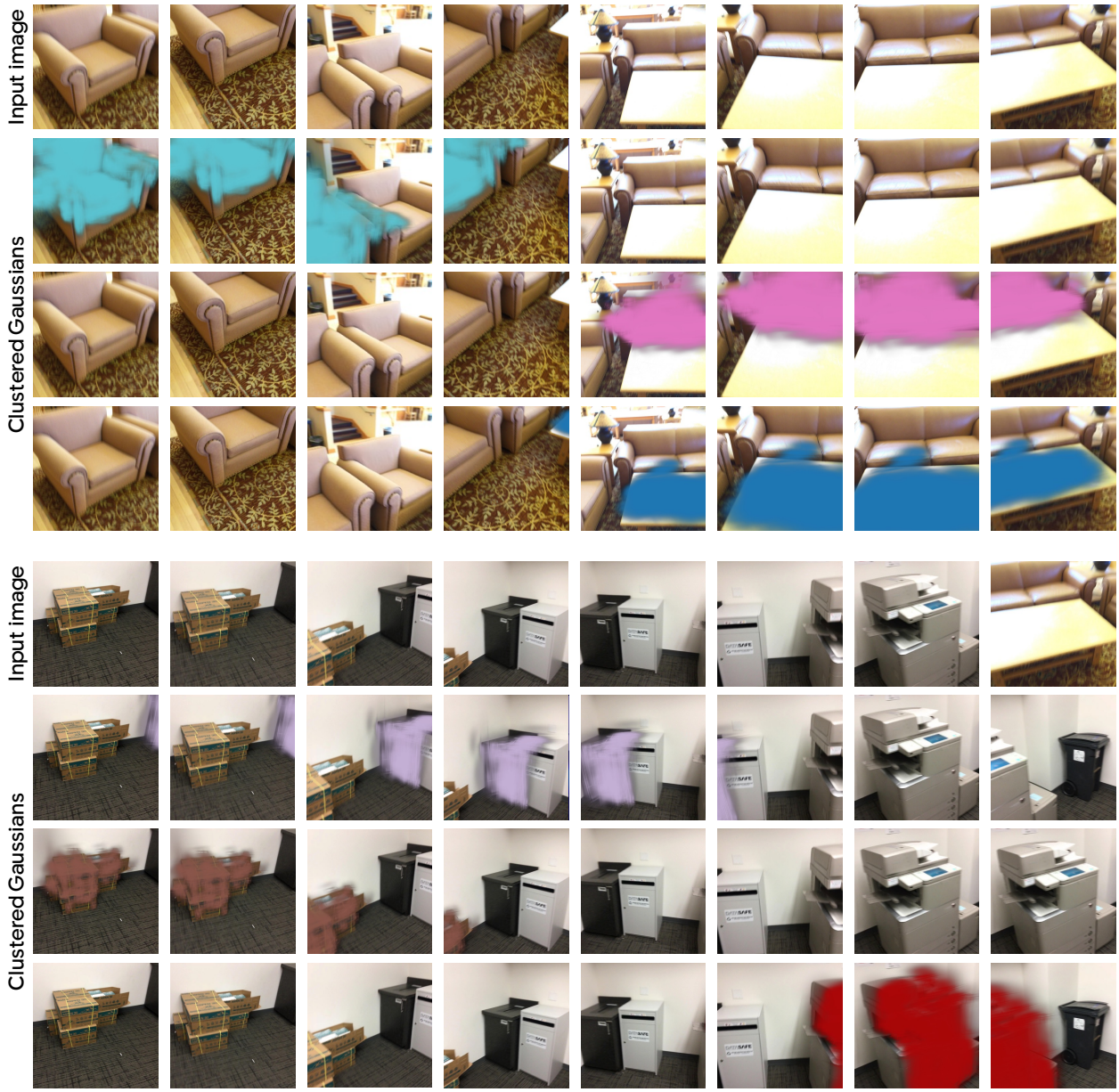

Figure 6: **Visualization of clustered Gaussian from summary tokens.** We visualize the Gaussians clustered from a set of Gaussian summary tokens via K-means over the summary tokens. Similar summary tokens are mapped to the same object without any explicit clustering supervision.

C.3. Additional quantitative evaluation of semantic feature structuring

The PCA visualizations in Fig. 2 and the cross-view correspondence analysis in Appendix B provide qualitative and geometric evidence that joint reconstruction training improves the LLM's image features. To obtain a more direct quantitative measure of whether the features become more *semantically* structured, we conduct an unsupervised semantic segmentation probe following the standard evaluation methodology of (Cuttano et al., 2026; Hamilton et al., 2022).

Specifically, on the ScanNet scenes of the SQA3D evaluation set (which provide per-image 2D semantic labels), we extract the LLM's image features at layer ℓ for all 32 input views and apply K-means clustering *jointly* over all views' tokens, with k set to the number of ground-truth NYU40 classes present in the scene. Since the

Table 9: Unsupervised semantic segmentation probe. We cluster the LLM's image features via K-means and measure mIoU against ground-truth ScanNet labels. DINOv3 (Siméoni et al., 2025) is evaluated under the same protocol as an upper-bound reference.

Method	mIoU
Baseline (no summary tokens)	26.54
Imagine3D-LLM (Ours)	**35.43**
DINOv3 (Siméoni et al., 2025)	37.62

Table 10: Reconstruction quality on SQA3D test scenes. We compare the rendering quality of our Gaussian summary tokens against the compact Gaussian teacher (Veicht et al., 2026) used during training. Higher is better for PSNR and SSIM; lower is better for LPIPS.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Gaussian teacher (Veicht et al., 2026)	20.60	0.76	0.44
Imagine3D-LLM (Ours)	17.61	0.74	0.49

resulting clusters carry no labels, we follow the standard unsupervised segmentation protocol: each cluster is assigned to its maximally-overlapping ground-truth class via majority vote. After assignment, cluster labels are upsampled to the label resolution and we compute mIoU against the ground-truth semantic masks, aggregated across all views of each scene. As DINOv3 (Siméoni et al., 2025) features are well-known to be semantically structured, we additionally evaluate DINOv3 under the same protocol as an upper-bound reference.

As reported in Tab. 9, our model's image features achieve 35.43 mIoU, a substantial improvement over the baseline (26.54) that closes much of the gap to DINOv3 (37.62). This result provides direct evidence that the feature space itself, not only its downstream usage, becomes more semantically structured through joint reconstruction training. Importantly, this structuring emerges as a byproduct of the reconstruction objective—the image features are never directly supervised with semantic labels—confirming that learning to reconstruct propagates object-level structure into the LLM's visual representations.

C.4. Additional quantitative evaluation of reconstruction quality

While our primary goal is to improve 3D reasoning rather than novel view synthesis, here we further analyze rendering quality on the SQA3D test scenes and compare against the compact Gaussian teacher (Veicht et al., 2026) in Tab. 10. As expected, our reconstruction quality is somewhat below the teacher's, since our Gaussians are decoded from the hidden states of an LLM that is jointly optimized for language modeling and reconstruction, trained with much fewer iterations, and without any special design to match the teacher's rendering fidelity. However, the primary goal of our design is not to achieve high-fidelity rendering: the reconstruction objective serves as an inductive bias that teaches the model how the scene is structured (mental reconstruction), not as a novel view synthesis system (§ 3.2.1).

C.5. Additional analysis between Gaussian summary token and multi-view images

To accurately reconstruct the scene, it is crucial for the Gaussian summary tokens to effectively aggregate information from multi-view images. We conduct additional analysis to examine how Gaussian summary tokens attend to multi-view image features when generating 3D Gaussians. As shown in Fig. 7, each Gaussian summary token attends to corresponding regions of the same object across multiple views in the scene. This behavior emerges because a limited number of Gaussian summary tokens are compelled to aggregate multiple occurrences of the same object into a single representation, effectively clustering common objects without any explicit supervision. This emergent object-centric property is reminiscent of human perception, where the same object observed from multiple viewpoints is interpreted as a single entity to form a coherent 3D understanding of the scene.

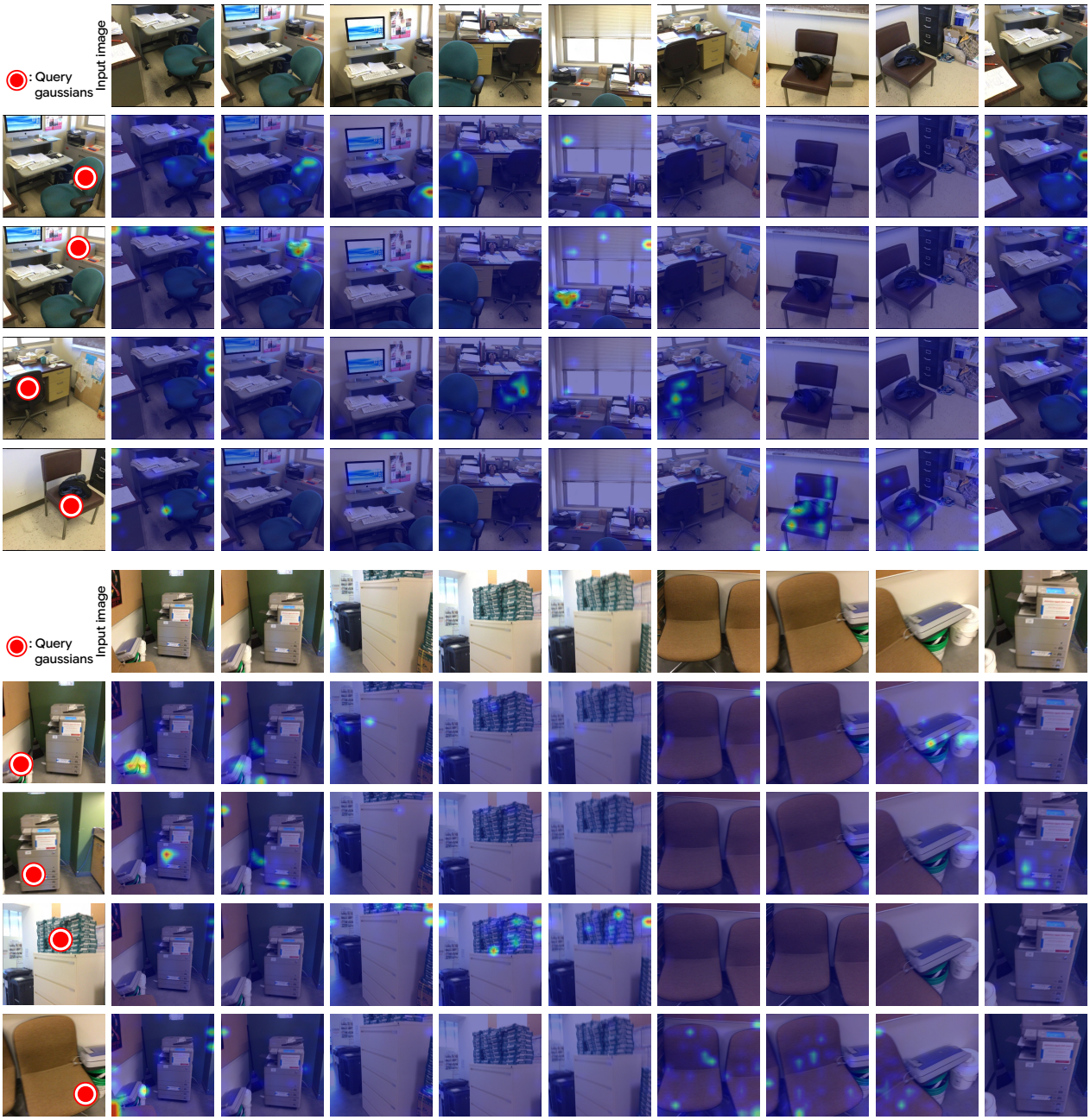

Figure 7: Visualization of attention maps between Gaussian summary tokens and images. We visualize the attention score map between a Gaussian summary token (a red dot indicating the decoded Gaussian) and the input images. Each Gaussian token attends to corresponding regions of the same object across multiple views.

Training setting	Peak VRAM (GB)
w/o $\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{distill}}$	40
Imagine3D-LLM (full)	44

Table 11: **Training-time computation cost.** Peak per-GPU VRAM measured during training under identical batch size and sequence length. The reconstruction and distillation losses introduce only a modest memory overhead while substantially improving 3D reasoning.

Table 12: **Inference cost comparison.** Peak GPU memory and per-sample inference time measured on the SPAR evaluation set. VLM3R-7B (Fan et al., 2025) additionally runs an external 3D foundation model (CUT3R (Wang et al., 2025c)) on all input frames before fusion.

Method	Peak Memory (GB)	Inference Time (ms)
Baseline (7B)	17.09	112.1
VLM3R-7B (Fan et al., 2025)	25.29	344
Imagine3D-LLM (7B)	20.43	176.9

C.6. Computation cost analysis.

We further analyze the additional training-time cost incurred by our reconstruction and distillation objectives. As reported in Tab. 11, augmenting the standard language modeling loss with $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{distill}}$ raises the peak per-GPU VRAM from 40 GB to 44 GB, a 10% increase under identical batch size and sequence length. This modest overhead is a direct consequence of our design choice to keep the number of Gaussian summary tokens compact ($M = 2592$, § 3.2.1). To make this concrete, pixel-aligned feed-forward 3DGS estimators (Charatan et al., 2024; Chen et al., 2024d; Hong et al., 2024; Ye et al., 2024) predict one Gaussian per input pixel, which for our setting of 32 images at 384×384 resolution would amount to roughly 4.7M Gaussians per scene. In contrast, Imagine3D-LLM decodes only $M \times G = 2592 \times 32 \approx 83$K Gaussians, less than 1.8% of the pixel-aligned count. This compactness keeps the rasterization, the teacher forward pass, and the gradient buffers comparatively lightweight, allowing the reconstruction objective to fit alongside MLLM finetuning at a manageable cost. The overhead is also confined to training, as the compact Gaussian teacher (An et al., 2026) is discarded at inference time and the Gaussian decoder head $\mathcal{H}(\cdot)$ is only invoked when reconstructions are explicitly required. Given the substantial gains in 3D reasoning across all evaluated benchmarks (Tab. 1, 2, 3), we view this as a favorable trade-off, indicating that imagining the scene through a compact Gaussian abstraction is not only effective but also practically affordable.

Inference time cost. We further report peak GPU memory and per-sample inference time on the SPAR evaluation set for three models: (a) the baseline multi-image MLLM taking only images, (b) VLM3R-7B (Fan et al., 2025), which fuses image features with external CUT3R (Wang et al., 2025c) features, and (c) our model.

As shown in Tab. 12, Imagine3D-LLM uses 20.43 GB and 176.9 ms per sample, compared to the baseline's 17.09 GB and 112.1 ms. The additional overhead compared to the baseline comes almost entirely from the longer input sequence ($M = 2592$ additional summary tokens), as the Gaussian decoder head is only invoked when the user explicitly requests a reconstruction and the distillation teacher is discarded after training. Importantly, approaches like VLM3R that fuse external 3D foundation model features must run a separate model (CUT3R) on all input frames at inference before fusion, resulting in substantially higher memory (25.29 GB) and latency (344 ms). Our approach requires no external 3D model at inference, keeping the cost well below that of feature-fusion alternatives while still providing substantial reasoning gains.

C.7. Visualization of estimated Gaussians.

While our main objective is improving 3D reasoning rather than achieving photorealistic novel view synthesis, we additionally visualize the estimated Gaussians in 3D to verify that the Gaussian summary tokens indeed acquire a coherent 3D understanding of the scene. As shown in Fig. 8, the rendered views recover the overall layout of the scene and the rough placement of major objects (e.g., the central table surface and surrounding

seating arrangement), confirming that even when decoded from only $M = 2592$ summary tokens, the model is able to organize multi-view evidence into a spatially consistent 3D abstraction. We emphasize, however, that the rendering quality itself is not the primary concern of Imagine3D-LLM. Unlike feed-forward 3DGS estimators that are explicitly optimized for high-fidelity novel view synthesis (Charatan et al., 2024; Chen et al., 2024d; Hong et al., 2024; Ye et al., 2024; Jiang et al., 2025a), our Gaussian summary tokens are trained jointly with the language modeling objective and are bottlenecked to a small set of tokens, which inherently trades off pixel-level fidelity for compact, object-level abstraction.

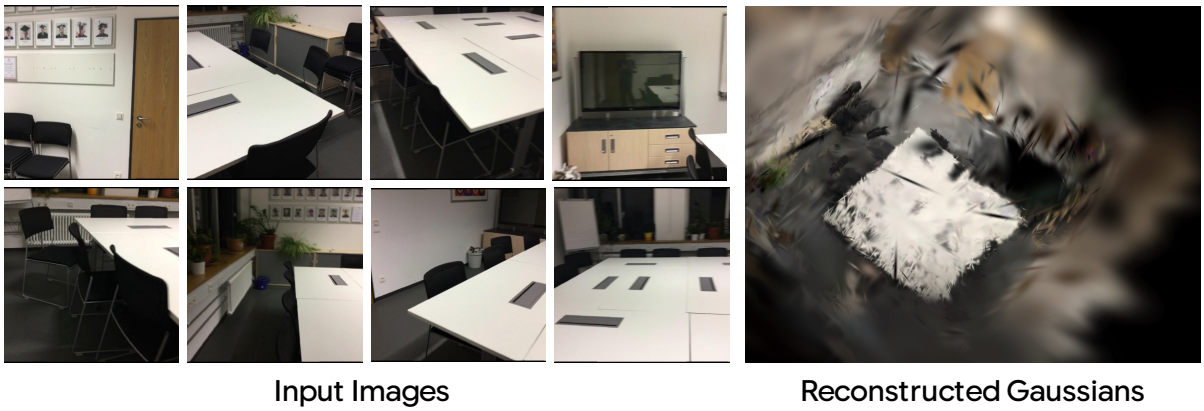

Figure 8: **Visualization of reconstructed scene with estimated Gaussians. (Top)**: input multi-view images. **(Bottom)**: a view rendered from the 3D Gaussians decoded by our Gaussian summary tokens. Although the rendering is coarse due to the compact $M = 2592$ token bottleneck, it preserves the overall scene layout and the placement of major objects, indicating that the tokens encode a coherent 3D abstraction of the scene.

D. Additional Related Works

Because our framework incorporates 3D Gaussian Splatting within an MLLM, several existing directions may appear superficially related to Imagine3D-LLM. We briefly discuss them here and clarify the key distinctions.

Feed-forward 3DGS as a feature lifter for MLLMs. SplatTalk (Thai et al., 2025) also pairs a feed-forward 3DGS estimator with an MLLM, but uses 3DGS as an external *feature lifter*: a frozen estimator (FreeSplat (Wang et al., 2024b)) lifts pretrained 2D features into a 3D Gaussian field that is then fed to the LLM as visual tokens. The MLLM itself never learns to reconstruct the scene. In contrast, Imagine3D-LLM equips the MLLM with learnable Gaussian summary tokens that jointly learn to reconstruct the scene and support language modeling, with no external 3DGS module at inference.

Per-scene 3DGS-based scene understanding. LangSplat (Qin et al., 2024), FMGS (Zuo et al., 2025), and M3 (Zou et al., 2025) distill 2D foundation features into 3D Gaussians optimized per scene, enabling open-vocabulary retrieval or segmentation. Imagine3D-LLM targets a different setting: a feed-forward generalist that answers natural-language questions about novel scenes in a single forward pass, without per-scene optimization.

Point-cloud-input MLLMs for 3D detection. SpatialLM (Mao et al., 2025) feeds 3D point clouds to an MLLM and outputs structured bounding-box coordinates and labels from a fixed vocabulary. Imagine3D-LLM instead targets free-form 3D understanding—spatial QA, situated reasoning, dense captioning, and grounding—through natural language interaction.

Latent geometric reasoning without explicit reconstruction. The most conceptually similar concurrent work is 3DThinker (Chen et al., 2025), which also argues for "thinking with 3D" before answering spatial questions. 3DThinker aligns latent reasoning tokens with features from a frozen 3D foundation model (VGGT (Wang et al.,

2025b)), placing it closer to methods that fuse geometry-foundation features into MLLMs (e.g., VLM3R (Fan et al., 2025), G^2VLM (Hu et al., 2025)). Imagine3D-LLM instead requires the Gaussian summary tokens to actually reconstruct the scene through differentiable rasterization, providing a stronger inductive bias toward 3D-aware internal representations (§3.3). Empirically, Imagine3D-LLM outperforms 3DThinker-7B by 5.2 points on SPAR-Bench overall, with larger margins on the medium- and high-level reasoning splits.

E. Limitations

While Imagine3D-LLM demonstrates that abstract reconstruction is an effective inductive bias for 3D-aware MLLMs, training efficiency remains a limitation. As discussed in § 3.2.2, our framework can learn Gaussian summary tokens without a pretrained compact Gaussian teacher (Veicht et al., 2026), but requires more training steps to achieve comparable downstream performance. The teacher therefore accelerates convergence rather than being a necessary component of our framework. Improving the efficiency of teacher-free training remains an important direction for future work.

Another limitation concerns the scope of our training scenes and the fixed representation budget. Our model is trained on indoor scenes, where a fixed number of Gaussian summary tokens and decoded Gaussians has not posed a major constraint in our experiments. However, larger and more complex outdoor scenes may require a greater representation capacity, and the current Gaussian budget could be insufficient to capture their spatial structure. Extending our framework to cover both indoor and outdoor environments, and dynamically adapting the number of Gaussian summary tokens to the scale and complexity of each scene, are promising directions for future work.