COG2GEN3D: SCULPTURING 3D SEMANTIC-GEOMETRIC COGNITION FOR 3D GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models have achieved success in producing semantically plausible 2D images, but it remains challenging in 3D generation due to the absence of spatial geometry constraints. Typically, existing methods utilize geometric features as conditions to enhance spatial awareness. However, these methods can only model relative relationships and are prone to scale inconsistency of absolute geometry. Thus, we argue that semantic information and absolute geometry empower 3D cognition, thereby enabling controllable 3D generation for the physical world. In this work, we propose Cog2Gen3D, a 3D cognition-guided diffusion framework for 3D generation. Our model is guided by three key designs: 1) **Cognitive Feature Embeddings.** We encode different modalities into semantic and geometric representations and extract logical representations. 2) **3D Latent Cognition Graph.** We structure representations into dual-stream semantic-geometric graphs and fuse them via common-based cross-attention to obtain a 3D cognition graph. 3) **Cognition-Guided Latent Diffusion.** We leverage the fused 3D cognition graph as the condition to guide the latent diffusion process for 3D Gaussian generation. Under this unified framework, the 3D cognition graph ensures the physical plausibility and structural rationality of 3D generation. Moreover, we construct a validation subset based on the Marble World Labs. Extensive experiments demonstrate that our Cog2Gen3D significantly outperforms existing methods in semantic fidelity and geometric plausibility. Our code will be released upon acceptance.

1 INTRODUCTION

Generative models have successfully synthesized semantically plausible 2D images, which are widely used in image editing (Zhang et al., 2023) and video generation (Blattmann et al., 2023). When applying these models to the physical world, some researchers rely on the semantic priors of 2D diffusion models and extend them to 3D generation through Score Distillation Sampling (Poole et al., 2022; Wang et al., 2023a), as illustrated in Fig. 1 (a). However, this paradigm is prone to structural collapse due to the lack of geometric constraints required by the physical world. Our purpose is to integrate geometric and semantic representations for high-quality 3D generation.

As shown in Fig. 1 (b), existing methods mainly incorporate geometric priors such as scene graphs (Wang et al., 2019; Dhamo et al., 2021; Zhai et al., 2023; Lin & Mu, 2024; Zhai et al., 2024) and layouts (Li et al., 2019; Bahmani et al., 2023; Po & Wetzstein, 2024; Yang et al., 2024; Chen et al., 2025) to enhance spatial awareness for 3D generation. For example, EchoScene (Zhai et al., 2024) utilizes scene graphs to define relational dependencies between objects, while Layout2Scene (Chen et al., 2025) relies on bounding box layouts to guide the coarse spatial arrangement of the scene. However, these approaches are limited to modeling 2D relative spatial relationships and fail to capture 3D absolute geometry. This leads to scale inconsistency and geometric collapse, making it difficult to satisfy the rigid constraints of the physical world. Therefore, we argue that the key to 3D generation for the physical world is integrating high-level semantics and absolute geometry.

To address the above issues, we propose that high-level semantics and 3D absolute geometry empower 3D cognition and enhance the spatial awareness of 3D generation. As illustrated in Fig. 1 (c), our design is grounded in two key insights: 1) Geometric features provide *essential physical plausibility*. Since geometric encoders capture dense correspondence and absolute metric information, we leverage these cues to enforce physical constraints and scale consistency. 2) Latent scene

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

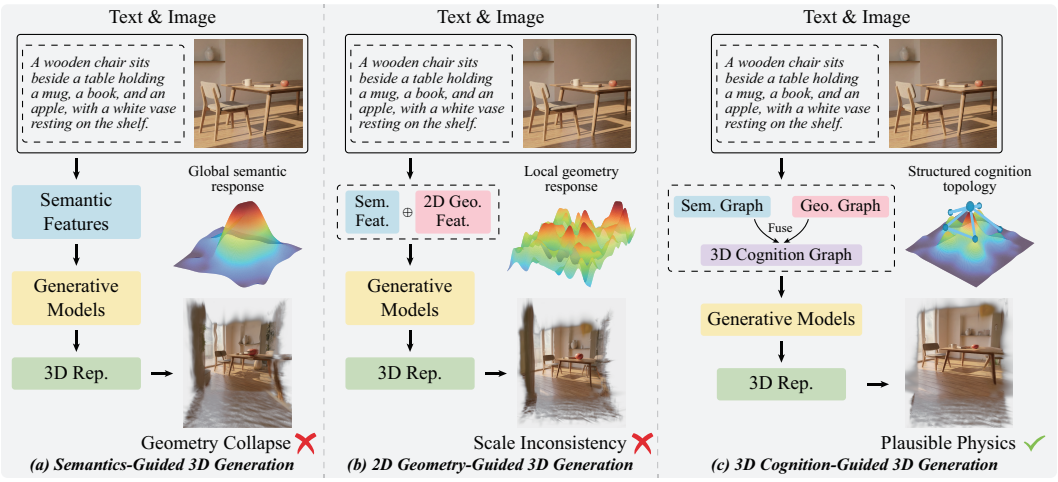


Figure 1: **Paradigm shift in 3D generation.** Semantic-guided methods rely on 2D priors and often produce physical violations such as object intersections. Geometry-guided methods add relative constraints but lack absolute metric awareness, leading to scale inconsistency. Our paradigm conditions generation on a 3D cognition graph, improving fidelity and physical plausibility.

graphs provide *structural rationality* and *superior reasoning robustness*. Their ability to robustly infer topological relationships in a latent space inspires our graph module to ensure coherent scene structures. Consequently, we design a common-based fusion integrating semantic and geometric knowledge to represent 3D cognition, enabling controllable 3D generation for the physical world.

In this work, we propose **Cog2Gen3D**, a novel 3D cognition-guided diffusion framework for 3D generation. Our model consists of three key components: 1) **Cognitive Feature Embeddings**. We encode different input modalities into semantic and geometric representations and further extract logical representation to serve as high-level guidance. 2) **3D Latent Cognition Graph**. We structure these representations into dual-stream semantic-geometric graphs and utilize the logical representation as a bridge to fuse them via common-based cross-attention, obtaining a unified 3D cognition graph. 3) **Cognition-Guided Latent Diffusion**. We leverage the 3D cognition graph as the condition to guide the diffusion process for 3D Gaussian generation. Through this design, our approach effectively awakens 3D cognition, ensuring that the generated scenes possess both high-fidelity semantics and plausible geometry. Our main contributions are summarized as follows:

- We propose **Cog2Gen3D**, an innovative framework that introduces 3D cognition to guide 3D generation, effectively bridging semantic priors with geometric constraints to support controllable 3D object and scene generation from visual and textual prompts.
- We observe that geometric features provide geometric consistency and latent graphs offer structural rationality. This motivates cognitive embeddings and a 3D latent cognition graph for sculpting robust 3D cognition that captures appearance attributes and spatial interactions.
- We develop a cognition-guided latent diffusion mechanism that steers 3D Gaussian generation using the cognition graph, ensuring semantic fidelity and geometric plausibility of the 3D scenes.
- We integrate a series of 3D datasets and construct a curated validation subset, collectively named the cognition scene graph 3D dataset (**CogSG-3D**), along with explicit scene graph labels for supervision. Extensive experiments have shown the significant effectiveness of our method.

2 Related Works

Semantics-Guided 3D Generation. 2D diffusion models provide rich semantic priors for 3D generation (Poole et al., 2022; Wang et al., 2023a; Chen et al., 2023; Wang et al., 2023b; Lin et al., 2023; Yi et al., 2024). For instance, DreamFusion (Poole et al., 2022) lifts 2D priors via Score Distillation Sampling (SDS) to optimize 3D models. However, these approaches often suffer from geometric collapse since they treat 3D generation as multi-view 2D in-painting, lacking intrinsic spatial perception. We introduce geometric cues to ensure appearance fidelity and structural rationality.

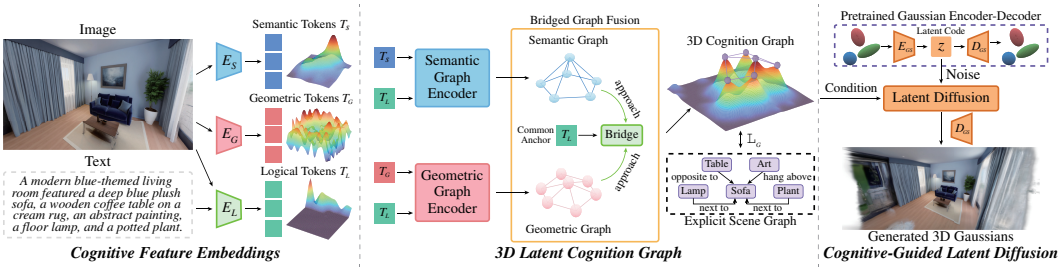


Figure 2: **Overview of Cog2Gen3D.** The model first extracts three types of cognitive tokens, representing semantic, geometric, and logical information (T_S, T_G, T_L). These tokens are structured into a 3D cognition, where the logical tokens act as a bridge for semantic-geometric alignment. Finally, the 3D cognition steers a latent diffusion process for 3D Gaussians.

2D Geometry-Guided 3D Generation. To mitigate structural inconsistencies, recent approaches utilize 2D geometric priors like scene graphs (Wang et al., 2019; Dhamo et al., 2021; Zhai et al., 2023; Lin & Mu, 2024) and layouts (Li et al., 2019; Bahmani et al., 2023; Po & Wetzstein, 2024; Yang et al., 2024) for global composition. For instance, EchoScene (Zhai et al., 2024) and Layout2Scene (Chen et al., 2025) respectively use topological structures and bounding boxes to constrain relative positioning. While these representations provide reasonable blueprints, they struggle to ensure physical plausibility and structural fidelity. This fundamentally arises from modeling 2D relative relationships rather than 3D absolute metric geometry. Consequently, our work captures 3D absolute geometry to sculpt structurally rigorous and physically plausible scenes.

Spatial Geometric Perception. Spatial geometric perception is crucial for structurally rigorous 3D generation, requiring absolute metric geometry rather than coarse topology. While models like VGGT (Wang et al., 2025) excel at encoding fine-grained 3D geometry, integrating this perception into generative models remains challenging. Existing Multimodal LLMs often rely on naive feature concatenation (Hong et al., 2023; Xu et al., 2024), failing to assimilate precise geometric constraints. Therefore, we construct a structured geometric graph fused with a corresponding semantic graph. This awakens 3D cognition, ensuring physically plausible and structurally rational 3D scenes.

3 Our Cog2Gen3D

Overview. As shown in Fig. 2, our Cog2Gen3D achieves semantically coherent and geometrically plausible 3D generation through the following three stages:

1) **Cognitive Feature Embeddings.** We transform input images I and texts T into disentangled semantic, geometric, and logical representations T_S, T_G, T_L:

$$T_S, T_G, T_L = E_{S,G,L}(I, T), \tag{1}$$

which serve as the fundamental building blocks for subsequent cognitive reasoning.

2) **3D Latent Cognition Graph.** We encode the extracted features into dual-stream graphs G_{sem}, G_{geo} and fuse them via common-based fusion with logical tokens T_L as a bridge:

$$G_{\text{sem}} = \text{GE}(T_S, T_L), \quad G_{\text{geo}} = \text{GE}(T_G, T_L), G_{\text{cog}} \quad = \text{ComFusion}(G_{\text{sem}}, G_{\text{geo}}, T_L). \tag{2}$$

The 3D cognition graph G_{cog} precisely captures both the extrinsic semantics and intrinsic geometry.

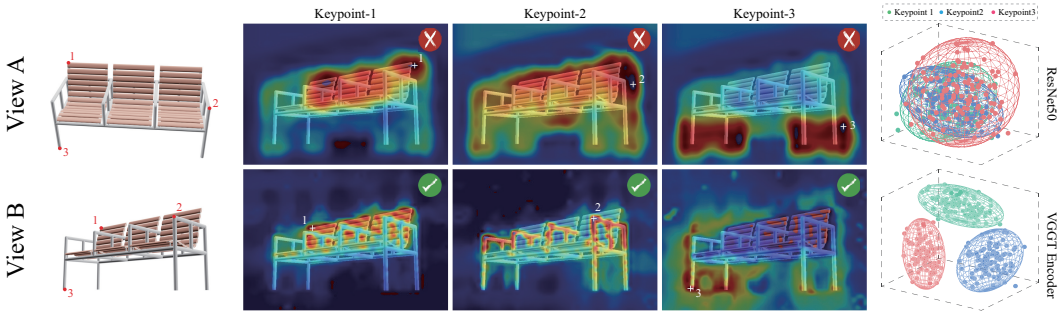
3) **Cognition-Guided Latent Diffusion.** The 3D cognition graph acts as a structural condition for latent diffusion, generating 3D Gaussians $\hat{\mathcal{G}}$ with high-fidelity appearance and rational structure:

$$\hat{\mathbf{z}}_0 = \text{LDM}(\mathbf{z}_T \mid G_{cog}), \hat{\mathcal{G}} = D_{GS}(\hat{\mathbf{z}}_0). \tag{3}$$

Remarks. The generated 3D Gaussians $\hat{\mathcal{G}}$ represent scenes with rational appearance and structure.

3.1 Cognitive Feature Embeddings

While conventional generative models rely predominantly on semantic priors, we establish a 3D cognition foundation by integrating semantics, geometry, and logical constraints.



(a) Input Images (b) Heatmap Visualization of Attention Maps (c) t-SNE Results

Figure 3: **Cross-view feature correspondence analysis.** The results demonstrate that the VGGT encoder has superior cross-view geometric consistency. This validates its capability of capturing absolute geometry cues, which motivates us to introduce the VGGT encoder as our geometric expert.

Semantic Encoder(E_S). To capture visual appearance, we use pre-trained ResNet50 (He et al., 2016) as E_S to extract high-level image features and project them into Semantic Tokens T_S, ensuring high-fidelity appearance in the generated 3D Gaussians.

Geometric Encoder (E_G). In Fig. 3, we evaluate ResNet50 (He et al., 2016) and VGGT (Wang et al., 2025) via cross-view attention between View A keypoints and View B spatial maps. While ResNet50 struggles with feature drift, VGGT distinctly separates keypoints, demonstrating robust cross-view consistency. Thus, VGGT serves as E_G, providing essential geometric grounding for rigorous structural plausibility and metric accuracy.

Logical Encoder (E_L). To bridge raw features and structural reasoning, we adopt CLIP ViT and CLIP Text encoders (Radford et al., 2021) as the logical encoder E_L. By processing image-text pairs, E_L extracts Logical Tokens T_L that encapsulate high-level relational contexts and abstract concepts, serving as the essential guidance for constructing the 3D cognition of the scene.

Tokens $\{T_S, T_G, T_L\}$ define the feature space for 3D Latent Cognition Graph construction.

3.2 3D Latent Cognition Graph

As Fig. 4 shows, explicit scene graphs suffer from severe feature divergence under noisy prompts. Our 3D latent cognition graph uses a dual-stream encoder and common-based fusion to implicitly sculpt cognitive tokens into a noise-resistant and structurally rigorous representation.

Dual-Stream Latent Graph Encoder. We construct semantic and geometric graphs sharing a topological encoding paradigm but differing in tokens and positional embeddings.

Semantic Graph Construction. The semantic query Q_S concatenates tokens T_S with a 2D positional embedding $\text{PE}(x_p, y_p)$. Cross-attention between Q_S and T_S extracts n initial nodes N_i^S. Logical tokens T_L guide this attention to formulate semantic edges E_{ij}^S. An MLP updates the nodes using messages m_{ij}^S to obtain the semantic graph G_{sem}:

$$m_{ij}^S = \text{MLP}(N_i^S, N_j^S, E_{ij}^S), \quad G_{sem} = N_i^S + \sum_j m_{ij}^S. \tag{4}$$

Geometric Graph Construction. To model absolute 3D metrics, the geometric query Q_G employs a 3D positional embedding $\text{PE}(x_q, y_q, z_q)$, where z_q is a learnable embedding capturing spatial geometry. Following the semantic pipeline, cross-attention extracts nodes N_i^G and utilizes T_L to formulate edges E_{ij}^G. An MLP then updates them to produce G_{geo}. Driven by z_q, the geometry graph G_{geo} effectively models absolute 3D metric relationships between entities.

Common-based Cross-Attention Fusion. Semantic and geometric graphs capture coherent appearance and precise geometry, respectively, in separate feature spaces. Both use logical tokens T_L to formulate edges E_{ij}, providing a common topological foundation.

Treating T_L as a unifying anchor, we project it into a logical query $\mathbf{Q}_L$. Semantic and geometric nodes are concatenated along the feature dimension to formulate keys $\mathbf{K}_{SG}$ and values $\mathbf{V}_{SG}$:

$$\mathbf{Q}_L = T_L \mathbf{W}_Q, \mathbf{K}_{\text{SG}} = \left[\mathbf{G}_{\text{sem}} \parallel \mathbf{G}_{\text{geo}}\right] \mathbf{W}_K, \mathbf{V}_{\text{SG}} = \left[\mathbf{G}_{\text{sem}} \parallel \mathbf{G}_{\text{geo}}\right] \mathbf{W}_V. \tag{5}$$

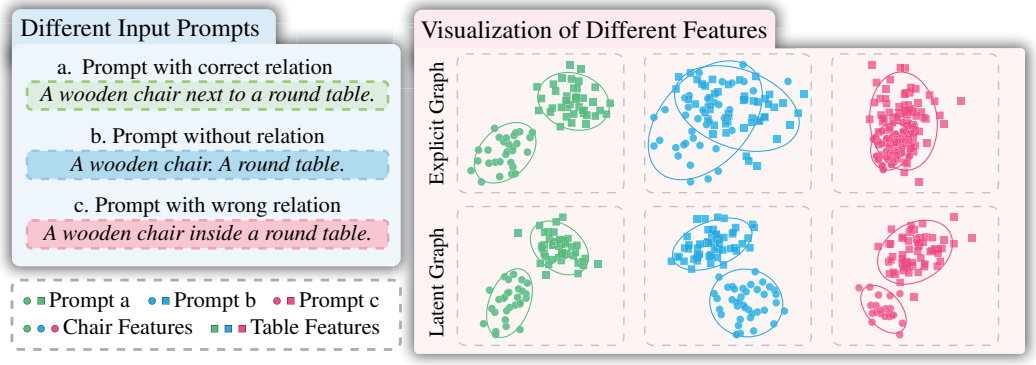


Figure 4: **Feature distributions under prompt perturbations.** Explicit graphs diverge significantly given missing or incorrect relations. Conversely, our latent scene graph maintains a stable, distinctly clustered distribution, demonstrating superior robustness to noisy prompts.

where $[\cdot||\cdot]$ denotes concatenation, and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable linear projections. The 3D cognition graph nodes $\mathbf{N}^{cog}$ are computed via scaled dot-product cross-attention:

$$\mathbf{N}^{cog} = \text{Softmax}\left(\frac{\mathbf{Q}_L \mathbf{K}_{SG}^T}{\sqrt{d_k}}\right)\mathbf{V}_{SG}, \tag{6}$$

where d_k is the scaling factor representing the channel dimension. Here, the logical query $\mathbf{Q}_L$ acts as a bridge, adaptively aligning corresponding semantic textures and structural constraints. Finally, a residual connection with LayerNorm (LN) and an MLP stabilize the fused representation:

$$\hat{\mathbf{N}}^{\text{cog}} = \text{LN}(T_L + \mathbf{N}^{\text{cog}}), \quad \mathbf{G}_{\text{cog}} = \text{MLP}(\hat{\mathbf{N}}^{\text{cog}}) + \hat{\mathbf{N}}^{\text{cog}}. \tag{7}$$

Ultimately, this fusion yields a holistic 3D cognition graph $\mathbf{G}_{cog}$ that balances semantic coherence with geometric rationality to structurally guide the 3D generative process.

3.3 Cognition-Guided Latent Diffusion

We generate semantically faithful and geometrically plausible 3D scenes through cognition-guided latent diffusion in a compressed 3D Gaussian space.

Conditioned Diffusion Process. To avoid the prohibitive cost of diffusion on explicit 3D representations, we operate in latent space. The forward process progressively adds Gaussian noise ϵ to ground-truth $\mathbf{z}_0$, yielding noisy latent $\mathbf{z}_t$ at timestep t. During reverse denoising, our fused cognition graph $\mathbf{G}_{cog}$ replaces text or layouts as the structural condition for noise prediction:

$$\hat{\epsilon} = \epsilon_\theta(\mathbf{z}_t, t, \mathbf{G}_{cog}). \tag{8}$$

$\mathbf{G}_{cog}$ mitigates ambiguity and layout distortion associated with 2D-prior-based generation.

Latent-Gaussian Encoder-Decoder. We use a pre-trained latent-Gaussian encoder-decoder (Fig. 2). The encoder E_{GS} compresses Gaussians $\mathcal{G}$ into compact latent $\mathbf{z}_0 = E_{GS}(\mathcal{G})$ for forward noising; after conditioned denoising, the decoder D_{GS} reconstructs $\hat{\mathcal{G}} = D_{GS}(\hat{\mathbf{z}}_0)$. This symmetric architecture efficiently learns complex spatial distributions for high-fidelity generation.

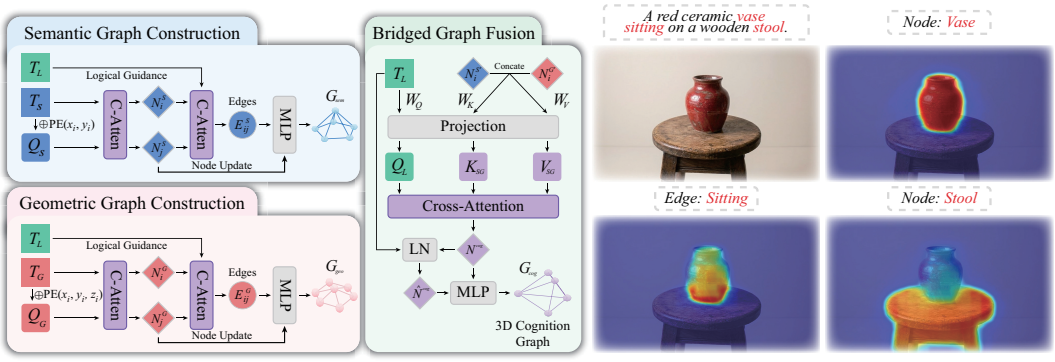
3.4 Optimization

The training loss of our Cog2Gen3D consists of three components: a latent diffusion loss $\mathcal{L}_{diff}$, an explicit node grounding loss $\mathcal{L}_G$, and a 3D Gaussian reconstruction loss $\mathcal{L}_{recon}$.

Latent Diffusion Loss $\mathcal{L}_{diff}$. To train the denoising network, we employ the standard diffusion objective, minimizing MSE between the added noise ϵ and predicted noise $\hat{\epsilon}$:

$$\mathcal{L}_{diff} = \mathbb{E}_{\mathbf{z}_0,\epsilon,t}\left[||\epsilon - \hat{\epsilon}||_2^2\right]. \tag{9}$$

Explicit Node Grounding Loss $\mathcal{L}_g$. To avoid bottlenecking spatial reasoning with sparse explicit relations, we discard edge supervision and solely anchor node semantics. To bridge the density gap between latent nodes and K explicit entities, we dynamically select the top-K latent nodes via



(a) Cognition Graph Construction Modules (b) Visualization of Cognition Graph Attention

Figure 5: **Architectural details and interpretability of the 3D Latent Cognition Graph.** (a) The pipeline for constructing the cognition graph from cognitive tokens. (b) The correspondence visualization showing how graph components precisely align with 3D entities and spatial boundaries.

cross-attention ranking. A classification head yields semantic probabilities p_i^{sem}, supervised using a Cross-Entropy loss against the explicit labels y_i^{sem}:

$$\mathcal{L}_g = \frac{1}{K}\sum_{i=1}^{K} \mathrm{CE}(p_i^{sem}, y_i^{sem}). \tag{10}$$

This minimalist top-K supervision ensures semantic fidelity while granting the latent edges absolute freedom to autonomously infer complex 3D topology.

3D Gaussian Reconstruction Loss $\mathcal{L}_{recon}$. To guarantee metric precision and visual fidelity, we supervise the Gaussians $\hat{\mathcal{G}}$ using the ground-truth $\mathcal{G}$. This supervision is applied in the image space to ensure multi-view consistency. For a set of viewpoints v, we render the predicted Gaussians into images $\hat{I}_v$ and compare them against the GT renders I_v, calculating L1 loss and D-SSIM loss:

$$\mathcal{L}_{\text{recon}} = \sum_{v} \Big[\lambda_{\text{L1}} \|\hat{I}_v - I_v\|_1 + \lambda_{\text{ssim}}\big(1 - \text{SSIM}(\hat{I}_v, I_v)\big)\Big] \tag{11}$$

Total Loss. The entire framework is jointly optimized end-to-end using a weighted sum of the three losses. The training objective is formulated as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{diff} + \lambda_2 \mathcal{L}_g + \lambda_3 \mathcal{L}_{recon}, \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters that balance the relative contributions of diffusion fidelity, structural cognition, and visual reconstruction during end-to-end optimization.

4 Training Datasets and Pipeline

4.1 Our CogSG-3D Dataset

As Fig. 6 shows, to effectively train Cog2Gen3D, we propose CogSG-3D, a comprehensive dataset aggregating prominent public 3D datasets and self-built data from Marble World Labs (World Labs, 2025). It covers object-level (ShapeNet (Chang et al., 2015), Objaverse-XL (Deitke et al., 2023), OmniObject3D (Wu et al., 2023), Pix3D (Sun et al., 2018), ABO (Collins et al., 2022), ModelNet (Wu et al., 2015)) and scene-level (ScanNet (Dai et al., 2017), 3D-Front (Fu et al., 2021)) categories, ensuring diverse semantic and geometric distributions. To standardize these heterogeneous formats, our pipeline encompasses three steps: 1) *Image prompt rendering:* Normalizing canonical poses to render 2D multi-view visual prompts. 2) *Text prompt acqui-*



(a) CogSG-3D Dataset Statistical (b) Self-built Dataset Visualization

Figure 6: **Statistics and examples of our CogSG-3D dataset.** (a) Dataset composition across object- and scene-level sources. (b) Representative self-built samples with text prompts, images, scene graphs, and 3D Gaussians.

Table 1: Text-to-3D comparison on T3Bench.

Method	T3Bench (He et al., 2023)			
	Single Obj. ↑	Single w/ Surr. ↑	Multi Obj. ↑	Average ↑
DreamFusion (Poole et al., 2022)	24.4	19.8	11.7	18.7
Magic3D (Lin et al., 2023)	37.0	35.4	25.7	32.7
SJC (Wang et al., 2023a)	24.7	19.8	11.7	18.7
Fantasia3D (Chen et al., 2023)	26.4	27.0	18.5	24.0
ProlificDreamer (Wang et al., 2023b)	49.4	44.8	35.8	43.3
GaussianDreamer (Yi et al., 2024)	54.0	48.6	34.5	45.7
Ours	**58.3**	**57.9**	**53.6**	**56.6**

Table 2: Image-to-3D object generation comparison.

Method	ShapeNet (Chang et al., 2015)			OmniObject3D (Wu et al., 2023)		
	FID ↓	KID (%) ↓	MMD (%) ↓	FID ↓	KID (%) ↓	MMD (‰) ↓
EG3D (Chan et al., 2022)	28.12	1.25	7.14	41.56	1.13	28.21
GET3D (Gao et al., 2022)	38.62	1.85	6.45	49.41	1.53	13.57
DiffRF (Müller et al., 2023)	98.53	6.14	8.59	147.59	8.82	16.07
DiffTF (Cao et al., 2023)	26.58	1.15	6.15	25.36	0.81	6.64
DiffGS (Zhou et al., 2024)	23.45	1.05	5.31	27.35	0.84	7.05
LN3Diff (Lan et al., 2024)	21.51	1.08	5.25	21.42	0.73	4.87
Ours	**14.54**	**0.67**	**3.25**	**15.94**	**0.58**	**3.02**

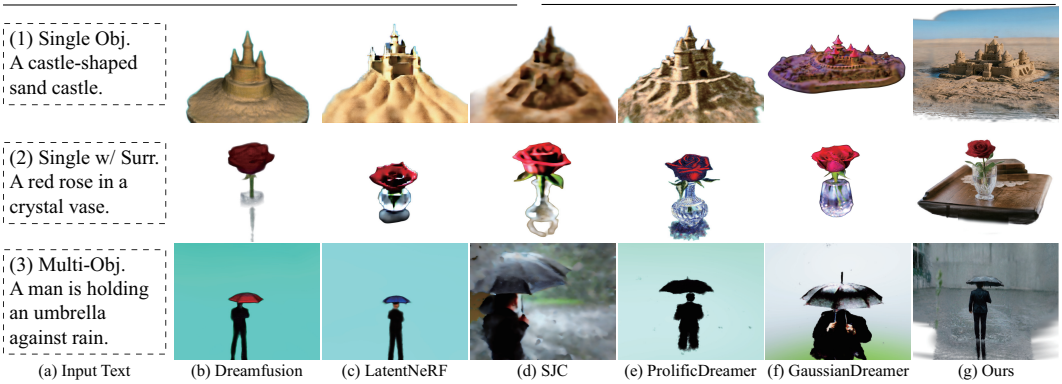


Figure 7: Visualization comparison of Text-to-3D on the T3Bench dataset.

sition: Leveraging Vision-Language Models (Gemini) to generate dense descriptive texts for unannotated data. 3) *Unified 3D Gaussian representation:* Converting 3D ground truths into optimizable 3D Gaussians to overcome traditional mesh constraints.

4.2 Training Pipeline

We design a progressive three-stage training paradigm to ensure stable model convergence and achieve optimal semantic-geometric alignment throughout the training process:

Stage 1: Geometric-Latent Alignment. This stage aims to acquire latent representations of 3D scenes and a robust decoder. We pretrain a Gaussian auto-decoder where the encoder E_{GS} maps 3D Gaussians into a compact latent variable z and the decoder D_{GS} is simultaneously optimized to flawlessly reconstruct the 3D Gaussians from z, preserving rich semantic and geometric details.

Stage 2: Cognitive-Generative Alignment. This stage targets at obtaining semantically and geometrically coherent 3D latents guided by multimodal priors. With the autoencoder frozen, we train the Latent Diffusion Model (LDM) in the z-space. The denoising network is conditioned on the 3D Latent Cognition Graph to enforce strict structural constraints.

Stage 3: End-to-End Fine-Tuning. This stage harmonizes the generative latent space with the explicit 3D rendering space. We unfreeze the decoder D_{GS} and jointly optimize it with the LDM end-to-end. This mitigates feature mismatch between synthesized latents and the decoder's expected input distribution, significantly enhancing the overall visual fidelity and geometric precision.

5 Experiments

5.1 Implementation Details

Our 3D cognition graph has 64 hidden nodes, adjustable to scene complexity. Diffusion uses $T = 1000$ training timesteps and a 50-step DDIM sampler for efficient inference. AdamW optimizes the total loss with initial learning rate 1×10^{-5} and Eq. (12) weights $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.8$. All training and testing use 8 NVIDIA A800 GPUs.

5.2 Comparison Experiments

Text-to-3D Generation. We benchmark on T3Bench (He et al., 2023) against SOTA methods (Poole et al., 2022; Wang et al., 2023a; Chen et al., 2023; Wang et al., 2023b; Lin et al., 2023; Yi et al., 2024), reporting quality scores at three complexity levels and their average. Cog2Gen3D leads every metric, with notable multi-object margins (Table 1, Fig. 7). Baselines often exhibit geometric

Table 3: Image-to-3D scenes comparison.

Method	3D-Front (Fu et al., 2021)		
	Chamfer Dist. ↓	F-Score ↑	IoU ↑
LucidDreamer (Chung et al., 2023)	0.083	50.79	0.536
SSR (Chen et al., 2024)	0.140	39.76	0.311
Gen3DSR (Ardelean et al., 2025)	0.123	40.07	0.363
REPARO (Han et al., 2025)	0.129	41.68	0.339
MIDI (Huang et al., 2025)	0.080	50.19	0.518
EchoScene (Zhai et al., 2024)	0.105	45.62	0.458
Layout2Scene (Chen et al., 2025)	0.094	48.36	0.492
Ours	**0.063**	**58.43**	**0.682**

Table 4: Cognitive tokens ablation.

Cognitive Tokens			User Study ↑		
Semantic	Geometric	Logical	SemF	GeoP	RelC
✓			4.12	2.38	2.15
	✓		2.23	4.08	2.74
		✓	2.41	2.58	3.92
✓	✓		4.38	4.35	2.82
✓		✓	4.42	2.75	4.28
	✓	✓	3.17	4.31	4.34
✓	✓	✓	**4.65**	**4.58**	**4.62**

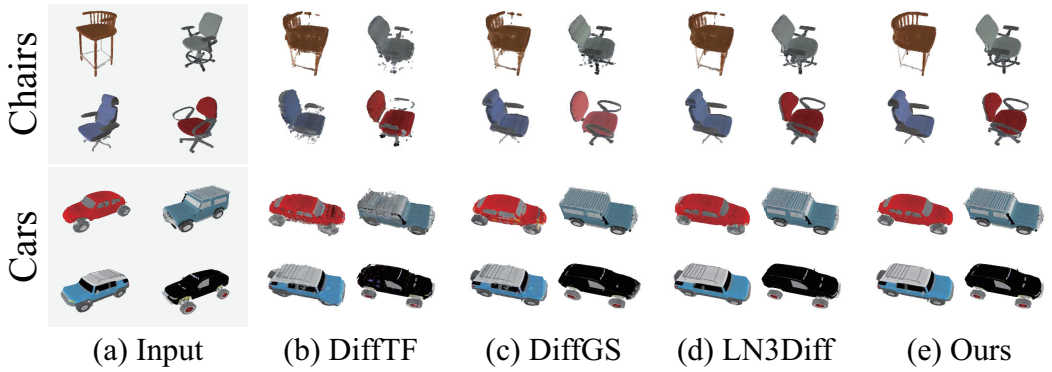


Figure 8: Visualization comparison of Image-to-3D objects on the ShapeNet dataset.

distortions or structural collapse in complex scenes; our graph constraints enable robust, high-fidelity 3D generation with precise geometries and coherent relationships.

Image-to-3D Object Generation. On ShapeNet (Chang et al., 2015) and OmniObject3D (Wu et al., 2023), Cog2Gen3D outperforms baselines (Chan et al., 2022; Gao et al., 2022; Müller et al., 2023; Cao et al., 2023; Zhou et al., 2024; Lan et al., 2024) on FID, KID, and MMD (Table 2, Fig. 8). Its cognition graph's semantic-geometric perception preserves visual fidelity and plausible geometry.

Image-to-3D Scene Generation. We evaluate complex scene generation on 3D-Front (Fu et al., 2021) and CogSG-3D against semantic (Chung et al., 2023; Chen et al., 2024; Ardelean et al., 2025; Han et al., 2025; Huang et al., 2025) and 2D geometry-guided (Zhai et al., 2024; Chen et al., 2025) baselines using Chamfer Distance (CD), F-Score, and IoU. Cog2Gen3D consistently leads all metrics (Table 3). In Fig. 10, baselines' limited spatial awareness causes chaotic layouts and structural collapse, whereas our model uses semantic and absolute geometric cognitive features to synthesize highly realistic, well-organized environments.

5.3 Ablation Study

Effectiveness of Cognitive Feature Embeddings. In the user study (Table 4), ablating semantic, geometric, or logical tokens severely degrades Semantic Fidelity (SemF), Geometric Plausibility (GeoP), and Relational Coherence (RelC), respectively. These distinct, complementary forms of cognition are all essential for high-quality 3D scene synthesis.

Superiority of Latent Graph Construction. Flattening the graph into a token sequence degrades performance (Table 5), indicating lost 3D spatial dependencies. Our graph's semantic-geometric interactions improve structural rationality and geometric plausibility.

Impact of Dual-Stream Scene Graphs. Ablating either the semantic or geometric stream noticeably degrades performance (Fig. 9). Omitting the semantic graph compromises texture fidelity, while removing the geometric graph causes severe structural distortions. Both streams are thus indispensable for high-quality, structurally coherent 3D generation.



Figure 9: Ablation of dual-stream scene graphs.

Table 5: Graph construction ablation.

Settings	3D-Front (Fu et al., 2021)		
	CD ↓	F-Score ↑	IoU ↑
w/o Graph	0.089	49.16	0.503
w/. Graph	**0.063**	**58.43**	**0.683**

Table 6: Geometry encoders.

Settings	3D-Front (Fu et al., 2021)		
	CD ↓	F-Score ↑	IoU ↑
ResNet50	0.091	48.75	0.498
CLIP ViT-L	0.076	53.88	0.591
VGGT	**0.063**	**58.43**	**0.683**

Table 7: Fusion strategy.

Settings	3D-Front (Fu et al., 2021)		
	CD ↓	F-Score ↑	IoU ↑
Concat.	0.085	50.41	0.523
Weighted	0.072	55.12	0.635
Common	**0.063**	**58.43**	**0.683**



(a) Input Image (b) Gen3DSR (c) LucidDreamer (d) Ours

Figure 10: Visualization comparison of Image-to-3D Scenes on 3D-Front and CogSG-3D.

5.4 Discussion

Correspondence between the Cognition Graph and 3D Scenes. Attention maps between generated 3D scenes and node/edge embeddings reveal high-attention regions precisely aligned with target objects and spatial boundaries (Fig. 5), confirming the graph as a structured cognitive map localizing semantic-geometric entities and linking abstract prompts to 3D scenes.

Influence of the Geometry Perception Backbone. Among geometry encoders (Table 6), VGGT (Wang et al., 2025) achieves optimal performance by excelling at extracting robust geometric features and spatial structures, providing precise absolute geometry guidance for diffusion and significantly enhancing generated scenes' geometric plausibility.

Effectiveness of Common-based Fusion Strategy. Common-based fusion aligns semantic and geometric features while preserving priors, outperforming concatenation and weighted fusion in semantic fidelity and geometric plausibility (Table 7).

Limitations. Cog2Gen3D generates high-quality, geometrically plausible 3D scenes but remains limited in dynamic 4D generation: its 3D cognition graph and Gaussian representation (Kerbl et al., 2023) only capture static constraints, lacking temporal modeling of motion or complex topological evolution. Future work will integrate spatio-temporal graphs and 4D Gaussian Splatting (Wu et al., 2024) for spatially grounded, temporally consistent dynamic scene generation.

6 Conclusion

In this work, we proposed Cog2Gen3D, a 3D cognition-guided diffusion framework sculpturing semantic-geometric cognition for high-quality 3D generation. By integrating absolute geometric priors and semantic constraints, our method resolves the challenges of scale inconsistency and the absence of spatial awareness prevalent in existing models. We introduced cognitive feature embeddings to map multi-modal information into a unified semantic-geometric space, and proposed the 3D latent cognition graph to capture complex spatial topological relationships and structural dependencies within 3D environments. Furthermore, we constructed the CogSG-3D dataset, providing extensive explicit scene graph and 3D Gaussian annotations to support training. Extensive experiments across tasks including text-to-3D, image-to-3D object, and complex scene generation demonstrate our approach's superiority and affirm its effectiveness in modeling the physical world.

AI USE STATEMENT

No generative AI tools were used in the research, implementation, experimentation, analysis, or preparation of this work. The authors take full responsibility for all content presented in this work.

REFERENCES

Andreea Ardelean, Mert Özer, and Bernhard Egger. Gen3dsr: Generalizable 3d scene reconstruction via divide and conquer from a single view. In *2025 International Conference on 3D Vision (3DV)*, pp. 616–626. IEEE, 2025.

Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas Guibas, and Andrea Tagliasacchi. Cc3d: Layout-conditioned generation of compositional 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7171–7181, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Ziang Cao, Fangzhou Hong, Tong Wu, Liang Pan, and Ziwei Liu. Large-vocabulary 3d diffusion model with transformer. *arXiv preprint arXiv:2309.07920*, 2023.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133, 2022.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Minglin Chen, Longguang Wang, Sheng Ao, Ye Zhang, Kai Xu, and Yulan Guo. Layout2scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. *arXiv preprint arXiv:2501.02519*, 2025.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22246–22256, 2023.

Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. In *2024 International Conference on 3D Vision (3DV)*, pp. 1456–1467. IEEE, 2024.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21126–21136, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16352–16361, 2021.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances in neural information processing systems*, 35:31841–31854, 2022.

Haonan Han, Rui Yang, Huan Liao, Jiankai Xing, Zunnan Xu, Xiaoming Yu, Junwei Zha, Xiu Li, and Wanhua Li. Reparo: Compositional 3d assets generation with differentiable 3d layout alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25367–25377, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T3bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023.

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. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23646–23657, 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.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *ECCV*, 2024.

Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023.

Chenguo Lin and Yadong Mu. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. *arXiv preprint arXiv:2402.04717*, 2024.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12663–12673, 2023.

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4328–4338, 2023.

Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. In *2024 International Conference on 3D Vision (3DV)*, pp. 651–663. IEEE, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

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*, pp. 8748–8763. PmLR, 2021.

Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2974–2983, 2018.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12619–12629, 2023a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023b.

World Labs. Marble: A multimodal world model. `https://www.worldlabs.ai/blog/marble-world-model`, November 2025. Accessed: 2026-03-02.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 803–814, 2023.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2024.

Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. *Advances in Neural Information Processing Systems*, 37:82060–82084, 2024.

Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6796–6807, 2024.

Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems*, 36:30026–30038, 2023.

Guangyao Zhai, Evin Pınar Örnek, Dave Zhenyu Chen, Ruotong Liao, Yan Di, Nassir Navab, Federico Tombari, and Benjamin Busam. Echoscene: Indoor scene generation via information echo over scene graph diffusion. In *European Conference on Computer Vision*, pp. 167–184. Springer, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. Diffgs: Functional gaussian splatting diffusion. *Advances in Neural Information Processing Systems*, 37:37535–37560, 2024.

APPENDIX

In this supplementary material, we provide more implementation details of our 3D latent cognition graph module, Gaussian encoder and decoder, and latent diffusion model in Sec. A. Then we conduct additional ablation experiments and discussions and further explain the limitations in Sec. B. Moreover, we demonstrate more visualization results of text-to-3D generation, image-to-3D generation and the results of generalization experiments in Sec. C.

A IMPLEMENTATION DETAILS

A.1 GAUSSIAN ENCODER AND DECODER

As shown in Fig. 11, we use a simplified yet effective Variational Autoencoder (VAE) to compress unstructured 3D Gaussian Splatting (3DGS) primitives into a compact latent space and reconstruct them with high fidelity, without the exorbitant computational cost of deep transformer stacks.

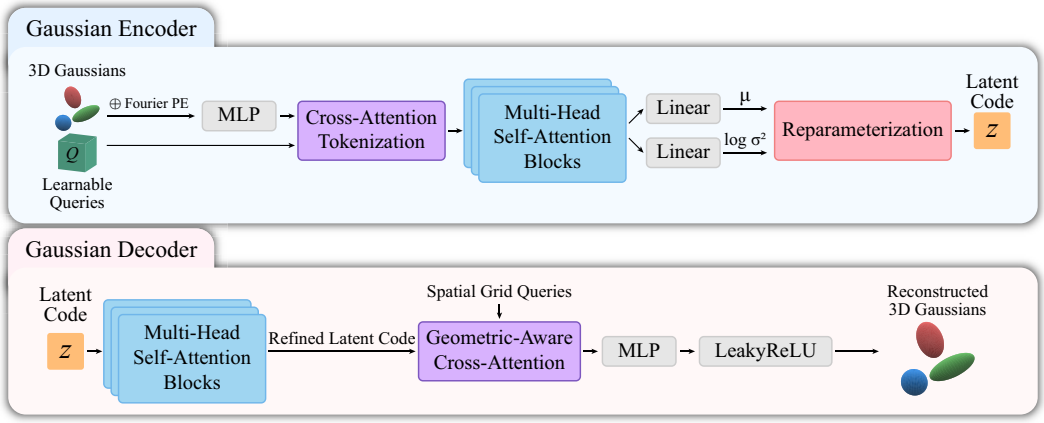


Figure 11: Implementation details of Gaussian encoder and decoder.

Gaussian Encoder. The encoder takes the raw 3DGS parameters as input. To preserve high-frequency geometric details, we first apply Fourier positional encoding to the spatial coordinates of each Gaussian. These augmented features are then independently mapped to a higher dimension using a 2-layer MLP. To effectively aggregate the varying number of unordered Gaussian primitives, we tokenize the scene using a Cross-Attention module. Specifically, we initialize $K = 512$ learnable canonical queries, which attend to the MLP-encoded Gaussian features to extract a fixed-size, compact representation. This tokenized sequence is subsequently processed by a shallow Transformer encoder consisting of 4 Multi-Head Self-Attention (MHSA) blocks to capture global semantic-geometric correlations. Finally, two parallel linear layers map the output tokens to the latent mean μ and log-variance $\log \sigma^2$, from which the latent code z is sampled via the reparameterization trick.

Gaussian Decoder. The decoder symmetrically reconstructs the continuous 3DGS parameters from the sampled latent code z. The latent tokens are first refined through another 4-layer MHSA block to restore spatial and semantic structural dependencies. To project these discrete latents back into the continuous 3D space, we utilize a geometric-aware decoding strategy. Instead of relying on deep autoregressive or complex continuous querying mechanisms, we deploy a final Cross-Attention module. Here, a set of target queries—anchored to a coarse predefined $16 \times 16 \times 16$ spatial grid—attends to the refined latent tokens z. The output of this cross-attention layer is directly fed into a shared 3-layer MLP equipped with LeakyReLU activations. This tail MLP maps the aggregated features to the final continuous attributes of the reconstructed 3DGS primitives.

A.2 LATENT DIFFUSION MODEL

To circumvent the computational bottleneck of high-resolution generation, our diffusion process operates entirely within a compact Gaussian latent space.

Latent Structural Conditioning. The cognition graph $\mathbf{G}_{cog}$ is explicitly instantiated as a sequence of rendered warped frames and their corresponding valid masks. To align with the latent diffusion paradigm, these inputs are projected into the compact latent space using a pre-trained encoder. The resulting encoded features are concatenated along the channel dimension and zero-padded to form a unified latent structural condition, $\mathbf{c}_{cog}$, which is injected into the denoising network as guidance.

Denoising Network Backbone. The denoising network ϵ_θ is parameterized by a scalable Diffusion Transformer. The noisy latent $\mathbf{z}_t$, timestep t, and the structural condition $\mathbf{c}_{cog}$ are patchified and processed through a series of multi-head self-attention and cross-attention blocks. This transformer-based formulation enables the network to effectively capture the long-range temporal dependencies and complex spatial correlations required for high-fidelity scene dynamics.

B Ablation Study and Discussion

B.1 Effectiveness of 3D Position Embedding

To validate the critical role of explicit 3D spatial priors in establishing the geometric cognition capabilities of Cog2Gen3D, we conduct an ablation study on the 3D position embedding. For the experimental setup, we compare two distinct configurations: completely omitting the position embedding and our default setting where the 3D position embedding is explicitly integrated into our proposed geometric graph. Table 8 reveals that the absence of spatial priors within the Geometric Graph leads to a severe degradation in the reconstruction quality of unstructured primitives. Conversely, equipping the Geometric Graph with the 3D position embedding achieves optimal generative performance. This thoroughly demonstrates that incorporating explicit 3D coordinates into the Geometric Graph is essential for the network to capture spatial dependencies.

Table 8: Effectiveness of 3D position embedding.

Settings	3D-Front (Fu et al., 2021)		
	Chamfer Distance ↓	F-Score ↑	IoU ↑
w/o 3D Position Embedding	0.082	50.04	0.515
w/. 3D Position Embedding	**0.063**	**58.43**	**0.683**

B.2 Impact of the Number of Nodes in Cognition Graph

To evaluate the representational capacity of our Cognition Graph, we analyze the impact of the number of implicit nodes. As shown in Table 9, there is a clear performance trade-off governed by this count. Insufficient nodes restrict the network's degrees of freedom, limiting its ability to model intricate spatial layouts and fine geometric details. Conversely, excessive nodes introduce redundant computational overhead and increase the risk of overfitting, yielding diminishing returns or even performance degradation. Thus, an optimal intermediate node count best balances capacity and generalization, effectively facilitating the sculpturing of 3D semantic-geometric cognition.

B.3 Influence of the Inference Step of Diffusion

We analyze the influence of inference steps on 3D generation quality to understand our model's generation dynamics. As shown in Table 10, varying denoising steps reveals a clear fidelity-efficiency trade-off. Insufficient steps lead to incomplete denoising and a noticeable performance drop, failing to adequately map noise to the structured latent space. Conversely, excessive steps substantially increase inference time with only marginal visual gains. Thus, we set $T = 50$ as the optimal balance, ensuring high-fidelity sculpturing of 3D semantic-geometric cognition without unnecessary computational overhead associated with excessive diffusion steps.

Table 9: Impact of the number of nodes in cognition graph.

Number of Graph Nodes	3D-Front (Fu et al., 2021)		
	Chamfer Distance ↓	F-Score ↑	IoU ↑
16	0.104	41.26	0.437
32	0.096	45.83	0.464
48	0.071	53.42	0.592
64	**0.063**	**58.43**	**0.683**
80	0.065	57.93	0.667
96	0.066	57.18	0.641

Table 10: Influence of the inference step of diffusion.

Inference Step	3D-Front (Fu et al., 2021)			Inference Time
	Chamfer Distance ↓	F-Score ↑	IoU ↑	
25	0.083	47.36	0.524	105s
50	**0.063**	**58.43**	**0.683**	132s
75	0.063	58.43	0.683	155s
100	0.063	58.43	0.683	187s
200	0.063	58.43	0.683	274s

B.4 Weight Sensitivity of Model Loss

To investigate the impact of different loss components on training dynamics, we analyze the weight sensitivity of our multi-objective loss function, as shown in Fig. 12. The primary objective of this experiment is to identify a weight configuration that enables the total loss to descend as rapidly as possible while converging to the absolute minimum. Our overall loss comprises the diffusion loss $\mathcal{L}_{diff}$, the auxiliary loss $\mathcal{L}_g$, and the reconstruction loss $\mathcal{L}_{recon}$, which are modulated by the hyper-parameters λ_1, λ_2, and λ_3, respectively. Empirical evaluations demonstrate that setting $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.8$ yields the optimal training dynamics. This specific configuration achieves the fastest convergence rate and the lowest final loss value, ensuring a robust balance between the core generation process and necessary auxiliary regularization.

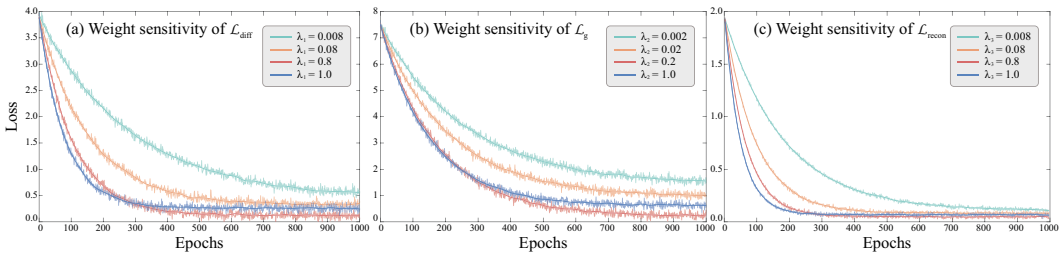


Figure 12: The weight sensitivity of model losses.

B.5 Discussion on Computational Efficiency

To evaluate the practical applicability of our approach, we compare its computational efficiency against recent baselines on the 3D-Front (Fu et al., 2021) dataset, as summarized in Table 11. Methods like Gen3DSR (Ardelean et al., 2025) and MIDI (Huang et al., 2025) require less memory but perform poorly. Optimization-heavy frameworks such as EchoScene (Zhai et al., 2024) and Lay-

out2Scene (Chen et al., 2025) demand substantial computational resources and prolonged inference times. In contrast, our proposed method effectively balances generation quality and resource utilization. Ultimately, these results demonstrate that our framework maintains highly competitive computational efficiency while achieving state-of-the-art 3D scene generation performance.

Table 11: Discussion on computational efficiency.

Methods	Memory Consumption (GB)	Inference Time (min)
Gen3DSR (Ardelean et al., 2025)	16.74	8.7
REPARO (Han et al., 2025)	21.56	4.1
MIDI (Huang et al., 2025)	12.53	1.2
EchoScene (Zhai et al., 2024)	38.72	38.6
Layout2Scene (Chen et al., 2025)	45.28	56.6
Ours	40.32	3.5

B.6 Limitations and Possible Solutions

While our proposed Cog2Gen3D demonstrates strong performance in generating high-quality and geometrically plausible static 3D scenes, it still faces several challenges. These challenges primarily stem from inherent design choices in the architecture and the nature of the generative priors, which can impact its applicability in highly dynamic or massive-scale real-world environments. Below, we elaborate on the key limitations and outline potential directions for future research to address them.

A core component of our model for capturing structural layouts is the 3D Semantic-Geometric Cognition graph, which aggregates spatial information to ensure structural consistency. However, this static design introduces constraints that may not adequately handle all types of environments or motions. Specifically, we analyze the following possible limitations on model performance:

Absence of Temporal Modeling. As our cognition graph and the 3D Gaussian Splatting (Kerbl et al., 2023) representation are designed exclusively for static constraints, the framework currently struggles with dynamic 4D generation. It fails to capture motion, topological evolution, or complex physical interactions as a dynamic scene evolves over time.

Scalability Constraints. When scaling from bounded indoor scenes (e.g., the 3D-Front dataset) to unbounded or massive environments, the complexity of the cognition graph and the required number of Gaussian primitives grow significantly, leading to potential memory and optimization bottlenecks.

Dependency on Diffusion Priors. The generative diversity and geometric accuracy are inherently bounded by the distribution biases of the underlying pre-trained diffusion models. Highly unconventional or out-of-distribution (OOD) layouts may occasionally result in geometric or textural artifacts.

To address these limitations, we propose several avenues for future research:

Spatio-Temporal Extension. To enable dynamic scene generation, we plan to extend our framework into a Spatio-Temporal Cognition Graph and upgrade the rendering to 4D Gaussian Splatting (Wu et al., 2024). We will also explore memory-efficient dynamic rendering to ensure that training remains computationally feasible on standard GPUs.

Hierarchical Modeling and LoD. To address scalability, future iterations will introduce a hierarchical cognition graph to decouple macro-level layout planning from micro-level detailing, with adaptive level-of-detail (LoD) pruning of Gaussian primitives to strictly control memory use.

Robust Priors and Test-Time Adaptation. To mitigate dataset biases, we aim to integrate stronger multi-modal geometric priors (e.g., advanced depth and normal estimators) into the optimization loop, and explore zero-shot test-time adaptation techniques to enhance OOD generalization.

C Comparison Experiments

C.1 Text-to-3D Generation

In Fig. 13 and Fig. 14, we present the visual results of text-to-3D generation on T3Bench (He et al., 2023) and our CogSG-3D datasets. We compare our Cog2Gen3D with SOTA methods (Poole et al., 2022; Lin et al., 2023; Wang et al., 2023a; Chen et al., 2023; Wang et al., 2023b; Yi et al., 2024). By sculpturing robust 3D semantic-geometric cognition, our Cog2Gen3D successfully disentangles complex object compositions and synthesizes fine-grained geometric details. Consequently, our approach achieves state-of-the-art text-to-3D generation performance with exceptional structural integrity and alignment between text and 3D content.

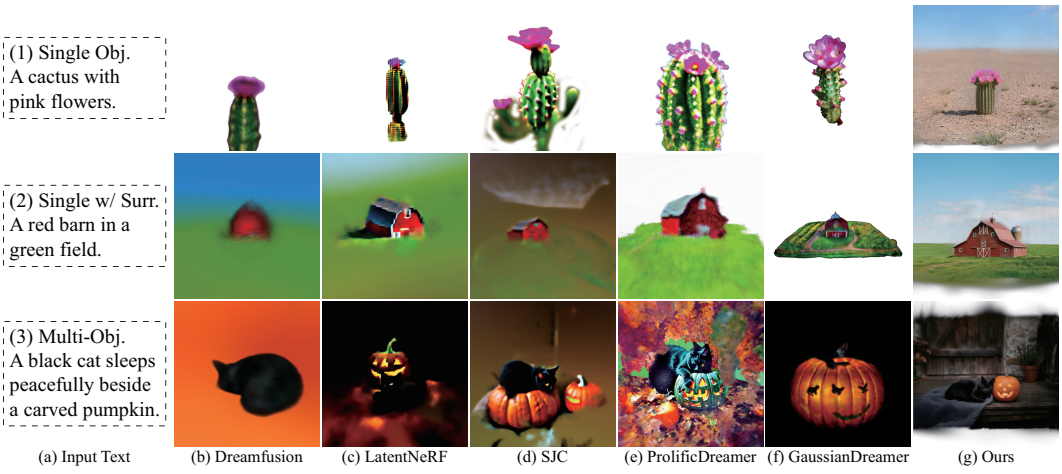


Figure 13: Text-to-3D comparison on T3Bench.

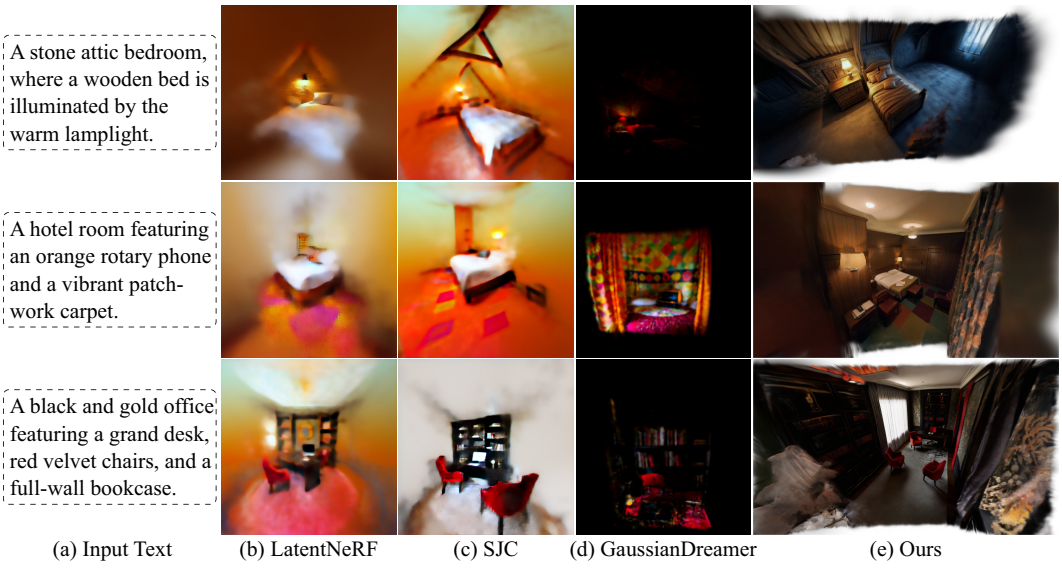


Figure 14: Text-to-3D comparison on CogSG-3D.

C.2 Image-to-3D Object Generation

In Fig. 15, we present the visual results of image-to-3D object generation on our CogSG-3D datasets. Compared to LucidDreamer (Chung et al., 2023), our Cog2Gen3D exhibits a distinct advantage in faithfully translating 2D visual priors into coherent 3D structures. Empowered by the sculptured 3D semantic-geometric cognition, our framework accurately recovers fine-grained geometric details while avoiding the multi-view inconsistencies often observed in the baseline. Con-

sequently, Cog2Gen3D achieves superior image-to-3D generation quality, characterized by remarkable structural integrity and strict alignment with the reference images.

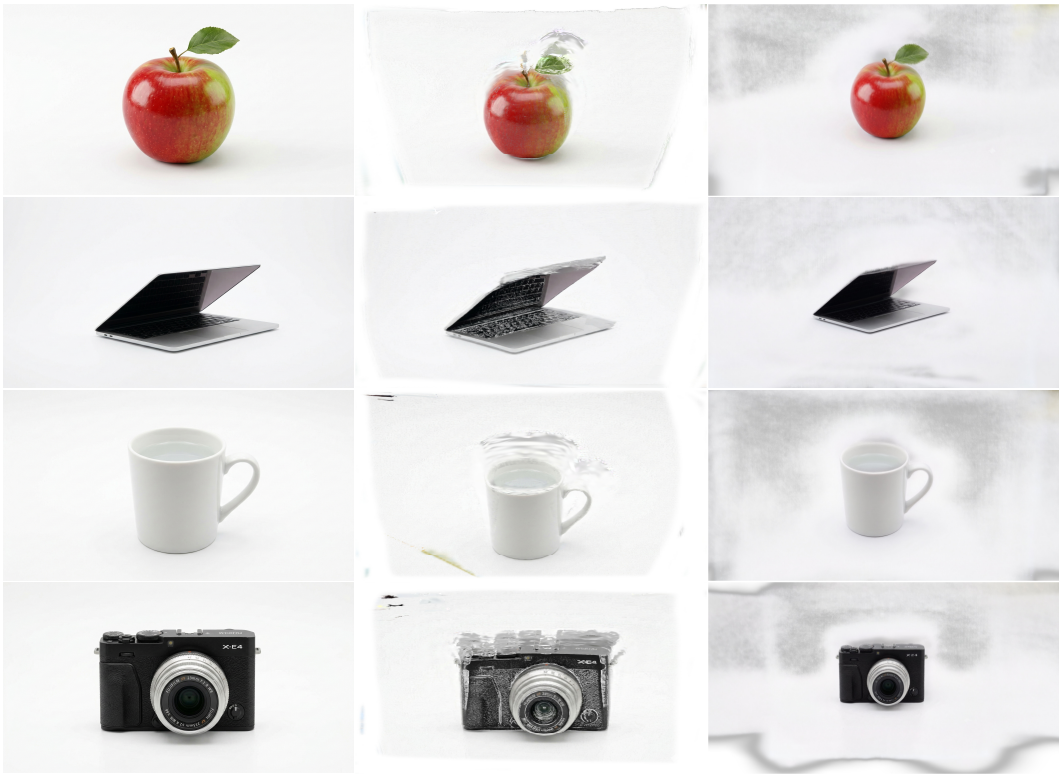


Figure 15: Image-to-3D objects comparison on CogSG-3D.

C.3 Image-to-3D Scene Generation

In Fig. 16, we present the visual results of image-to-3D scene generation on our CogSG-3D datasets. We compare our Cog2Gen3D with state-of-the-art methods, including LucidDreamer (Chung et al., 2023) and Gen3DSR (Ardelean et al., 2025). Compared to these baselines, which often struggle with geometric distortions and structural collapse when expanding complex scene layouts, Cog2Gen3D demonstrates a significant advantage in faithfully translating 2D visual priors into coherent 3D environments. Empowered by the sculptured 3D semantic-geometric cognition, our framework effectively captures both the global spatial arrangements and the intricate local textures of the scene. Consequently, our approach achieves superior image-to-3D scene generation quality, delivering exceptional multi-view consistency, spatial structural fidelity, and strict reference-image alignment.

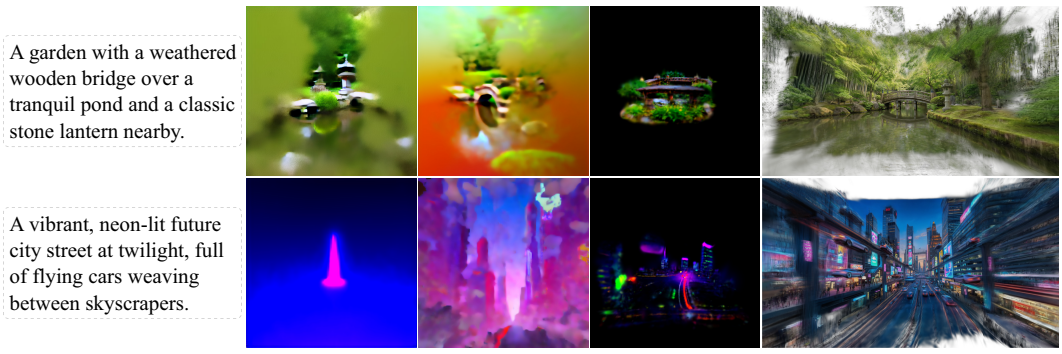
C.4 Generalization Experiments

To evaluate the zero-shot generalization capabilities of our framework, we conduct qualitative experiments using unseen, in-the-wild internet data. As shown in Fig. 17 and Fig. 18, we compare Cog2Gen3D with text-to-3D baselines (Metzer et al., 2023; Wang et al., 2023a; Yi et al., 2024) and the image-to-3D baseline (Chung et al., 2023). For out-of-distribution prompts or complex unconstrained images, the baseline methods frequently suffer from geometric collapse, blurry textures, or distorted topologies. In contrast, by relying on the robust 3D semantic-geometric cognition sculptured during training, Cog2Gen3D demonstrates exceptional adaptability. It synthesizes structurally coherent and geometrically sharp 3D assets strictly aligned with diverse in-the-wild inputs without requiring domain-specific fine-tuning, demonstrating superior generalization in both tasks.



(a) Input Image (b) Gen3DSR (c) LucidDreamer (d) Ours

Figure 16: Image-to-3D scenes comparison on CogSG-3D.



(a) Input Text (b) LatentNeRF (c) SJC (d) GaussianDreamer (e) Ours

Figure 17: Text-to-3D generalization comparison on in-the-wild text prompts.



(a) Input Image (b) LucidDreamer (c) Ours

Figure 18: Image-to-3D generalization comparison on in-the-wild image prompts.