VLM-DREAMER: VLM-IMAGINED BI-DIRECTIONAL INPAINTING FOR SINGLE-IMAGE 360 SCENE GENERATION

Ting-Wei Huang[1] *Fu-En Yang*[2] *Min-Hung Chen*[2] *Yen-Yu Lin*[1] *Yu-Lun Liu*[1]

[1]National Yang Ming Chiao Tung University [2]NVIDIA

ABSTRACT

This paper presents VLM-Dreamer, a novel three-stage pipeline for generating an enriched and consistent 360-degree 3D scene from a single image. VLM-Dreamer addresses the limitations of existing approaches by integrating a Vision-Language Model (VLM) to overcome the constraints of fixed text prompts. By leveraging the VLM's reasoning capabilities, we can imagine plausible surrounding objects that naturally co-occur with the input, allowing us to generate semantically rich and spatially consistent scene descriptions. Specifically, VLM-Dreamer first synthesizes a consistent 360-degree scene image. It then uses an inpainting diffusion model, guided by the VLM-generated descriptions, to produce a coarse 3D Gaussian Splatting representation. At the final stage, it introduces bi-directional inpainting to refine the scene, enhancing cross-view alignment and semantic coherence. Through extensive experiments, we demonstrate that VLM-Dreamer achieves superior performance across multiple metrics, effectively balancing creative imagination with spatial fidelity to produce diverse and high-quality 3D scenes. The source code is available at https://github.com/Lalalalex/VLMDreamer.git.

1. INTRODUCTION

3D scene reconstruction is a core task in computer vision and graphics, typically requiring many multi-view images to recover scene geometry and appearance. In practice, such multiple inputs are often unavailable, making it challenging to generate a complete 3D scene from a single image. Despite applications in content creation, virtual reality, and gaming, this task remains difficult due to limited 3D training data and the inherent ambiguity of sparse input.

Recent works [1, 2, 3] address this by directly synthesizing coherent 3D representations from a single image using powerful generative models and learned priors. Many methods, such as [4, 5, 6], also use a text prompt to guide generation. However, relying on a static prompt for large or 360-degree scenes often causes content redundancy, inaccurate placement, and unnatural arrangements. For example, as shown in Figure 1, Invisible Stitch [5] produces excessive bookshelves, while RealmDreamer [6] embeds a bear within a table. Moreover, such generation lacks spatial reasoning, as objects are often placed without considering contextual coherence or spatial plausibility, resulting in low diversity and unnatural visual arrangements.

Text prompt: *"The reading room features tall bookshelves lined with numerous books, creating a cozy atmosphere."*

Text prompt: *"A bear sitting in a classroom with a hat on, realistic, 4k image, high detail."*

Fig. 1. **Using a static text prompt for 3D scene generation often suffers from content redundancy and inaccurate spatial arrangement.** (**Top**) Invisible Stitch [5] produces excessive bookshelves. (**Bottom**) RealmDreamer [6] embeds a bear within a table.

To address these challenges, we propose a three-stage 3D scene generation pipeline that balances creative reasoning with spatial fidelity. Our method leverages a vision-language model (VLM) to imagine contextually appropriate surrounding objects from a single input image, guiding the generation of a coherent 360-degree scene. Bi-directional refinement ensures semantic consistency and alignment across views. Our contributions are threefold: We integrate VLM-based generative reasoning into 3D scene synthesis, enriching and diversifying the generated content from a single image. We design a three-stage pipeline that balances imagination with spatial fidelity, producing diverse, consistent, and high-quality results. Extensive evaluations on the WorldScore dataset [8] demonstrate superior and balanced performance across multiple metrics.

2. RELATED WORK

2.1. 3D Representation

An effective 3D representation is crucial for balancing quality, efficiency, and flexibility in scene generation. 3D Gaussian Splatting (3DGS) [7], a state-of-the-art method using anisotropic Gaussian primitives, provides high-fidelity, real-time rendering. We adopt 3DGS as the 3D representation in

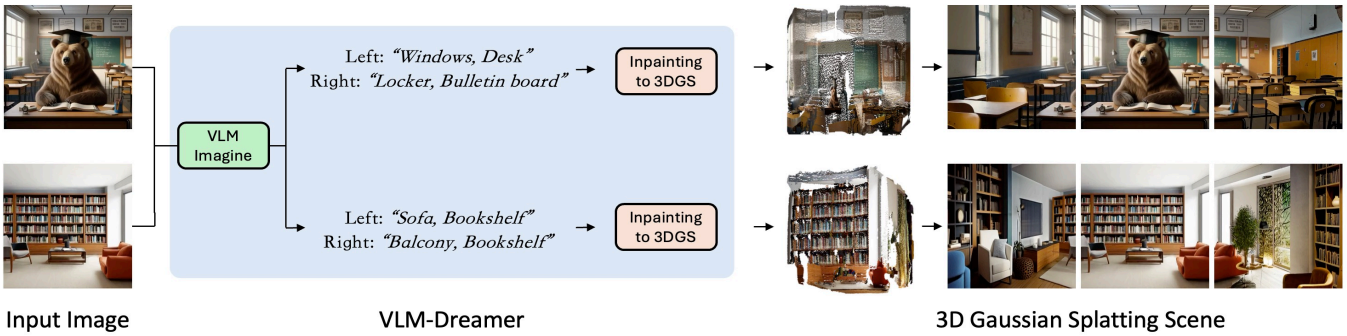

Fig. 2. **Given an input image, our method generates a 360 Scene using 3D Gaussian Splatting (3DGS) [7].** Starting from the input, we leverage a vision-language model (VLM) to imagine plausible surrounding objects, guiding the inpainting process to reconstruct the full scene. Our approach results in novel view synthesis exhibiting consistency, diversity, and high quality.

this work.

2.2. Scene Generation

3D scene generation aims to synthesize spatially coherent and visually realistic environments from sparse inputs such as images or text, with applications in gaming, virtual reality, and simulation. Current research typically falls into three paradigms:

Panorama-to-3D Scene Generation. Methods like [3, 9, 10] generate a 360-degree panorama and lift it into a 3D representation, enforcing multi-view consistency. However, they often suffer from low resolution, blurry geometry, and limited content diversity, especially in peripheral regions.

Inpainting-to-3D Scene Generation. These approaches [4, 5, 2, 11, 6] iteratively synthesize multi-view images from an initial view using inpainting guided by a text prompt. While they achieve high single-view quality, they often lack global geometric constraints, resulting in limited 3D consistency and cross-view misalignment.

Video-to-3D Scene Generation. Methods [1, 12, 13] reconstruct 3D scenes from video sequences, leveraging multi-view and temporal information for improved coherence. However, they require extensive data and computation, and maintaining global consistency in long or 360-degree videos remains challenging.

Our method combines these paradigms by using geometrically consistent representations as a foundation for semantic reasoning and powerful inpainting for high-quality synthesis. By employing a vision-language model (VLM) to imagine plausible surrounding objects, it enriches scenes with contextually relevant details, enabling not only coherent and high-quality but also diverse 3D scene generation.

2.3. MultiDiffusion

MultiDiffusion [14] extends diffusion models to support multiple spatially localized prompts, enabling better spatial control and regional consistency by fusing intermediate predictions across conditions. Building on this, we propose a bi-directional inpainting strategy that leverages multi-prompt conditioning with spatial weighting to improve semantic alignment and global consistency in 3D scene refinement.

3. PROPOSED METHOD

The proposed VLM-Dreamer is a three-stage pipeline for generating a 360-degree scene from a single image. As illustrated in Figure 3, Stage 0 uses a pre-trained panorama or video diffusion model to synthesize multi-view images as semantic initialization, which a vision-language model (VLM) analyzes to imagine surrounding content and produce 360-aware scene descriptions. Stage 1 performs uni-directional inpainting from the input view guided by the VLM prompts, generating a coarse scene with high semantic diversity. Stage 2 refines the scene via bi-directional inpainting, where each view is influenced by its neighbors and their prompts, combined through a linear interpolation weighted sum to enhance local consistency. The final 360-degree scene is output in 3D Gaussian Splatting (3DGS) [7] format.

3.1. 360-Aware Scene Description Generation

This stage aims to generate a semantically rich and spatially consistent 360-degree scene description. The description then serves as the textual condition for inpainting in the subsequent stages. As shown in Figure 3a, given an input image, we apply a pre-trained panorama or video diffusion model to synthesize a set of multi-view images, which broadly cover the semantic context of a 360 scene. In this work, we use the panorama model [3] for its stable and good performance. Although these diffusion models ensure good spatial consistency across views, their outputs often suffer from low resolution, blurry textures, or structural artifacts due to inherent generative limitations. As such, we do not use these images directly for 3D reconstruction but treat them as semantic observations for high-level scene knowledge extraction.

To obtain semantic context, we leverage the VLM [15] to "imagine" what lies beyond each view. To this end, for each multi-view image P_{i-1}, we query the VLM with a specific

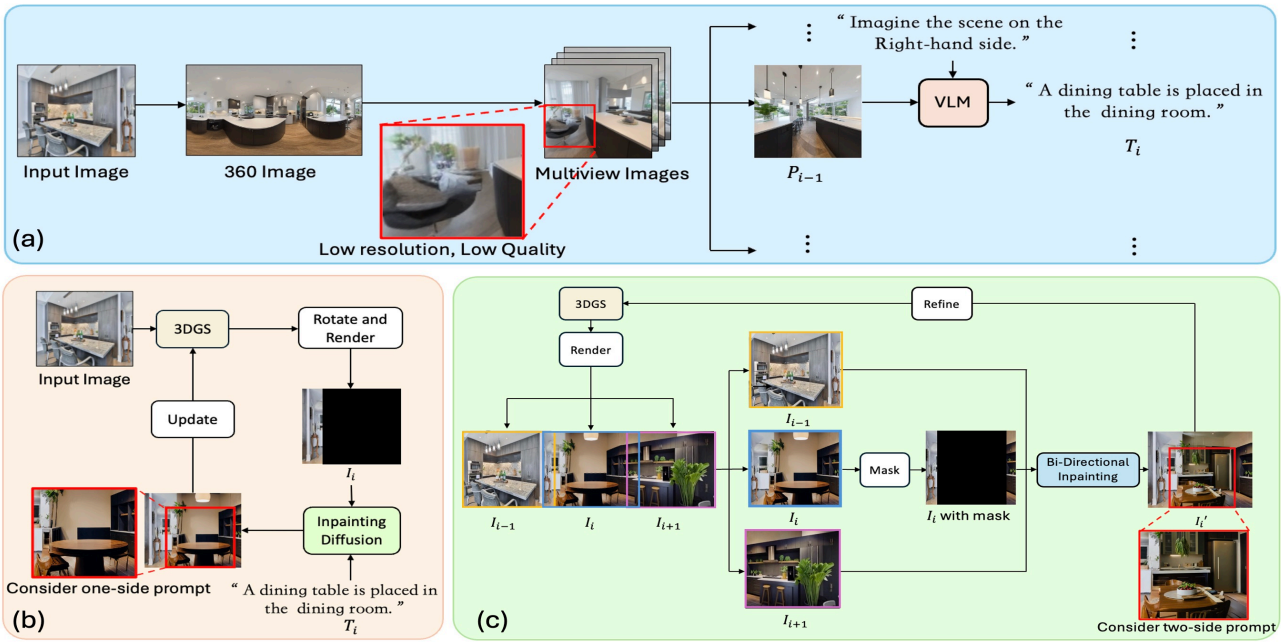

Fig. 3. **Our three-stage pipeline.** (a) Stage 0: 360-Aware Scene Description Generation. It uses a VLM to obtain 360-aware, enriched, and semantically consistent textual prompts to describe the full scene. (b) Stage 1: Uni-Directional Coarse Scene Generation. These textual prompts are utilized to guide initial, one-way inpainting to produce a realistic and spatially consistent 3DGS representation of the scene. (c) Stage 2: Bi-Directional 3DGS Refinement. It refines the scene's alignment and consistency through a two-way inpainting process.

prompt (e.g., "what might appear to the right of this view?"). This process provides the adjacent view P_i with a directional textual description T_i.

The design of grounding each imagination step on globally aware multi-view inputs is crucial, as it ensures that the resulting descriptions remain contextually coherent with the entire scene. If we allow the VLM to imagine the scene sequentially, conditioning each new description solely on the previous view, the content would likely drift semantically over time and gradually deviate from the original scene.

By leveraging the imagination capabilities of the VLM, this stage enhances the diversity of the scene's content. The resulting 360-degree scene description $\{T_i\}$ is both semantically rich and spatially aligned, and it serves as the foundation for the next stage of coarse scene generation.

3.2. Uni-Directional Coarse Scene Generation

The goal of this stage is to synthesize a coarse 360 scene by leveraging the previously generated 360-aware scene descriptions as textual conditions for an inpainting model.

Using an inpainting model in this stage offers a key advantage: Most inpainting models are trained to complete missing regions with high visual fidelity, guided by both the input image's context and text prompts. While they can produce realistic and detailed outputs, they inherently lack 3D consistency across views. We solve this by conditioning the inpainting process on our 360-aware scene descriptions. These descriptions provide semantically diverse and spatially coherent cues, which enable the inpainting model to produce images that are not only realistic and detailed but also consistent across viewpoints.

As shown in Figure 3b, this stage begins by creating an initial 3DGS scene from the input image. We then iteratively rotate the current viewpoint by a small angle and render a novel view I_i, which exposes previously unseen areas. For each rendered view I_i, we retrieve the corresponding direction scene description T_i from the previous stage, and jointly input (I_i, T_i) into the pre-trained inpainting diffusion model [16]. The model synthesizes the result that aligns with the semantic context described in T_i, and we use this output to further train and update the 3DGS. This process is repeated iteratively until the entire 360 field of view is covered, resulting in an initial coarse 3DGS scene.

3.3. Bi-Directional 3DGS Refinement

The primary goal of this stage is to further improve the alignment of the coarse 3DGS scene generated in the previous stage. Although the initial 3DGS synthesis covers a complete 360 view, each inpainting operation is conditioned only on a uni-directional prompt. This approach inherently limits the scene's global consistency. To address this limitation, we propose a Bi-Directional Inpainting strategy inspired by MultiDiffusion [14], which enables conditioning on multiple textual prompts simultaneously.

Bi-Directional Inpainting. MultiDiffusion guides the generation process by using multiple, spatially localized condi-

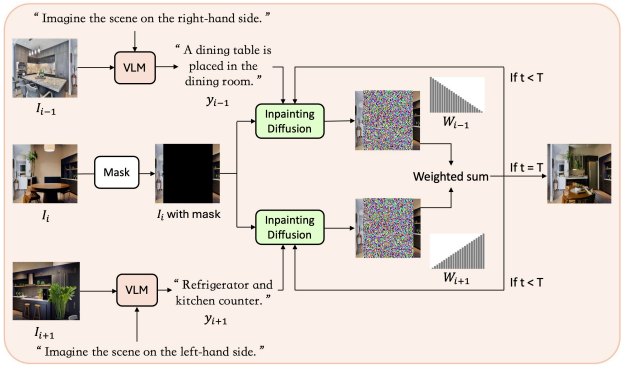

Fig. 4. **Proposed bi-directional inpainting.** We independently denoise based on conditions from two directions and fuse the results via a linear interpolation weighted sum, producing an output that incorporates semantic information from both directions.

tions. It works by fusing the denoising trajectories from a shared pre-trained diffusion model. During each denoising step t, the model computes multiple intermediate predictions, each guided by a different condition. For areas where these predictions overlap, a spatially-aware weighted average is used to blend them into a single consistent output. Specifically, the fused noise prediction ϵ_t at step t is computed as:

$$\epsilon_t = \sum_{i=1}^{n} \frac{F_i^{-1}(W_i)}{\sum_{j=1}^{n} F_j^{-1}(W_j)} \otimes F_i^{-1}(\epsilon_\theta(z_t \mid y_i)), \quad (1)$$

where z_t denotes the latent at denoising step t, z is the initial noise latent, and n is the total number of spatial regions (or conditions). $\epsilon_\theta(z_t \mid y_i)$ denotes the prediction from region i conditioned on prompt y_i, W_i is the spatial blending weight, and F_i^{-1} is the inverse mapping that transforms local predictions back to the global coordinate system. This formulation lets the model incorporate multiple conditioning signals in a spatially coherent way, preserving the integrity of the diffusion process while enabling region-specific guidance.

Building upon MultiDiffusion, we specialize its formulation for our bi-directional inpainting by using directional textual prompts from neighboring views, as illustrated in Figure 4. For each inpainting region, we generate two intermediate predictions: One is conditioned on the left-direction prompt y_{i-1}, imagined by the VLM from the perspective of the left neighbor, while the other is on the right-direction prompt y_{i+1}, imagined from the right neighbor's perspective. During the denoising process at each timestep, these two prompt-conditioned predictions, $\epsilon_\theta(z_t \mid y_{i-1})$ and $\epsilon_\theta(z_t \mid y_{i+1})$, are spatially blended using linear interpolation weights W_{i-1} and W_{i+1}, which vary smoothly across the horizontal axis. As illustrated in Figure 4, the weights are defined by a linear interpolation parameter $\alpha(x) \in [0, 1]$ that varies along the horizontal axis x, such that:

$$W_{i-1}(x) = 1 - \alpha(x) \text{ and } W_{i+1}(x) = \alpha(x). \quad (2)$$

The final fused noise prediction at each step is computed as a linear interpolation weighted sum of the two predictions:

$$\epsilon_t = \epsilon_\theta(z_t \mid y_{i-1}) \otimes W_{i-1} + \epsilon_\theta(z_t \mid y_{i+1}) \otimes W_{i+1}. \quad (3)$$

This interpolation is performed at the noise level rather than the pixel level.

Bi-Directional 3DGS Refinement Pipeline. Based on our bi-directional inpainting strategy shown in Figure 3c, for each target view i, we render three images from the coarse scene: the current view I_i and its adjacent views I_{i-1} and I_{i+1}. Before inpainting, a central region of I_i is masked to create a hole for refinement. As detailed in Figure 4, we then use I_{i-1} and I_{i+1} to generate text prompts y_{i-1} and y_{i+1}, imagined by a VLM based on the neighboring views. These prompts are processed using the directional inpainting method to generate a refined image I_i'. This result better captures the intermediate view semantics by jointly considering contextual information from both the left and right directions, and is used to update the 3DGS scene, enhancing view-to-view consistency and improving global scene coherence.

3.4. Implementation Details

All experiments are conducted on a single RTX 3090 GPU (24GB). We use Fooocus [16] for high-fidelity inpainting and LLaVA [15] to generate 360-aware scene descriptions from visual input. The final 3D scene is reconstructed via 3D Gaussian Splatting (3DGS) [7]. The entire pipeline is implemented in PyTorch as a unified and efficient framework.

4. EXPERIMENTS

4.1. Experimental Settings

Dataset. We conduct experiments on a subset of the WorldScore dataset [8], a benchmark for 3D scene generation that provides diverse real-world scenes, evaluated under a consistent 360 trajectory.

Inspired by previous work [6], we randomly select a subset of 20 scenes from this dataset for our evaluation. Each scene is comprised of a single image and a corresponding scene-level text description, which allows us to evaluate both image- and text-driven scene generation methods.

Competing Methods. We compare our method with five representative baselines [4, 5, 3, 2, 1]. LucidDreamer, Invisible Stitch, and VistaDream iteratively expand a 3DGS scene using inpainting-based diffusion. DreamScene360 generates a panorama from a text prompt for 3DGS training, while FlexWorld synthesizes a video from an image to train a 3DGS model.

Evaluation Metrics. We evaluate 360-degree scene generation on three aspects: *Quality*, *Diversity*, and *Alignment*. For quality, we measure perceptual fidelity using no-reference metrics: CLIP-IQA [17], CLIP-Aesthetic [18], BRISQUE [19], and NIQE [20]. For diversity, we use Inception Score [21], Intra-LPIPS [22], and the Obj Score,

Table 1. We evaluate our method and five representative baselines across eight criteria within three key aspects: image quality, diversity, and alignment. Best, second best, and third best results are highlighted.

Method	Image Quality				Diversity			Alignment
	Aes ↑	CLIPIQA↑	BRISQUE↓	NIQE ↓	IS ↑	Obj↑	IntraLPIPS↑	FrameCLIP↑
LucidDreamer [4]	5.049	0.710	33.746	4.981	1.760	11.050	0.598	0.922
Invisible Stitch [5]	4.338	0.565	36.444	4.181	1.437	7.850	0.704	0.870
DreamScene360 [3]	4.315	0.318	41.597	5.660	1.931	6.800	0.619	0.914
Vistadream [2]	4.841	0.649	51.101	6.639	1.949	10.650	0.655	0.883
FlexWorld [1]	4.905	0.421	25.150	4.174	2.410	9.550	0.620	0.875
VLM-Dreamer (Ours)	5.070	0.725	24.637	3.859	2.601	12.750	0.726	0.914

Table 2. Ablation studies. LM" indicates the use of the VLM to imagine surrounding objects, "SD" provides a 360-aware scene description for inpainting, "BD" is bi-directional inpainting, and "LI" applies linear interpolation weighting. The final model combines these components for consistent 360-degree scene generation.

Version	Components				Image Quality				Diversity			Alignment
	VLM	SD	BD	LI	Aes ↑	CLIPIQA↑	BRISQUE↓	NIQE ↓	IS ↑	Obj↑	IntraLPIPS↑	FrameCLIP↑
Ver 1					5.012 ± 0.099	0.718 ± 0.094	25.462 ± 5.310	4.032 ± 0.563	2.062 ± 0.302	11.250 ± 1.226	0.699 ± 0.049	**0.919 ± 0.013**
Ver 2	✓				5.020 ± 0.095	0.711 ± 0.112	26.412 ± 5.971	3.991 ± 0.468	**2.671 ± 0.417**	**13.450 ± 2.301**	**0.731 ± 0.052**	0.896 ± 0.031
Ver 3	✓	✓			5.037 ± 0.093	0.718 ± 0.083	25.018 ± 4.605	3.886 ± 0.507	2.448 ± 0.369	12.300 ± 1.526	0.711 ± 0.059	0.905 ± 0.018
Ver 4	✓	✓	✓		4.983 ± 0.109	0.714 ± 0.105	24.872 ± 5.268	3.987 ± 0.547	2.472 ± 0.382	12.550 ± 2.184	0.720 ± 0.039	0.907 ± 0.015
Ver 5	✓	✓	✓	✓	**5.070 ± 0.098**	**0.725 ± 0.084**	**24.637 ± 5.412**	**3.859 ± 0.462**	2.601 ± 0.396	12.750 ± 2.018	0.725 ± 0.045	0.914 ± 0.017

computed via a pretrained segmentation model [23] to count distinct object categories. For alignment, we evaluate local semantic consistency across adjacent frames using FrameCLIP, the CLIP similarity between consecutive views.

4.2. Comparisons with Baselines

Table 1 compares VLM-Dreamer with five baselines. LucidDreamer [4] and Invisible Stitch [5] rely on fixed text prompts, achieving high alignment but low diversity due to repeated structures. VistaDream [2] uses a VLM to generate prompts for the current view; however, it only describes the immediate view without further imagination, and overall scene diversity remains limited. DreamScene360 [3] generates panoramas for 3D scenes, achieving strong alignment but sacrificing detail and object variety. FlexWorld [1] produces coherent sequences from videos, but quality degrades over long 360 trajectories. In contrast, VLM-Dreamer consistently balances quality, alignment, and diversity. By leveraging 360-degree semantic features from panorama and video diffusion and the imagination capability of a VLM, it ensures global consistency, high-quality local generation, and diverse yet coherent 3D scenes.

4.3. Ablation Studies

For our ablation studies, we evaluate VLM-Dreamer and its degenerated variants, and report the average and standard deviation of all metrics in Table 2. Ver1 uses a fixed prompt without a VLM, achieving high alignment but low diversity and repetitive content. Ver2 introduces the VLM but omits a global scene description, generating each view based only on the preceding frame; diversity improves, but semantic drift reduces alignment. Ver3 adds a 360-aware scene description to guide uni-directional inpainting, improving both diversity and alignment. Ver4 applies bi-directional inpainting, and Ver5 further combines it with linear interpolation weighting, refining spatial consistency and enhancing alignment across the scene.

4.4. Limitations

Our method uses Vision-Language Models (VLMs) to infer and imagine surrounding content from an input image. While VLMs provide rich semantic guidance, their outputs are biased by the training data, and they may struggle with rare or ambiguous objects.

5. CONCLUSION

In this work, we present VLM-Dreamer, a novel three-stage pipeline for 360-degree 3D scene generation from a single input image. By integrating a vision-language model (VLM), our method effectively leverages semantic reasoning to imagine plausible surrounding objects and greatly enriches scene content beyond the visible input. Experiments on the WorldScore dataset show that VLM-Dreamer achieves superior and balanced performance across multiple metrics, underscoring the potential of integrating VLM understanding into generative models. This work suggests a new direction for future research that moves beyond what's visible to what's imaginable.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under Grants NSTC 112-2222-E-A49-004-MY2, 113-2628-E-A49-023-, 114-2221-E-A49-038-MY3, and 112-2221-E-A49-090-MY3, and by the NVIDIA Taiwan AI Research & Development Center (TRDC). Yu-Lun Liu acknowledges support from the Yushan Young Fellow Program by the Ministry of Education (MOE), Taiwan.

7. REFERENCES

[1] Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li, "FlexWorld: Progressively expanding 3d scenes for flexiable-view synthesis," *arXiv preprint arXiv:2503.13265*, 2025.

[2] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang, "VistaDream: Sampling multiview consistent images for single-view scene reconstruction," in *ICCV*, 2025.

[3] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi, "Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting," in *ECCV*, 2024.

[4] 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.

[5] Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht, "Invisible stitch: Generating smooth 3d scenes with depth inpainting," *arXiv preprint arXiv:2404.19758*, 2024.

[6] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi, "RealmDreamer: Text-driven 3d scene generation with inpainting and depth diffusion," *International Conference on 3D Vision (3DV)*, 2025.

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3D gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, 2023.

[8] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu, "WorldScore: A unified evaluation benchmark for world generation," *arXiv preprint arXiv:2504.00983*, 2025.

[9] Yukun Huang, Yanning Zhou, Jianan Wang, Kaiyi Huang, and Xihui Liu, "DreamCube: 3d panorama generation via multi-plane synchronization," *arXiv preprint arXiv:2506.17206*, 2025.

[10] Yikun Ma, Dandan Zhan, and Zhi Jin, "FastScene: Text-driven fast 3d indoor scene generation via panoramic gaussian splatting," in *IJCAI*, 2024.

[11] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu, "WonderWorld: Interactive 3d scene generation from a single image," *arXiv:2406.09394*, 2024.

[12] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian, "ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis," *arXiv preprint arXiv:2409.02048*, 2024.

[13] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang, "Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion," in *ICCV*, 2025.

[14] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel, "MultiDiffusion: Fusing diffusion paths for controlled image generation," *arXiv preprint arXiv:2302.08113*, 2023.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[16] Lvming Zhang, "Fooocus," https://github.com/lllyasviel/Fooocus, 2023.

[17] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy, "Exploring CLIP for assessing the look and feel of images," in *AAAI*, 2023.

[18] Christoph Schuhmann, "CLIP+MLP aesthetic score predictor," https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

[19] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE TIP*, 2012.

[20] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Sign. Process. Letters*, 2013.

[21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," in *NeurIPS*, 2016.

[22] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang, "Few-shot image generation via cross-domain correspondence," in *CVPR*, 2021.

[23] Mupparaju Sohan, Thotakura Sai Ram, and Ch Venkata Rami Reddy, "A review on yolov8 and its advancements," in *ICDICI*, 2024.