This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Camera Control for Text-to-Image Generation via Learning Viewpoint Tokens

Xinxuan Lu Charless Fowlkes Alexander C. Berg
University of California, Irvine
{xinxul1, fowlkes, bergac}@uci.edu

Abstract

Current text-to-image models struggle to provide precise camera control using natural language alone. In this work, we present a framework for precise camera control with global scene understanding in text-to-image generation by learning parametric camera tokens. We fine-tune image generation models for viewpoint-conditioned text-to-image generation on a curated dataset that combines 3D-rendered images for geometric supervision and photorealistic augmentations for appearance and background diversity. Qualitative and quantitative experiments demonstrate that our method achieves state-of-the-art accuracy while preserving image quality and prompt fidelity. Unlike prior methods that overfit to object-specific appearance correlations, our viewpoint tokens learn factorized geometric representations that transfer to unseen object categories. Our work shows that text-vision latent spaces can be endowed with explicit 3D camera structure, offering a pathway toward geometrically-aware prompts for text-to-image generation. Project page: https://randdl.github.io/viewtoken_control/.

1. Introduction

Controllable image generation with precise camera viewpoint specification is an increasingly important capability for modern generative models. While many text-to-image models [9, 19, 34] have demonstrated remarkable progress in semantic fidelity and visual realism, they struggle to follow even simple geometric instructions such as "back view", "30° left-side view", or "45° top-down perspective." Natural language is expressive but inherently ambiguous and discrete for viewpoint specification, and current models often hallucinate incorrect poses, collapse to biased canonical angles, or produce inconsistent geometry across trials. To overcome these limitations, we introduce a method that augments text prompts with explicit, fine-grained camera control, enabling precise specification of viewpoint (Fig. 1).

Prior attempts at viewpoint control remain limited. 3D-aware generative models [20, 35, 39] and Novel View Synthesis [17, 25, 31, 52] require additional inputs beyond text prompts as summarized in Tab. 1. View-NeTI [3] learns object and viewpoint tokens by training on multi-view images for each object. Compass Control [27] makes progress by learning viewpoint tokens with text prompts but only supports azimuth control. Its attention-masking strategy confines viewpoint cross-attention to a local region, which can inhibit global scene understanding and may lead to overfitting to specific training objects. Overall, learning camera information within text prompts remains underexplored.

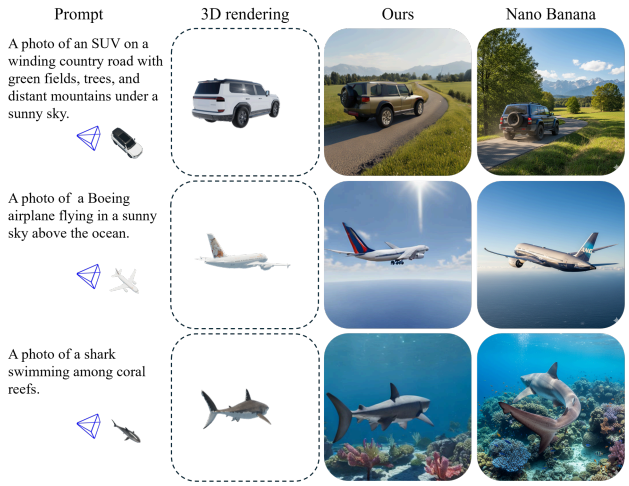


Figure 1. **Our model vs. Gemini 2.5 Flash Image (Nano Banana)** [9]. Our encoded camera viewpoint tokens enable precise camera pose control, while Nano Banana often fails despite detailed descriptions: "angled diagonally to show its rear and left side from a rear three-quarter view (approx. 220° azimuth), with the camera slightly below eye level (10° elevation). It occupies approximately 60% of the image width, positioned in the center slightly to the left." See Supp. B for more results. Objects shown in prompts and 3D rendering column are only shown to illustrate the desired viewpoints.

Approach. Our approach adds a camera viewpoint specification to a text prompt and learns how to integrate this information for image generation. The camera is param-



(a) 45° to the left (b) 45° to the right (c) 30° to the right

Figure 2. **GPT5 viewpoint failures.** Generated by GPT5 [26] using "A white sedan seen from 45°/30° to the left/right of the front view". All three prompts result in nearly identical orientations.

eterized relative to the object and its front in order to help provide a consistent notion of viewpoint, for example, "left" and "right".

These camera parameters are encoded into learnable viewpoint embeddings and concatenated with the text embeddings as the input to a text-to-image generation backbone [7, 34, 42] and jointly trained or fine-tuned. This allows the resulting trained model to condition on semantic content and explicit camera viewpoint specifications during text-to-image generation.

The choice of data used in this joint training is important to avoid overfitting or collapse. To achieve this we construct a dataset with two parts. One part—the large rendered dataset—uses rendered 3D models of objects. Using this alone for training can cause the text-to-image models to collapse, "forgetting" how to render more complex scenes and follow detailed text prompts. To avoid this, we add a second part—photorealistic augmented images—consisting of full scenes containing an object in a known pose. These are generated by prompting a commercial image generation system with a rendered object in a known pose as well as a text description of the object and the scene. We use the same two-part dataset throughout our experiments.

We demonstrate the effectiveness of our approach by fine-tuning multiple text-to-image generation models, while simultaneously learning a lightweight encoder for each that maps camera parameters to token embeddings (Fig. 3).

Through quantitative and qualitative experiments, we show that our method achieves state-of-the-art viewpoint accuracy while maintaining high image fidelity and robust generalization to unseen objects. Compared to prior work—such as View-NeTI [3], which learns object-specific tokens, and Compass Control [27], which can overfit to training appearance—our method learns viewpoint token embeddings that are more independent of object identity or shape. Finally, our canonical camera-object framework and two-part dataset design also enable a global understanding of scene geometry, contributing to more consistent foreground–background relationships and more reliable viewpoint control.

Table 1. **Comparison with previous work** in terms of input type and camera control.

Model	Input Type
Wonder3D [20]	Image + Camera
Zero-123 [17]	Image + Camera
OneDiffusion [13]	Image + Camera
Stable-Virtual-Camera [52]	Image + Camera
Controlnet-Depth [50]	Image + Depth
View-NeTI [3]	Object Token + Camera Token
Compass Control [27]	Text + Azimuth Token
Ours	Text + Camera Token

Our contributions are:

- State-of-the-art camera control in both range and accuracy while preserving image quality and fidelity to text prompts, outperforming both novel-view generation and prior token-based methods.
- Context-preserving training that enables the model to learn viewpoint cues with a global understanding of the scene rather than isolated object cutouts.
- Two-part dataset design: a high volume of canonically aligned renderings provide strong geometric supervision, while a low volume of photorealistic augmentations maintain realism and appearance diversity in generations.
- Experiments showing that this approach is robust, using the same training data and approach for different text-to-image generation models, and that it has improved generalization to unseen object categories with less overfitting.

2. Related Work

Text-to-Image (T2I) Generation. Large-scale diffusion-based text-to-image models [28, 29, 33, 34, 37] have achieved unprecedented progress in realism and semantic alignment, while MAR [14] provided a strong alternative based on autoregressive image encoding and generation. More recently, extensions toward unified multimodal models [6, 41–43] further integrate image understanding and generation by learning a unified vision-language space in both input and output. Despite many advances, such models still struggle to provide precise spatial or geometric control, as natural language offers only implicit viewpoint descriptions, and training data are heavily biased toward front-facing views or common compositions. To overcome these limitations, our approach embeds viewpoint information into the text prompt for better geometry understanding.

Controllable Image Generation. To enhance control, many works augment text prompts with auxiliary structural inputs such as depth, edges, or segmentation masks, such as ControlNet [50] and T2I-Adapter [24]. Recent works also explore 2D and 3D layout-guided generation [15, 22, 48]. However, methods that condition on structural cues inher-

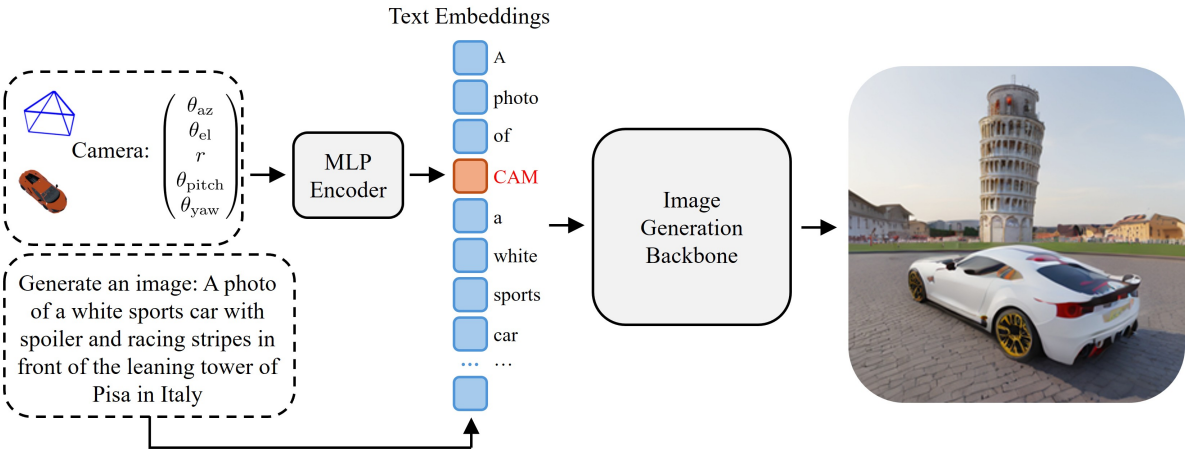


Figure 3. **Architecture overview.** An MLP encoder maps camera parameters to a token embedding that is processed jointly with text tokens for viewpoint-conditioned image generation. We fine-tune the image generation model [7, 34, 42] jointly with the camera token encoder. The rendered red car in the prompt is only shown to illustrate the desired viewpoint.

ently require explicit 3D reference inputs at inference (e.g., depth or edges) thereby limiting flexibility and reducing usability in real-world settings. In contrast, our method achieves fine-grained spatial control from parameterized camera tokens appended to text inputs without relying on additional geometric information.

3D Generative Models. Traditional 3D-aware generative models [4, 46] integrate explicit geometry representations such as NeRF [23] to synthesize viewpoint-consistent images. Subsequent text-to-3D methods [16, 30] leverage score distillation sampling to distill pretrained 2D diffusion models into 3D representations. More recently, many works [20, 35, 36, 38, 39] directly regress 3D structures from a single or sparse set of images. However, these methods lack a consistent canonical understanding of an object's front-facing orientation, and the resulting 3D models may lack the nuance and realistic texture seen in 2D image generation. In contrast, our work generates viewpoint-conditioned images that preserve high-fidelity appearance and provide contextual consistency between object and the scene background.

Novel View Synthesis. Classic NVS methods [12, 23] reconstruct a scene from multiple calibrated images and enable rendering from unseen viewpoints. More recent works [5, 13, 17, 18, 25, 31, 40, 49, 52] leverage 2D diffusion models to generate multi-view images or 3D structure from one or a few input views. However, these methods require one or more input images and thus cannot provide direct camera control for *ab initio* text-to-image generation. Our approach bridges this gap by embedding explicit viewpoint tokens into the text prompt, providing a one-step viewpoint-conditioned text-to-image generation model without additional reference images.

Viewpoint-Conditioned Generation. Recent work has begun to explore viewpoint control within text-to-image generation models. PreciseCam [2] targets scene-level camera control instead of objects. Diffusion-as-Shader [10] addresses relative camera control in video generation rather than T2I. View-NeTI [3] learns disentangled object and viewpoint tokens but requires object-specific multi-view supervision and struggles to produce geometrically consistent novel views without such data. Compass Control [27] introduces compass tokens that condition generation on azimuth rotations without needing multi-view inputs. However, Compass Control's controllability is limited to a single rotation axis and it generalizes poorly to unseen objects. Its attention localization strategy also prevents it from learning a global understanding of the scenes for different viewpoints. Our work extends this line by enabling flexible and accurate camera control over multiple camera parameters and stronger generalization to new objects and prompts.

3. Method

Our method is designed to work with any text-to-image model that operates on text embeddings as inputs. As illustrated in Fig. 3, given a text prompt and explicit camera parameters $\boldsymbol{\theta}$, we generate a parametric viewpoint embedding token in the same input space as the text tokens. The combined text and viewpoint token embeddings are processed jointly through the model to generate viewpoint-conditioned images. We discuss our object-camera system in Sec. 3.1, how we encode the viewpoint into token embeddings in Sec. 3.2, and our dataset setup in Sec. 3.3.

3.1. Viewpoint Parameterization

We adopt an object-centric system where the object is fixed at the origin, and the front of the object always faces along

the positive x-axis of the world coordinates. This gives us consistent "left/right" and "front/back" in natural language across all objects. The camera is allowed to move freely to capture different views of the object.

We parameterize the camera viewpoint using a factorized 5-parameter representation:

$$\boldsymbol{\theta} = (\theta_{\text{az}}, \theta_{\text{el}}, r, \theta_{\text{pitch}}, \theta_{\text{yaw}}) \in \mathbb{R}^5 \tag{1}$$

where $(\theta_{\text{az}}, \theta_{\text{el}}, r)$ defines the position of the camera with a spherical coordinate. The radius r is specified in units of the object diameter. $(\theta_{\text{pitch}}, \theta_{\text{yaw}})$ defines the relative camera rotation with respect to the direction from the camera position to the origin. Positive θ_{pitch} represents camera tilting down, and positive θ_{yaw} represents camera tilting left. We assume the camera and object "up" directions are aligned (i.e., $\theta_{\text{roll}} = 0$) and a fixed focal length (FoV of $\sim 55°$).

3.2. Viewpoint Token Encoding

We use a parametric viewpoint token that encodes camera view using a lightweight MLP. Given the 5-parameter viewpoint representation $\boldsymbol{\theta} = (\theta_{\text{az}}, \theta_{\text{el}}, r, \theta_{\text{pitch}}, \theta_{\text{yaw}})$, we first apply a parameter encoding function:

$$\phi(\boldsymbol{\theta}) = [\sin(\theta_{\text{az}}), \cos(\theta_{\text{az}}), \theta_{\text{el}}, r, \theta_{\text{pitch}}, \theta_{\text{yaw}}] \in \mathbb{R}^6, \tag{2}$$

where azimuth is encoded via sine and cosine to handle periodicity, radius is normalized to $[0, 1]$, and elevation, pitch, and yaw are directly used as radian values. We then map these encoded parameters to a token embedding via a 3-layer MLP with ReLU activations:

$$\mathbf{e}_{\text{view}} = \text{MLP}_{\text{view}}(\phi(\boldsymbol{\theta})) \in \mathbb{R}^d \tag{3}$$

We insert the viewpoint token adjacent to the object description, allowing precise geometric information to flow through the model's attention mechanism alongside text.

3.3. Dataset Setup

For the large rendered dataset, we manually select 3,111 objects across four categories (animals, vehicles, people, and furniture) from TexVerse [51], a large-scale 3D asset collection. We align each object to a canonical front-facing orientation so that $\theta_{\text{az}} = 0, \theta_{\text{el}} = 0$ correspond to the object's front view. To ensure diverse yet natural perspectives, we sample cameras randomly: $r \in [\frac{4}{3}, 2]$ object size, $\theta_{\text{az}} \in [0, 2\pi)$, $\theta_{\text{el}} \in [0, \pi/4]$, and $\theta_{\text{pitch}}, \theta_{\text{yaw}} \in [-\pi/12, \pi/12]$. We render 120 viewpoints per object, yielding around 373K images with transparent background.

To create a second dataset with photorealistic augmented images, we select 800 high-quality objects from the large dataset and render each at 20 random viewpoints for background augmentation. We use Nano Banana [9] to edit the rendered images to include diverse backgrounds and object appearances while following the original rendered pose. We sample a diverse set of detailed descriptions for the objects to encourage better prompt alignment of our method (e.g., "a horse with a golden body and pale mane", "a sports car with sleek body and white racing strips") and diverse background prompts. We manually filter out implausible results, yielding approximately 6.6K augmented images ($\sim$ 8 viewpoints per object). During training, we sample equally from the rendered and photorealistic datasets. See Supp. C for further dataset construction details.

4. Experiments

4.1. Training Details

We use Harmon [42], a unified multimodal model with an LLM backbone [47] and a MAR decoder [14], as our primary T2I backbone. We fine-tune the backbone jointly with the viewpoint MLP using the backbone's standard image generation loss. We initialize from a pretrained Harmon checkpoint [44] and fine-tune for 7,500 iterations with a batch size of 192 using AdamW [21]. We apply separate learning rates: a higher rate of 2×10^{-4} for the newly introduced ViewpointMLP and a lower rate of 2×10^{-5} for the pretrained Harmon LLM and MAR decoder. Training takes approximately 28 hours on a single NVIDIA A100 (80 GB).

4.2. Evaluation Metrics

We evaluate our method on two aspects: (i) viewpoint accuracy for each camera parameter, and (ii) prompt alignment using CLIP similarity [32] and the GenEval benchmark [8].

Viewpoint Accuracy. To measure geometric fidelity, we train a viewpoint regressor following similar evaluation protocol of Compass Control [27]. The regressor achieves a mean azimuth error of $4.16°$ on images synthesized with ControlNet and Canny edges of rendered 3D objects. We provide further details of the regressor in Supp. F.

Prompt Alignment. To verify that viewpoint conditioning preserves the model's text-to-image generation ability, we evaluate on general benchmarks. CLIP similarity quantifies semantic correspondence between generated images and prompts, while the *Single Object* and *Color* cases in GenEval assess object presence and descriptive fidelity.

4.3. Testing Dataset

We evaluate on 11 "easy" test objects from Compass Control [27] and 26 additional "diverse" objects spanning animals, vehicles, furniture, people, and mythical creatures. Eleven of these objects do not appear in our training data, testing cross-category generalization. For each "diverse" object, we generate three descriptive phrases and combine them with background prompts to test the generalization abilities. Each object-background pair is rendered with 10 random viewpoints, totaling 5,550 test samples.

To test robustness and object-background consistency on

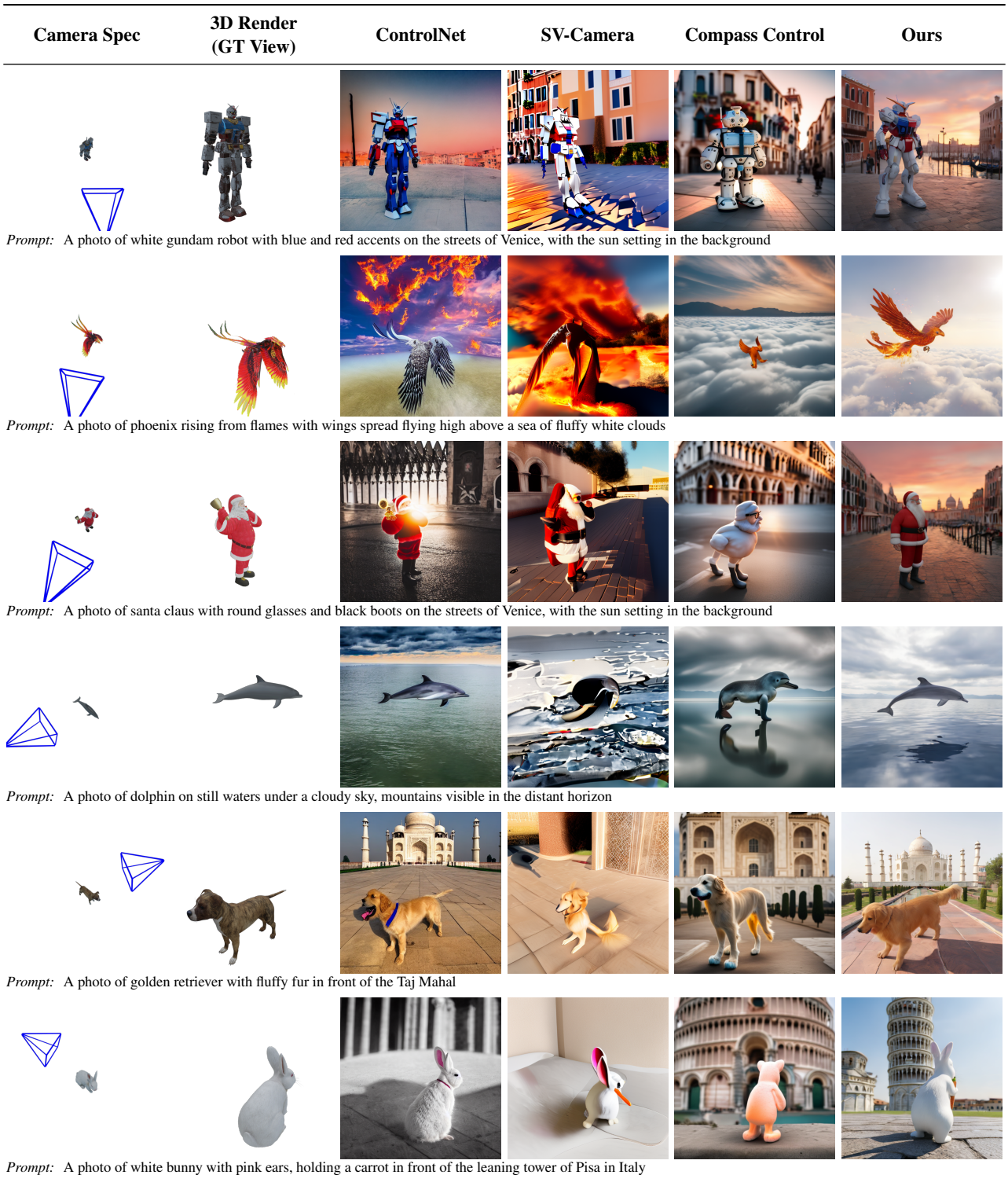


Figure 4. **Qualitative comparison across methods.** Each row shows images and the corresponding prompt. The first two columns visualize the ground-truth camera frustum and a rendering of a 3D object similar to the prompt description. The depth-map of the rendered object is used as an oracle to guide ControlNet but *not used by the other models*. The remaining columns show results from different methods: ControlNet [50], Stable-Virtual-Camera (SV-Camera) [52], Compass Control [27], and our method. Our approach achieves precise viewpoint control while maintaining high image quality and prompt fidelity.

Table 2. **Quantitative comparison** of camera pose fidelity and CLIP score. Methods are evaluated on mean and median angular errors (degrees), radius error (normalized by object size), and CLIP prompt-image similarity. Our approach achieves the best performance across all metrics among models without using oracle geometry information.

Method	Input Type	Azimuth↓		Elevation↓		Radius↓		Yaw↓	Pitch↓	CLIP↑
		Mean	Median	Mean	Median	Mean	Median	Mean	Mean	Mean
ControlNet [50]	Image + Oracle Depth	25.65	5.25	5.77	4.21	0.09	0.07	0.80	0.94	0.3307
SV-Camera [52]	Image + Camera	54.89	23.52	9.05	6.99	0.20	0.18	2.89	2.29	0.2596
Compass [27]	Text + Azimuth Token	31.07	11.60	14.49	12.58	0.17	0.14	2.03	2.61	0.3433
Ours (Harmon)	Text + Camera Token	**18.11**	**7.63**	**7.62**	**6.03**	**0.11**	**0.09**	**1.25**	**1.38**	**0.3555**

Table 3. **Azimuth error breakdown** across 11 "easy" and 26 "diverse" objects.

Method	Whole set	Easy set	Diverse set
ControlNet [50]	25.65	23.22	26.87
SV-Camera [52]	54.89	62.08	51.29
Compass [27]	31.07	18.62	37.29
Ours (Harmon)	**18.11**	**16.22**	**19.06**

Table 4. **GenEval benchmarks** for single object and color adherence.

Model	GenEval [8] ↑	
	Single Obj.	Colors
SD2.1 [34]	98	85
Compass Control [27]	85.62	58.78
Decrease from SD2.1	*-14.28*	*-33.82*
Harmon [42]†	99.90	92.60
Ours (Harmon)	94.38	76.60
Decrease from Harmon	***-5.52***	***-16.00***

†Using checkpoint from [44].

challenging camera angles, we construct an additional test set with 2 back views ($\theta_{az} \in [\frac{3}{4}\pi, \frac{5}{4}\pi]$) and 2 high-elevation views ($\theta_{el} = \frac{2\pi}{9}$) per object–background combination, totaling 2,220 samples.

4.4. Baselines

- **ControlNet-Depth** [50] (text + depth). Provides an oracle baseline with perfect geometry by using depth maps from 3D objects placed on a ground plane.
- **Stable-Virtual-Camera (SV-Camera)** [52] (image + camera). Performs novel-view synthesis given a front-view image input, which we generate using ControlNet-Plus [45] with depth map conditioning.
- **Compass Control** [27] (text + object azimuth). Encodes azimuth orientation tokens and uses 2D bounding boxes for localization. For fair comparison, we extract 2D boxes from test object renderings as additional conditions.

4.5. Quantitative Results

Viewpoint Accuracy. Table 2 summarizes quantitative results. Our method has lower errors than Compass Control and Stable-Virtual-Camera across all five camera parameters. ControlNet-Depth performs slightly better on a few parameters due to oracle access to depth information. Table 3 shows a breakdown of the azimuth error on the "easy" set and the "diverse" set. Compass Control shows a large discrepancy between performance on the 11 "easy" and 26 "diverse" objects, while our method maintains low azimuth errors on both sets, demonstrating stronger generalization.

Prompt Alignment. As shown in Tabs. 2 and 4, our method achieves higher GenEval scores than Compass Control and a higher CLIP similarity than all baselines. Our approach maintains better prompt fidelity of the backbone model than Compass Control does. Compass Control often fails to generate the correct object or color specified in the prompt, reflecting overfitting to its training distribution.

Challenging Viewpoints. Table 5 reports performance on challenging back-view and high-elevation configurations. Our method retains superior accuracy under these extreme conditions, whereas Compass Control degrades sharply. ControlNet-Depth achieves high pitch accuracy due to explicit geometric supervision, but its text alignment remains weaker. These results demonstrate our method's robustness to rare and difficult viewpoints.

Generalization Ability. We quantify the overfitting problem of Compass Control by calculating the percentage of overfitting to its training objects. Among the three testing objects ("Santa Claus", "dolphin", and "rabbit"), Compass Control overfits to lions, ostriches, shoes, sofas, and teddy bears 94.2% of the time, showing that it overfits to category-specific correlations rather than learning factorized viewpoint representations. In contrast, our method has no obvious overfitting among the three objects and consistently produces semantically correct outputs across all categories, indicating effective disentanglement between viewpoint and object identity. See Supp. E for a detailed report.

Table 5. **Challenging viewpoints.** Camera pose accuracy and CLIP score. Values in parentheses show differences compared to the main test set.

Method	Azimuth↓			Elevation↓	Radius↓	Yaw↓	Pitch↓	CLIP↑
	Whole	Easy	Diverse					
ControlNet [50]	33.74(+8.09)	30.31(+7.09)	35.45(+8.58)	5.51(-0.26)	0.10(+0.01)	0.77(-0.03)	1.92(+0.98)	0.3162(-0.0145)
Compass [27]	39.07(+8.00)	28.66(+10.04)	44.27(+6.98)	21.21(+6.72)	0.19(+0.02)	1.77(-0.26)	4.79(+2.18)	0.3421(-0.0012)
Ours (Harmon)	**23.27**(+5.16)	**21.56**(+5.34)	**24.11**(+5.05)	**9.82**(+2.20)	**0.12**(+0.01)	**1.27**(+0.02)	**2.57**(+1.19)	**0.3522**(-0.0033)



(a) Rendered (b) ControlNet (c) Compass (d) Ours

Figure 5. **Comparisons on a high-angle view.** Prompt: A photo of a sedan in an ancient Greek temple ruin, with broken columns and weathered stone steps.



(a) 0° (b) 20° (c) 40°

Figure 6. **Results at varying camera elevations**: 0, 20, 40 degrees. The horizon line changes with the camera elevation.

4.6. Qualitative Results

Figure 4 compares methods qualitatively across diverse camera viewpoints and text prompts. ControlNet preserves object contours via access to accurate geometry but often fits geometrically incoherent content into shapes without semantic awareness. Stable-Virtual-Camera struggles with occluded regions, sometimes producing invalid shapes for novel viewpoints. Compass Control reproduces correct azimuths for seen categories but overfits heavily for novel objects—e.g., generating "Santa Claus" as an animal, "dolphin" as a four-legged creature, "rabbit" as a teddy bear. In contrast, our method generalizes well across novel categories (e.g., Gundam, Phoenix, Santa Claus), demonstrating that our viewpoint tokens capture better geometric conditions independent of object semantics.

High-elevation cases further highlight this distinction: other methods fail to follow the prompt or the viewpoint. This limitation arises from Compass Control's restricted cross-attention mechanism to local object regions and T2I backbone's training bias towards eye-level viewpoints. However, our method learns viewpoint conditioning jointly across foreground and background, producing globally coherent compositions, as shown in Figures 5 and 6. This global understanding demonstrates the potential of learning 3D geometry information in text prompts for image generation.

We demonstrate our method's robustness and generalization further on objects that do not exist in reality. Figure 7 presents three examples generated from imaginative text prompts. Our method produces visually plausible and diverse images that faithfully follow both the prompt semantics and the specified viewpoints. It further shows that we preserve the backbone models' understanding of text and images while embedding viewpoint tokens into the input space for text-to-image generation.

To demonstrate the extensibility of our framework, we retrain a variant of our method on the Compass Control two-object dataset. As shown in Fig. 8, it allows independent control over each object's orientation.

Table 6. **Backbone Variation and Ablation study**

Method	Azimuth	Elevation	Radius	Yaw	Pitch
Ours (Harmon)	18.11	7.62	0.11	1.25	1.38
Ours (SD2.1)	19.16	6.82	0.11	1.18	1.16
Ours (SD3.5)	12.85	8.09	0.17	2.75	1.97
Plücker rays	21.61	8.43	0.19	1.29	1.51
12D matrix	24.44	8.74	0.20	4.89	4.66
sinusoidal encoding	60.90	9.05	0.13	1.69	1.78
No rendered subset	22.98	9.34	0.20	4.84	4.93
Freezing backbone	40.19	8.47	0.14	1.83	2.07
More tokens	18.03	7.45	0.13	1.80	1.87

4.7. Ablations and Variations

The following experiments are summarized in Table 6. **Backbone.** To isolate the contribution of our method from the Harmon backbone, we train two variants using Stable Diffusion 2.1 [34] and Stable Diffusion 3.5 [7] with the same camera encoding architecture. They both achieve comparable viewpoint accuracy, confirming that the geometric generalization stems from our method and dataset

rather than the backbone alone.

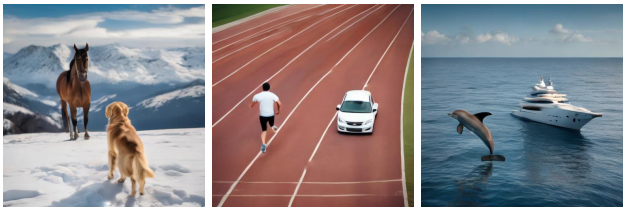
Viewpoint Encoding. We compare our factorized encoding against Plücker rays, 12D camera matrices, and sinusoidal positional encodings, which all underperform our encoding. We hypothesize that while Plücker rays excel at dense pixel-wise correspondence for multi-view generation and sinusoidal positional encodings capture high-frequency details, our single-view T2I setting benefits from disentangled semantic orientation signals that are easier for the model to learn. High-frequency components will introduce training instability and the entangled representation will make it difficult to separate camera position (azimuth and elevation) from camera rotation (yaw and pitch).

Other Ablations. Removing the rendered subset of the training data leads to a substantial accuracy drop, confirming its importance for learning geometric consistency. Fine-tuning both the LLM and MAR modules is also crucial, suggesting the backbone model [42] does not have 3D geometry-aware representations in its text input space. Adding additional tokens yields no observable benefit.



(a) (b) (c)

Figure 7. **Non-existent objects.** They use the same viewpoint as Fig. 1. (a): A small car made of vines and flowers on a countryside road, (b): A flying car with wings made of energy ribbons flying through a storm of glowing auroras over the Arctic, (c): An origami elephant standing on a wooden desk under soft sunlight.



(a) (b) (c)

Figure 8. **Multi-object viewpoint control.** (a) Golden retriever and horse, azimuth: 170°, -10°. (b) Running man and sedan, azimuth: -160°, 20°. (c) Dolphin and yacht, azimuth: 120°, -60°.

4.8. Limitations

Even though our method is capable of generating diverse viewpoints, the T2I backbones have a strong prior toward eye-level, horizontally centered views, particularly for well-known landmarks (e.g., "Taj Mahal"). This bias can cause the model to favor centered horizons for prompts with landmarks. We also occasionally observe degraded generations in human faces and fine structural details. Figure 9 illustrates the three representative failure cases. Our dataset currently covers elevation angles only in $[0°, 45°]$ and excludes roll rotations and intrinsics, as synthesizing reliable photorealistic data for rare viewpoints remains challenging. Even with ground-truth 3D renderings as reference, Nano Banana [9] often fails under extreme viewpoints (Fig. 10).



(a) (b) (c)

Figure 9. **Examples of failure cases.** (a–b) Red circles highlight errors; (c) Misaligned background viewpoints.



(a) (b) (c) (d)

Figure 10. Examples of **Nano Banana** [9] dataset augmentation failing at extreme elevation (75°, a–b) and roll (30°, c–d). (a, c) Reference; (b, d) augmented output.

5. Conclusion

We present a method for precise camera viewpoint control in text-to-image generation through learnable viewpoint tokens. By fine-tuning an image generation model on curated 3D renderings with photorealistic augmentation, we achieve state-of-the-art viewpoint accuracy while preserving image quality and prompt fidelity. Compared to previous work, our approach expands from azimuth control to flexible camera control and encourages global scene understanding, which is important for generating images with a consistent background and object viewpoint. Our method and dataset design work for different backbones and promote generalization to unseen object categories. Overall, results demonstrate that text prompts can internalize explicit 3D camera structure through simple parametric encoding, opening a new pathway toward geometrically-aware text-to-image generation systems.

Acknowledgements. This work was supported in part by the DARPA Perceptually enabled Task Guidance (PTG) Program under contract number HR00112220005, and by funding from the UCI CS Department.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 2

[2] Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2724–2733, 2025. 3

[3] James Burgess, Kuan-Chieh Wang, and Serena Yeung-Levy. Viewpoint textual inversion: Discovering scene representations and 3d view control in 2d diffusion models. In *European Conference on Computer Vision*, pages 416–435. Springer, 2025. 1, 2, 3

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 3

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 3

[6] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3, 7

[8] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023. 4, 6

[9] Google. Gemini 2.5 flash image (nano banana). https://aistudio.google.com/models/gemini-2-5-flash-image. Accessed: 2025-11-11. 1, 4, 8, 2

[10] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 3

[11] 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*, pages 770–778, 2016. 2

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 3

[13] Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2671–2682, 2025. 2, 3

[14] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 2024. 2, 4

[15] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023. 2

[16] 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*, pages 300–309, 2023. 3

[17] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3

[18] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[19] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1

[20] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 1, 2, 3

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4

[22] Léopold Maillard, Tom Durand, Adrien Ramanana Rahary, and Maks Ovsjanikov. Laconic: A 3d layout adapter for controllable image creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18046–18057, 2025. 2

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421, 2020. 3

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable

ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[25] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulo, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10258–10268, 2024. 1, 3

[26] OpenAI. Chatgpt (gpt-5). https://chat.openai.com/, 2025. Large language model. 2

[27] Rishubh Parihar, Vaibhav Agrawal, Sachidanand VS, and Venkatesh Babu Radhakrishnan. Compass control: Multi object orientation control for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2791–2801, 2025. 1, 2, 3, 4, 5, 6, 7, 8, 9

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

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

[31] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9925, 2024. 1, 3

[32] 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. 4

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 6, 7

[35] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1, 3

[36] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[38] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3

[39] Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint 2307.01097*, 2023. 1, 3

[40] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 3

[41] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025. 2

[42] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025. 2, 3, 4, 6, 8, 1

[43] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2

[44] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025. 4, 6

[45] xinsir6. ControlNetPlus: All-in-one controlnet for image generation and editing. https://github.com/xinsir6/ControlNetPlus, 2024. 6, 2

[46] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18430–18439, 2022. 3

[47] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei,

Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 4

[48] 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. 2

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 3

[50] 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*, pages 3836–3847, 2023. 2, 5, 6, 7, 8, 9

[51] Yibo Zhang, Shaoxiang Zhang, Size Wu, Kejun Qian, Dazhong Ji, Chen Change Loy, Wenbin Yang, and Guosheng Lin. Texverse: A universe of 3d objects with high-resolution textures. *arXiv preprint arXiv:2508.10868*, 2025. 4

[52] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 1, 2, 3, 5, 6, 7, 8, 9