

MUSES: 3D-Controllable Image Generation via Multi-Modal Agent Collaboration

Yanbo Ding^{1, 2}, Shaobin Zhuang^{3, 4}, Kunchang Li^{1, 2, 3}, Zhengrong Yue^{3, 4}, Yu Qiao^{1, 3}, Yali Wang^{1, 3*}

¹Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences

³Shanghai Artificial Intelligence Laboratory

⁴Shanghai Jiao Tong University
{yb.ding, yl.wang}@siat.ac.cn

Abstract

Despite recent advancements in text-to-image generation, most existing methods struggle to create images with multiple objects and complex spatial relationships in the 3D world. To tackle this limitation, we introduce a generic AI system, namely **MUSES**, for 3D-controllable image generation from user queries. Specifically, our MUSES develops a progressive workflow with three key components, including (1) Layout Manager for 2D-to-3D layout lifting, (2) Model Engineer for 3D object acquisition and calibration, (3) Image Artist for 3D-to-2D image rendering. By mimicking the collaboration of human professionals, this multi-modal agent pipeline facilitates the effective and automatic creation of images with 3D-controllable objects, through an explainable integration of top-down planning and bottom-up generation. Additionally, existing benchmarks lack detailed descriptions of complex 3D spatial relationships of multiple objects. To fill this gap, we further construct a new benchmark of T2I-3DisBench (3D image scene), which describes diverse 3D image scenes with 50 detailed prompts. Extensive experiments show the state-of-the-art performance for MUSES on both T2I-CompBench and T2I-3DisBench, outperforming recent strong competitors such as DALL-E 3 and Stable Diffusion 3. These results demonstrate a significant step forward of MUSES in bridging natural language, 2D image generation, and 3D world.

Code — <https://github.com/DINGYANB/MUSES>

Dataset — <https://huggingface.co/yanboding/MUSES>

Extended version — <https://arxiv.org/abs/2408.10605>

Introduction

Text-to-image generation (Rombach et al. 2022a; Ramesh et al. 2021; Midjourney 2024) is rapidly evolving in quality. However, such generation often struggles with detailed user queries of multiple objects in complex scenes. Several approaches have emerged to address this by compositional text-to-image synthesis (Yang et al. 2024b; Chefer et al. 2023; Feng et al. 2024, 2023). Unfortunately, they fail to accurately control 3D contents like object orientation and camera view, even though our real world is inherently three-dimensional. As shown in Fig. 1, when the prompt is “a

frog facing right is behind a rabbit facing forward”, existing methods collapse with unsatisfactory 3D arrangements. This raises a fundamental question: can we create images with precise 3D control to better simulate our world?

To answer this question, we draw inspiration from the workflow of 3D professionals. We observe that creating such images typically involves three key stages: scene layout planning, 3D objects design, and image rendering (Pharr, Jakob, and Humphreys 2023). This highlights the need for developing a systematic framework of 3D-controllable image creation, rather than relying on a single generation model. Therefore, we propose a generic AI system that automatically and precisely creates images with 3D controllable objects. We name it **MUSES**, since we consider human designers as our “Muses”, and mimic their workflows through a collaborative pipeline of multi-modal agents.

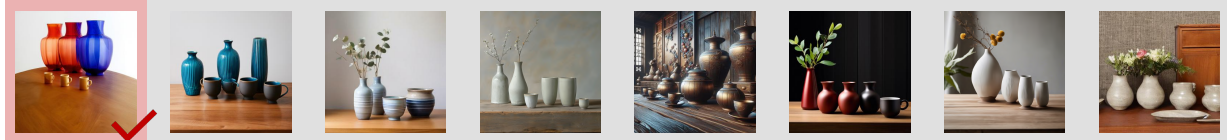
Our MUSES system, as depicted in Fig. 2, comprises three key agents that progressively achieve 3D-controllable image generation: (1) *Layout Manager*. First, we employ a Large Language Model (e.g., Llama3 (AI@Meta 2024)) to plan and assign 3D object locations based on user queries. Our innovative 2D-to-3D layout lifting paradigm first generates a 2D layout by in-context learning, then elevates it to a 3D layout via chain-of-thought reasoning. (2) *Model Engineer*. After obtaining the 3D layout, we introduce a model engineer to acquire 3D models of queried objects. To enhance its robustness, we design a flexible retriever that gathers 3D models through a decision tree approach, combining self-collected model shop retrieval, online search, and text-to-3D generation. Furthermore, to ensure orientation alignment with the planned 3D layout, we develop a novel aligner to calibrate object orientation by face-camera-view identification with CLIP (Radford et al. 2021). (3) *Image Artist*. Finally, we introduce an image artist to render 3D-controllable images. The 3D-aligned objects and their layouts are fed into Blender (Community 2018), which accurately assembles all the objects into 3D-to-2D condition images. These conditions are then used with ControlNet (Zhang, Rao, and Agrawala 2023) to generate the final image.

Our contributions are threefold. *First*, our MUSES is the first AI system for 3D-controllable image generation, to our best knowledge. It enables precise control over object properties such as object count, orientation, 3D spatial relationships, and camera view, potentially bridging the gap between

*Corresponding Author.

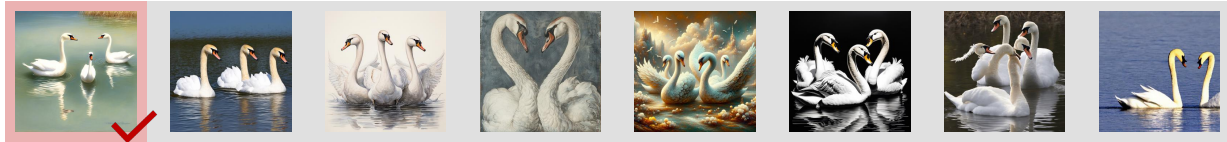
(a) Prompt containing **multiple objects**, **3D spatial relationships**, and **camera view**:

Three vases in the **back** and **three cups** in the **front**, evenly spaced on wooden table, **right view**.



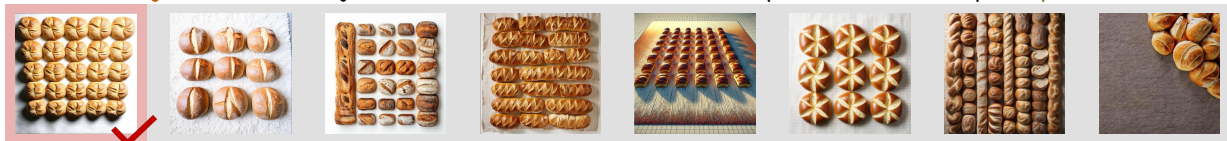
(b) Prompt containing **multiple objects**, **3D spatial relationships**, and **orientations**:

Three swans, one on the left facing **right**, one in the middle facing **forward**, one on the right facing **left**.



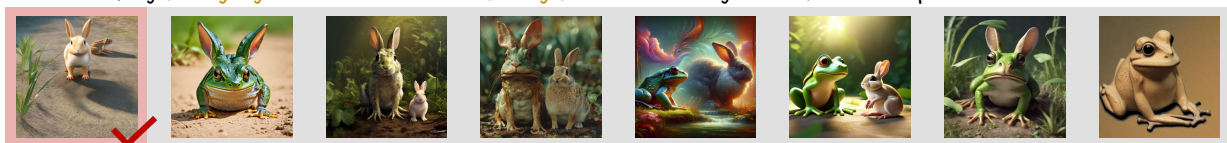
(c) Prompt containing **multiple objects**, **3D spatial relationships**, and **camera view**:

Twenty five bread arranged in **five horizontal lines**, fill the entire picture, on white carpet, **top view**.



(d) Prompt containing **3D spatial relationships**, **orientations**, and **camera view**:

A frog facing **right** is **behind** a rabbit facing **forward** on the ground, **front view**, photorealistic.



MUSES (Ours)

SD 3

Playground 2.5

Midjourney v6

DALLE3

Hunyuan-DiT

RPG-Diffusion

Structured Diffusion

Figure 1: **Comparison Results With Various Methods.** Our MUSES achieves the best, with object numbers highlighted in **brown**, object orientations in **yellow**, 3D spatial relationships in **blue**, and camera views in **green**, outperforming both open-sourced state-of-the-art methods and commercial API products, such as Stable Diffusion V3, DALL-E 3, and Midjourney v6.0.

image generation and world simulation. *Second*, MUSES is a distinct multi-agent collaboration for 3D-controllable image generation. Each agent of MUSES is an insightful and novel integration of multi-modal agents, allowing for top-down planning and bottom-up generation with robust control of 3D information. *Finally*, since existing benchmarks lack detailed descriptions of complex 3D information like object orientation and camera view, we further construct a new benchmark of T2I-3DisBench (3D image scene), which consists of 50 prompts involving multiple objects with diverse object orientations, 3D spatial relationships, and camera views across various complex 3D scenes. Extensive experiments demonstrate the superiority of MUSES on both T2I-CompBench and our T2I-3DisBench, where it consistently outperforms state-of-the-art competitors of both open-source models and commercial API products, including Stable Diffusion v3 (Esser et al. 2024), DALL-E 3 (Betker et al. 2023), and Midjourney v6.0 (Midjourney 2024).

Related Work

Controllable Image Generation. Before diffusion (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015), GAN-based (Creswell et al. 2018) methods such as ControlGAN (Lee and Seok 2019) and AttnGAN (Fang et al. 2022), in-

corporate text features via attention (Vaswani et al. 2017) modules to guide image generation. In recent years, Stable Diffusion series (Podell et al. 2023; Rombach et al. 2022b,a) have dominated the text-to-image generation market. Given that text-based control is insufficient for precise image generation, ControlNet (Zhang, Rao, and Agrawala 2023) introduced additional fine-grained conditions (e.g., depth maps). Additionally, methods such as Structured-Diffusion (Feng et al. 2023) and Attn-Exct (Chefer et al. 2023) fuse linguistic structures or attention-based semantic guidance into the diffusion process. Approaches like LayoutGPT (Feng et al. 2024) and LMD (Lian et al. 2023) use LLM to plan 2D layouts (bounding boxes), while RPG (Yang et al. 2024b) plans and assigns regions based on the user input. However, existing methods struggle to control 3D properties of objects. We instead plan 3D layouts and incorporate 3D models and simulations to achieve 3D controllable image generation.

LLM-Based Agents. LLMs like GPT series (Brown et al. 2020; Achiam et al. 2023) and Llama series (Touvron et al. 2023b,a) have revolutionized natural language processing (Chowdhary and Chowdhary 2020). MLLMs like LLaVA (Liu et al. 2024) and InternVL (Chen et al. 2024b) have enabled impressive performance on visual tasks. The combination of LLMs and MLLMs in multi-agent systems achieves remarkable success across various domains, including visual

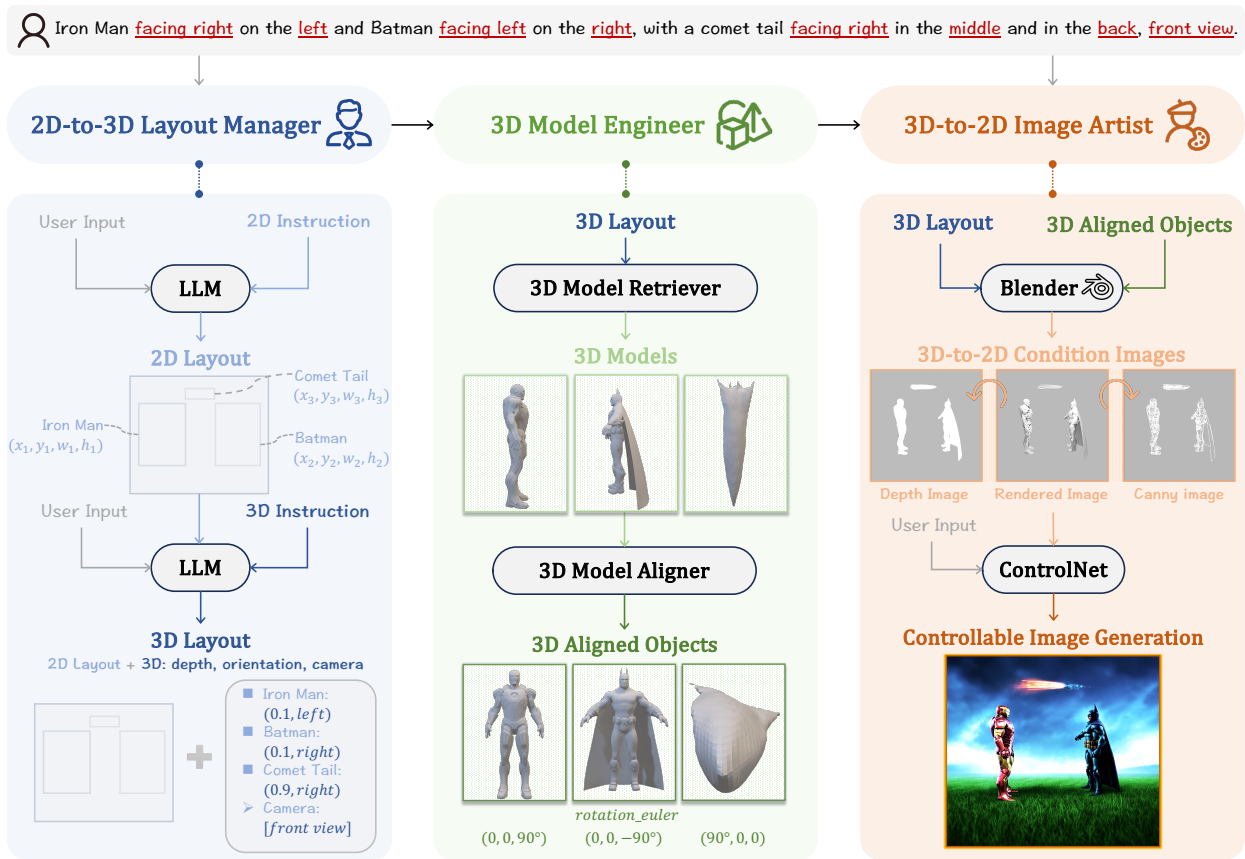


Figure 2: **Overview of our MUSES.** Based on the user input, Layout Manager first plans a 2D layout and lifts it to a 3D one. Then, Model Engineer acquires 3D models of query objects and aligns them to face the camera. Finally, Image Artist assembles all the 3D object models into visual conditions that are used for final controllable image generation.

understanding (Kelly et al. 2024; Wu et al. 2023a), gaming (Li et al. 2023; Gong et al. 2023), software development (Wu et al. 2023b; Qian et al. 2023), video generation (Yuan et al. 2024; Yang et al. 2024a), and autonomous driving (Wei et al. 2024; Palanisamy 2020). We focus on image generation via multi-agent collaboration. DiffusionGPT (Qin et al. 2024) uses LLM to select models in image generation. SLD (Wu et al. 2024) uses LLM for self-correcting the generated image. CompAgent (Wang et al. 2024) uses LLM to coordinate the image generation process into sub-steps. Unlike these works, we use LLM to plan and lift 2D layouts to 3D.

Method

In this section, we introduce our MUSES for 3D controllable image generation. As shown in Fig. 2, it is a generic AI system with a distinct multi-modal agent collaboration pipeline. Specifically, our MUSES consists of three collaborative agents including Layout Manager for 2D-to-3D layout lifting, Model Engineer for 3D object acquisition and calibration, Image Artist for 3D-to-2D image rendering.

Lifting: 2D-to-3D Layout Manager

To achieve precise 3D control, we first plan a 3D layout according to the user input. Specifically, we employ LLM (e.g.,

Llama3 (AI@Meta 2024)) as a layout manager due to its great power of linguistic understanding. To alleviate planning difficulty, we design a 2D-to-3D lifting paradigm for the progressive transition from 2D to 3D layout in Fig. 3.

2D Layout Planning via In-Context-Learning. We start by planning the 2D position of the objects. Apparently, asking LLM directly to generate an object layout is not ideal, as it may struggle with managing bounding boxes in images. Hence, we leverage In-Context Learning (Dong et al. 2022), allowing LLM to follow instructions with a few examples from a 2D layout shop. We use the NSR-1K dataset (Feng et al. 2024) as the 2D layout shop, since it contains over 40,000 entries with diverse prompts, objects, and the corresponding bounding boxes. As the original NSR-1K dataset lacks layouts in 3D scenes and complex scenes, we manually design and add some 3D layouts (details in Extended Version). First, we use CLIP (Radford et al. 2021) text encoder to compare similarities of user input and text prompts in the shop. Then, we select the top five 2D layouts with the highest similarity scores. Finally, we feed these contextual layouts along with instructions to LLM. Such a comprehensive approach results in a contextually relevant 2D layout, forming the foundation for the subsequent 3D lifting.

3D Layout Planning via Chain-of-Thought. Unlike previ-

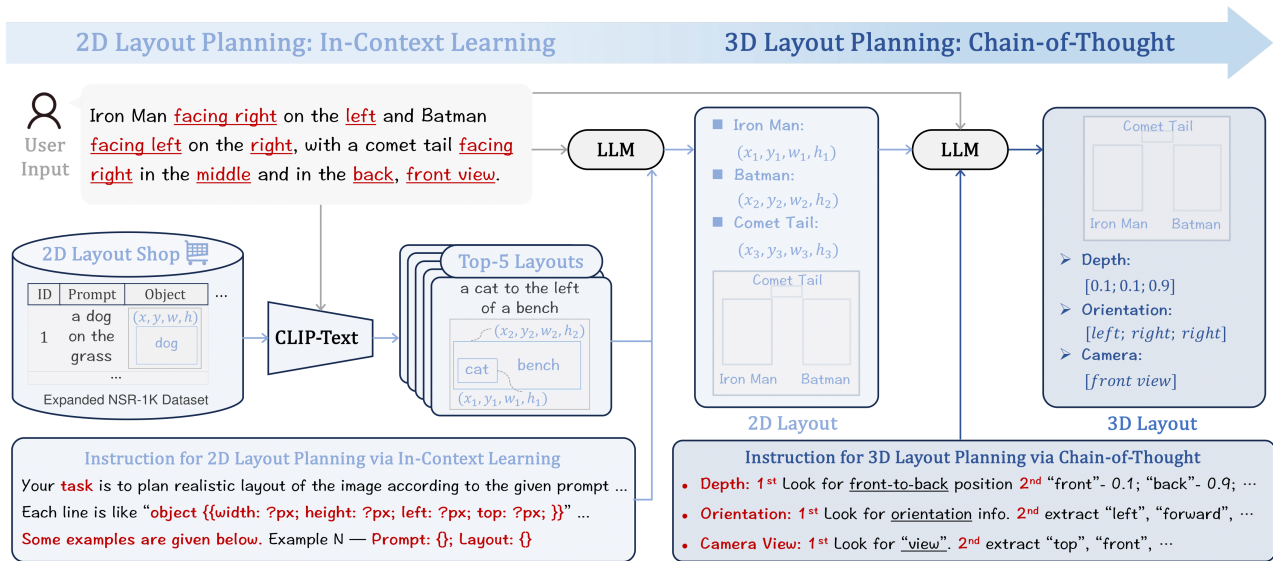


Figure 3: **2D-to-3D Layout Manager**. First, based on the user input, our layout manager employs the LLM to plan 2D layout through In-context Learning. Then, it lifts the 2D layout to 3D space via Chain of Thought reasoning.

ous approaches (Feng et al. 2024; Qu et al. 2023; Lian et al. 2023) that directly use 2D layout for image generation, we lift our 2D layout to 3D space. Specifically, With 2D layout and user input, we further ask LLM to plan 3D attributes, including depth, orientation, and camera view. To enhance complex planning capability, we design a chain-of-thought (CoT) prompt (Wei et al. 2022) for step-by-step reasoning on each attribute. Taking depth as an example, the first step instruction is to look for front-to-back position relationships in user input. Based on such relationships, the second-step instruction is to assign a depth value to each object, (e.g., “A is in front of B”: A’s depth is set to 0.1 and B’s depth is set to 0.9). The instructions for orientation and camera view are similar. We list all instructions in our Extended Version. Via such a concise manager, we can accurately exploit the 3D layout in user input for subsequent 3D simulation in Blender.

Calibrating: 3D Model Engineer

After planning the 3D layout, the next step is to acquire specified 3D models. Specifically, we introduce a Model Engineer which comprises two key roles: Model Retriever for acquiring 3D models in the 3D layout, Model Aligner for calibrating the orientations of 3D models to face camera.

3D Model Retriever: Retrieve-Generate Decision Tree. To obtain 3D models with efficiency and robustness, we design a concise Model Retriever via the decision tree of retrieval and generation in Fig. 4. Our motivation is that, 3D models from the internet have higher quality compared to those generated by text-to-3D. Hence, we prioritize the retrieval of existing 3D models in the decision tree, which not only enhances the overall visual quality of 3D models, but also reduces computational load in text-to-3D synthesis.

Specifically, our model retriever is based on a 3D model shop (300 3D models of 230 object categories) that is self-built in an online fashion. *First*, the model retriever looks

for 3D models of queried objects from the current shop, using the object name as the search key. *Second*, if it cannot find any 3D models of a queried object from this shop (e.g., Batman in Fig. 4), it will search online, e.g., on the professional website (Marius Kalytis 2011). For each object, there may exist numerous 3D models from such a website. Hence, we perform CLIP text encoder to calculate the similarity between the object name and the online item title, and select the 3D model with the highest similarity. *Third*, if online search also fails in finding suitable 3D models (e.g., Comet Halley in Fig. 4), it will employ a text-to-3D generation model like Shap-E (Jun and Nichol 2023) or 3DTopia (Hong et al. 2024) to synthesize corresponding 3D model. *Finally*, we add the newly found object models to our 3D model shop to enhance shop diversity and versatility of future usage. In such a manner, we obtain 3D models of query objects and remain up-to-date with the latest available high-quality models, ensuring both efficiency and robustness.

3D Model Aligner: Face-Camera Calibration. As 3D models are acquired from internet or generation, their orientations may not align with the expected ones in the planned 3D layout. To address this, we introduce a novel 3D Model Aligner, which can calibrate orientations of 3D models to face camera, for further usage along with 3D layouts. We propose to fine-tune CLIP as a binary classifier (Fig. 5) for its strong generalization capacity by large-scale pretraining.

Fine-tuning CLIP as a Face-Camera Classifier. First, we need to prepare the fine-tuning dataset. We randomly select 150 3D models from our 3D shop and import them into Blender with a standardized environment. For each 3D model, we perform multi-view rendering to generate a set of 2D images from various views, with different rotation_euler parameters in Blender. Then, we annotate images by tagging the description of “object name (faces / not face) camera”. To increase data diversity across 3D geometries, we

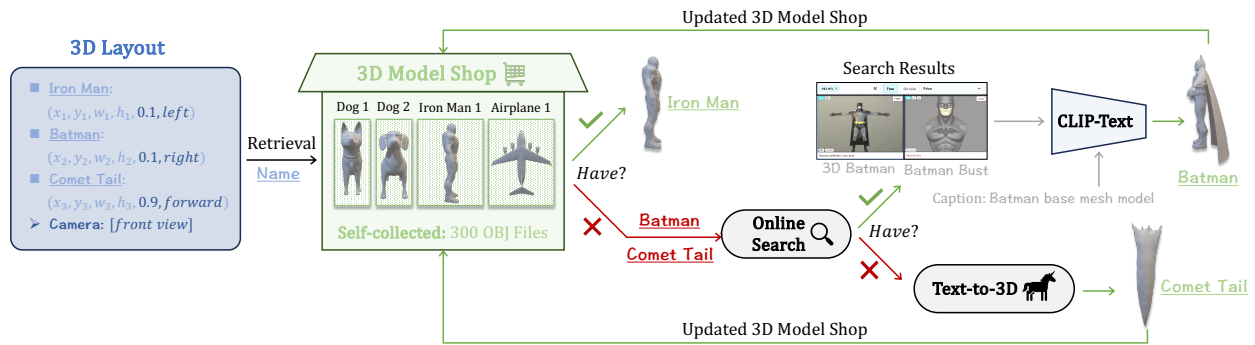


Figure 4: **3D Model Retriever.** We develop a retrieve-generate decision tree that automatically acquires 3D objects specified in the 3D layout from a self-collected model shop, based on a concise decision process of online search and text-to-3D generation.

randomly select 5 no-face-camera images as negative samples for each 3D model. To balance positive and negative samples during fine-tuning, we make the face-camera image as 5 copies for each 3D model. This results in a training set with 1500 image-text pairs, which are used for fine-tuning CLIP as a Face-Camera Classifier by contrastive language-image learning. After fine-tuning, we test it with an extra 1500 images (50% face camera, 50% not) from other 150 3D models. All test images are correctly classified.

Inferring Face-Camera Image of 3D Object Models. During the inference, we import 3D models of queried objects (from model retriever) into the same Blender environment used in training dataset generation. For each 3D model, we perform multi-view rendering to generate a comprehensive set of 2D images from various views. Then, we use the fine-tuned CLIP to identify the face-camera image of each 3D model, by comparing similarities between the rendered images and the text “object name faces camera”. Based on this image, we can effectively align 3D object models to face the camera through the configuration of the rotation parameter in Blender, which is used to generate correctly orientated objects in the subsequent 3D-to-2D image generation.

Rendering: 3D-to-2D Image Artist

So far, we have obtained a 3D layout and aligned 3D models in the user query. Given these 3D materials, we next introduce a concise image artist to create the 3D-controllable image, based on a 3D-to-2D rendering as shown in Fig. 2. More Blender rendering outputs can be found in our Extended Version. First, we assemble all the 3D object models into a complete scene based on the 3D layout. To ensure consistent and accurate 3D scene composition, we develop a comprehensive setting of parameter configuration in Blender such as settings of rendering environment, camera, and object (detailed Blender configurations are in Extended Version). Once the 3D scene is fully assembled, we use the engine, “CYCLES”, to convert the 3D scene into a 2D image. To enhance the control over the final image generation, we process this 2D rendering into two condition images, including (1) Depth Map by Blender’s Z-pass rendering (Ouza, Ulrich, and Yang 2017), representing 3D spatial relationships. (2) Canny Edge by OpenCV (Bradski 2000), highlighting contours. Finally, we leverage these 3D-to-2D condition im-

ages as fine-grained control, and use advanced image generation techniques like ControlNet (Zhang, Rao, and Agrawala 2023), to generate the final image with the user input. Via such a concise image artist, our MUSES can flexibly generate a 3D-controllable image that accurately reflects both 3D spatial details and semantic contents of the user input.

Experiment

Datasets and Metrics. We first conducted experiments on T2I-CompBench (Huang et al. 2023) due to its comprehensive evaluations of object count and spatial relationships. Since T2I-CompBench lacks detailed text prompts for object orientations and camera views, we further introduce T2I-3DisBench (3D image scene), a dataset of 50 textual prompts that encapsulate complex 3D information. We conducted both automatic and user evaluations on our T2I-3DisBench. Since the metrics of T2I-compBench are inadequate for assessing detailed 3D features, we uniquely employed Visual Question Answering (VQA) on InternVL (Chen et al. 2024b). Specifically, we fed instructions to InternVL, asking it to score the generated images considering four dimensions: count, orientation, 3D spatial relationship, and camera view. Additionally, we conducted user evaluation on our T2I-3DisBench, where participants scored the images based on the same four dimensions. More details about our T2I-3DisBench are shown in Extended Version.

Implementation Details. Our MUSES is modular and extensible, allowing for integration of various LLMs, CLIPs, and ControlNets. In our experimental setup, we employed Llama-3-8B (AI@Meta 2024) for 3D layout planning, ViT-L/14 for image/text encoding, ViT-B/32 for orientation calibration, and SD 3 ControlNet (Zhang, Rao, and Agrawala 2023) for controllable image generation. For Llama-3-8B, we set top_p to 0.1 and temperature to 0.2 to ensure precise, consistent, and reliable outputs. For SD 3 ControlNet, we set inference steps to 20 and control scales from 0.5 to 0.9, as discussed in parameter ablation in Extended Version. During evaluation, we select the best one from the five images of different control scales using benchmark metrics (e.g., T2I-CompBench and T2I-3DisBench). The Blender’s parameter conversion rules and settings are also specified in our Extended Version. We use Mini-InternVL 1.5 (Chen et al. 2024a) for automated evaluation on T2I-3DisBench. Exper-

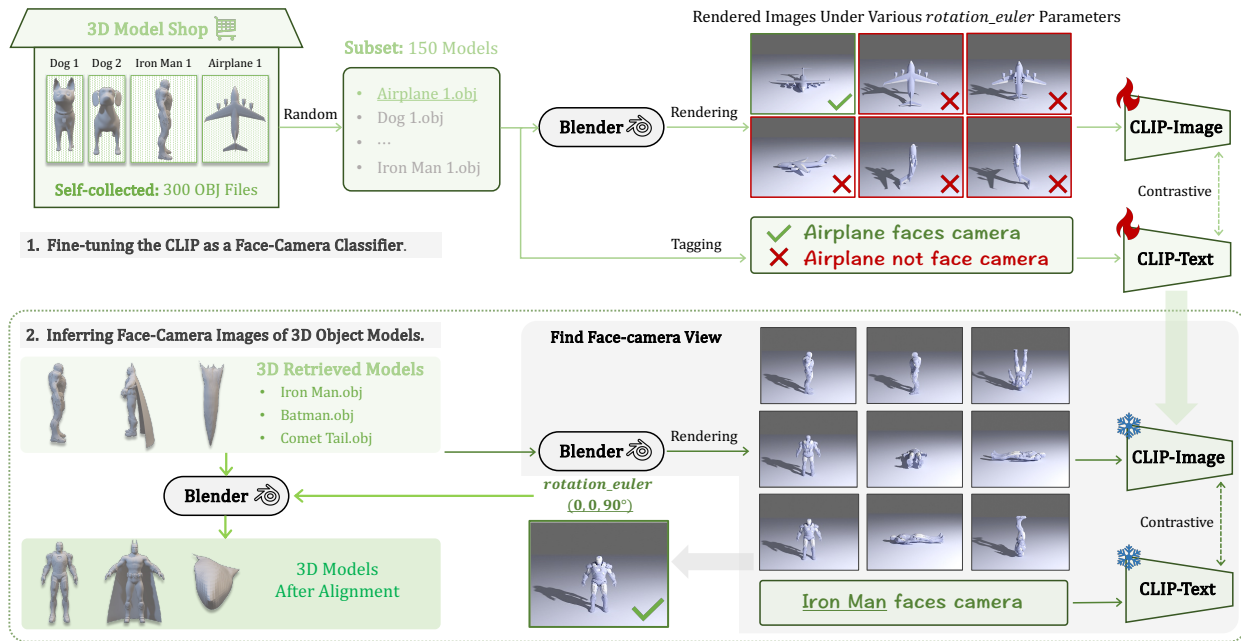


Figure 5: **3D Model Aligner**. It aligns 3D models with face-camera orientation, ensuring that the final orientation conforms to the planned 3D layout. First, we fine-tune CLIP as a Face-Camera Classifier, by a training set generated from our 3D model shop. Then, we use fine-tuned CLIP to identify the face-camera image of each object, aligning its 3D model to face the camera.

iments are conducted on 8 NVIDIA RTX 3090 GPUs.

SOTA Comparison on T2I-Compbench. As shown in Tab. 1, we compared our MUSES with two types of existing SOTA methods: specialized/multi-agent approaches in the upper part and generic models in the lower part. MUSES consistently outperforms both categories across all metrics, including object count, relationship, and attribute binding. Specifically, Our innovative 3D layout planning and 3D-to-2D image conditions enhance object relationship understanding, leading to best performance on spatial-related metrics. Additionally, precise object positioning boosts the numeracy score, and detailed 3D model shape guidance significantly improves attribute binding scores.

SOTA Comparison on T2I-3DisBench. As shown in Tab. 2, we conducted both automatic and user evaluations (in parentheses) on our T2I-3DisBench and compared two types of SOTA methods as well. *For automatic evaluation on InternVL-VQA metrics*, our MUSES consistently outperforms others, including the open-source SOTA model, Stable Diffusion V3, and closed-source API products like Midjourney v6.0. Obviously, existing approaches struggle with complex prompts containing 3D information (e.g., object orientation, camera view), highlighting the importance of our 3D-integration design. *For user evaluation*, We randomly selected 20 prompts from our T2I-3DisBench and engaged 30 participants to rate image accuracy on a scale of 0.0 (poor) to 1.0 (excellent) across four dimensions. The results show a strong preference for MUSES, demonstrating its effectiveness in handling complex 3D scenes.

Ablation Studies. As shown in Table 3, our full system achieves the best performance. Removing any of the components will result in performance degradation. (1) *Object*

Depth Planning is essential, as its removal leads to poor 3D spatial representation in Blender, affecting both datasets. (2) *Object Orientation Planning* is critical for prompts containing orientation information, with a score dropping on T2I-3DisBench when removed. (3) *Camera View Planning* affects performance mostly on T2I-3DisBench, since it contains camera information. (4) *Retrieve-Generate Decision Tree* significantly influences performance on both datasets, highlighting the importance of high-quality 3D models. (5) *Face-Camera Calibration* is the most important; its removal causes sharp performance drops as 3D models lose correct orientation specified in the 3D layout. (6) *CLIP Fine-tuning During Calibration* improves CLIP’s accuracy in determining object orientations, thus enhancing the orientation accuracy of our final image. (7) *Multiple Control Scales* effectively improves performance on both benchmarks. (8) *3D Layout Shop Expansion* helps the LLM generate 3D and complex layouts better; it obviously works. (9) *Co-ablation of Multiple Components* (last three columns) shows that removing the Model Engineer has the most significant impact, resulting in poor object shaping and orientation. Removing the Layout Manager also notably degrades performance. Removing all components will result in the lowest score. These findings demonstrate that each component is crucial for our 3D controllable image generation system.

Conclusion

We introduce MUSES, a multi-agent collaborative system for precise 3D controllable image generation. By integrating 3D layouts, models, and simulations, MUSES achieves fine-grained control over 3D object properties, including object orientation, 3D spatial relationship, and camera view.

Method	Attribute Binding			Object Relationship			Numeracy \uparrow
	Color \uparrow	Shape \uparrow	Texture \uparrow	2D-Spatial \uparrow	3D-Spatial \uparrow	Non-Spatial \uparrow	
LayoutGPT (Feng et al. 2024)	0.2921	0.3716	0.3310	0.1153	0.2607	0.2989	0.4193
Structured Diffusion (Feng et al. 2023)	0.4990	0.4218	0.4900	0.1386	0.2952	0.3111	0.4562
Attn-Exct (Chefer et al. 2023)	0.6400	0.4517	0.5963	0.1455	-	0.3109	-
GORS (Huang et al. 2023)	0.6603	0.4785	0.6287	0.1815	-	0.3193	-
RPG-Diffusion (Yang et al. 2024b)	0.6024	0.4597	0.5326	0.2115	0.3587	0.3104	0.4968
CompAgent (Wang et al. 2024)	0.7400	0.6305	0.7102	0.3698	-	0.3104	-
SDXL (Podell et al. 2023)	0.6369	0.5408	0.5637	0.2032	0.3438	0.3110	0.5145
PixArt- α (Chen et al. 2023)	0.6886	0.5582	0.7044	0.2082	0.3530	0.3179	0.5001
Playground v2.5 (Li et al. 2024a)	0.6381	0.4790	0.6297	0.2062	0.3816	0.3108	0.5329
Hunyuan-DiT (Li et al. 2024b)	0.6342	0.4641	0.5328	0.2337	0.3731	0.3063	0.5153
DALL-E 3 (Betker et al. 2023)	0.7785	0.6205	0.7036	0.2865	-	0.3003	-
SD v3 (Esser et al. 2024)	0.8085	0.5793	0.7317	0.3144	0.4026	0.3131	0.6088
MUSES (Ours)	0.8726	0.6812	0.8081	0.4756	0.4639	0.3226	0.7720

Table 1: **Evaluation Results on T2I-CompBench.** Our MUSES demonstrates the best performance on attribute binding, object relationship, and object count, outperforming SOTA methods, including multi-agent specialized methods, and generic models.

Method	Average Score \uparrow	Object Count \uparrow	Orientation \uparrow	3D Spatial Relationship \uparrow	Camera View \uparrow
Structured Diffusion (Feng et al. 2023)	0.1862 (0.15)	0.2160 (0.14)	0.1647 (0.08)	0.1773 (0.21)	0.1866 (0.18)
RPG-Diffusion (Yang et al. 2024b)	0.2753 (0.18)	0.3209 (0.18)	0.2533 (0.15)	0.3195 (0.25)	0.2075 (0.12)
LayoutGPT (Feng et al. 2024)	0.2348 (0.14)	0.2937 (0.14)	0.2058 (0.12)	0.2783 (0.18)	0.1613 (0.10)
Playground v2.5 (Li et al. 2024a)	0.2344 (0.19)	0.3587 (0.19)	0.1887 (0.17)	0.2071 (0.28)	0.1830 (0.15)
DALL-E 3 (Betker et al. 2023)	0.2627 (0.23)	0.3013 (0.15)	0.2363 (0.18)	0.2370 (0.29)	0.2757 (0.29)
Hunyuan-DiT (Li et al. 2024b)	0.2780 (0.22)	0.3496 (0.19)	0.2517 (0.21)	0.2598 (0.29)	0.2510 (0.18)
Midjourney v6.0 (Midjourney 2024)	0.3760 (0.35)	0.4470 (0.39)	0.3438 (0.27)	0.3518 (0.36)	0.3613 (0.38)
SD v3 (Esser et al. 2024)	0.4206 (0.36)	0.5383 (0.42)	0.3303 (0.26)	0.4600 (0.40)	0.3537 (0.33)
MUSES (Ours)	0.6156 (0.62)	0.7488 (0.61)	0.4709 (0.68)	0.7207 (0.62)	0.5220 (0.57)

Table 2: **Evaluation Results on our T2I-3DisBench.** Our MUSES consistently outperforms other methods across all metrics. Each cell displays both the InternVL-VQA metric value and the corresponding score from user evaluation (in parentheses).

Component	Choice												
Object Depth Planning	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
Object Orientation Planning	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
Camera View Planning	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✗
Retrieve-Generate Decision Tree	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗
Face-Camera Calibration	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✗
CLIP Fine-tuning During Calibration	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗
Multiple Control Scales	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗
3D Layout Shop Expansion	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
Results on T2I-CompBench	0.4639	0.4127	0.4609	0.4636	0.4234	0.3828	0.4289	0.3893	0.3925	0.3641	0.3427	0.3093	
Results on T2I-3DisBench	0.6156	0.4981	0.4938	0.5245	0.5360	0.4019	0.5026	0.5192	0.4405	0.3748	0.3661	0.3178	

Table 3: **Ablation Studies of Different Components on T2I-CompBench and on our T2I-3DisBench.** Where T2I-CompBench uses the 3D-spatial metric because it is most relevant to the 3D environment, and T2I-3DisBench uses the average InternVL-VQA score in terms of object count, object orientation, 3D spatial relationship, and camera view.

To evaluate such complex 3D image scenes more comprehensively, we construct a new benchmark named T2I-3DisBench. Experiments on T2I-CompBench and our newly introduced T2I-3DisBench demonstrate the state-of-the-art

performance of MUSES in handling complex 3D image scenes. Future work will focus on improving efficiency and expanding capabilities to control lighting conditions, and potentially expanding to video generation.

Acknowledgments

This work was supported by the National Key R&D Program of China(NO.2022ZD0160505), the National Natural Science Foundation of China under Grant(62272450), and the Joint Lab of CAS-HK.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- AI@Meta. 2024. Llama 3 Model Card.
- Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3): 8.
- Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.
- Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024a. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.
- Chowdhary, K.; and Chowdhary, K. 2020. Natural language processing. *Fundamentals of artificial intelligence*, 603–649.
- Community, B. O. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.
- Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.
- Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.
- Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.
- Fang, F.; Zhang, P.; Zhou, B.; Qian, K.; and Gan, Y. 2022. Atten-GAN: pedestrian trajectory prediction with gan based on attention mechanism. *Cognitive Computation*, 14(6): 2296–2305.
- Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. *arXiv:2212.05032*.
- Feng, W.; Zhu, W.; Fu, T.-j.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2024. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36.
- Gong, R.; Huang, Q.; Ma, X.; Vo, H.; Durante, Z.; Noda, Y.; Zheng, Z.; Zhu, S.-C.; Terzopoulos, D.; Fei-Fei, L.; et al. 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Hong, F.; Tang, J.; Cao, Z.; Shi, M.; Wu, T.; Chen, Z.; Wang, T.; Pan, L.; Lin, D.; and Liu, Z. 2024. 3DTopia: Large Text-to-3D Generation Model with Hybrid Diffusion Priors. *arXiv preprint arXiv:2403.02234*.
- Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 78723–78747.
- Jun, H.; and Nichol, A. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*.
- Kelly, C.; Hu, L.; Yang, B.; Tian, Y.; Yang, D.; Yang, C.; Huang, Z.; Li, Z.; Hu, J.; and Zou, Y. 2024. Visiongpt: Vision-language understanding agent using generalized multimodal framework. *arXiv preprint arXiv:2403.09027*.
- Lee, M.; and Seok, J. 2019. Controllable generative adversarial network. *Ieee Access*, 7: 28158–28169.
- Li, D.; Kamko, A.; Akhgari, E.; Sabet, A.; Xu, L.; and Doshi, S. 2024a. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation. *arXiv:2402.17245*.
- Li, H.; Chong, Y. Q.; Stepputtis, S.; Campbell, J.; Hughes, D.; Lewis, M.; and Sycara, K. 2023. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*.
- Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al.

- 2024b. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *arXiv preprint arXiv:2405.08748*.
- Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Marius Kalytis, D. L. 2011. CGTrader. <https://www.cgtrader.com>. Accessed: 2024-06-30.
- Midjourney. 2024. Midjourney v6.0. <https://www.midjourney.com>. Image generation AI.
- Ouzma, M.; Ulrich, M.; and Yang, B. 2017. A simple radar simulation tool for 3D objects based on blender. In *2017 18th International Radar Symposium (IRS)*, 1–10. IEEE.
- Palanisamy, P. 2020. Multi-agent connected autonomous driving using deep reinforcement learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.
- Pharr, M.; Jakob, W.; and Humphreys, G. 2023. *Physically based rendering: From theory to implementation*. MIT Press.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv:2307.01952*.
- Qian, C.; Cong, X.; Yang, C.; Chen, W.; Su, Y.; Xu, J.; Liu, Z.; and Sun, M. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.
- Qin, J.; Wu, J.; Chen, W.; Ren, Y.; Li, H.; Wu, H.; Xiao, X.; Wang, R.; and Wen, S. 2024. Diffusiongpt: LLM-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*.
- Qu, L.; Wu, S.; Fei, H.; Nie, L.; and Chua, T.-S. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 643–654.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Z.; Xie, E.; Li, A.; Wang, Z.; Liu, X.; and Li, Z. 2024. Divide and Conquer: Language Models can Plan and Self-Correct for Compositional Text-to-Image Generation. *arXiv:2401.15688*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Wei, Y.; Wang, Z.; Lu, Y.; Xu, C.; Liu, C.; Zhao, H.; Chen, S.; and Wang, Y. 2024. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15077–15087.
- Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Zhang, S.; Zhu, E.; Li, B.; Jiang, L.; Zhang, X.; and Wang, C. 2023b. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.
- Wu, T.-H.; Lian, L.; Gonzalez, J. E.; Li, B.; and Darrell, T. 2024. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6327–6336.
- Yang, D.; Hu, L.; Tian, Y.; Li, Z.; Kelly, C.; Yang, B.; Yang, C.; and Zou, Y. 2024a. WorldGPT: a Sora-inspired video AI agent as Rich world models from text and image inputs. *arXiv preprint arXiv:2403.07944*.
- Yang, L.; Yu, Z.; Meng, C.; Xu, M.; Ermon, S.; and Cui, B. 2024b. Mastering Text-to-Image Diffusion: Recapitulating, Planning, and Generating with Multimodal LLMs. *arXiv:2401.11708*.
- Yuan, Z.; Chen, R.; Li, Z.; Jia, H.; He, L.; Wang, C.; and Sun, L. 2024. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*.
- Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv:2302.05543*.