



Follow-Your-Instruction: A Comprehensive MLLM Agent for World Data Synthesis

Kunyu Feng^{1*}, Yue Ma^{2*}, Xinhua Zhang^{3*}

Boshi Liu⁴, Yikuang Yuluo⁵, Yinhan Zhang¹, Runtao Liu², Hongyu Liu², Zhiyuan Qin⁶, Shanhui Mo,

Qifeng Chen^{2†}, Zeyu Wang^{1,2†}

¹HKUST(GZ), ²HKUST, ³Tsinghua University, ⁴Peking University,

⁵Chongqing University, ⁶Beijing Innovation Center of Humanoid Robotics

Abstract

*With the growing demands of AI-generated content (AIGC), the need for high-quality, diverse, and scalable data has become increasingly crucial. However, collecting large-scale real-world data remains costly and time-consuming, hindering the development of downstream applications. While some works attempt to collect task-specific data via a rendering process, most approaches still rely on manual scene construction, limiting their scalability and accuracy. To address these challenges, we propose **Follow-Your-Instruction**, a Multimodal Large Language Model (MLLM)-driven framework for automatically synthesizing high-quality 2D, 3D, and 4D data. Our **Follow-Your-Instruction** first collects assets and their associated descriptions through multimodal inputs using the MLLM-Collector. Then it constructs 3D layouts, and leverages Vision-Language Models (VLMs) for semantic refinement through multi-view scenes with the MLLM-Generator and MLLM-Optimizer, respectively. Finally, it uses MLLM-Planner to generate temporally coherent future frames. We evaluate the quality of the generated data through comprehensive experiments on the 2D, 3D, and 4D generative tasks. The results show that our synthetic data significantly boosts the performance of existing baseline models, demonstrating Follow-Your-Instruction’s potential as a scalable and effective data engine for generative intelligence.*

1. Introduction

AI-generated content (AIGC) targets to generate creative and realistic content using a generative model. It has been

widely applied in the film industry, augmented reality, automated advertising, and creating content for social media. Recent achievements in the foundation models, such as the diffusion models [42, 47], Multimodal Large Language Models (MLLMs) [1, 40] have significantly enhanced the quality and flexibility of generated content. As the data-driving model, these models acquire strong prior knowledge from the large-scale training datasets, which enables them to easily handle challenging tasks, including multi-modal understanding [29, 43, 56, 57] and generation [8, 19, 22, 30, 31, 36, 51, 70], visual editing [12, 24, 32, 52, 53, 60, 69], animation [35, 58, 59, 68], and embodied robots [25, 67].

However, as AIGC applications continue to evolve toward more complex and fine-grained scenarios, the demand for high-quality, task-specific data has significantly increased. While most open-source foundation models are trained on large-scale but generic datasets such as LAION-400M [48] and WebVid-10M [6], these datasets typically lack the task-specific annotations required for fine-grained applications. For instance, tasks such as object removal demand accurate background masks, while 4D generation requires accurate camera trajectories. This absence of precise supervision signals often limits the direct applicability of these datasets to specialized tasks [64].

Currently, there are some early works [23, 46] to build task-specific datasets, particularly through rendering pipelines. Rendering engines such as Blender [13] enable fine-grained control over object layout, lighting conditions, and physical interactions, making them suitable for constructing datasets tailored to specific AIGC tasks. These synthetic datasets are often used to fine-tune powerful foundation models for improved performance in downstream applications. However, manually designing and curating such datasets remains a significant bottleneck, as it requires substantial human effort, domain expertise, and often struggles to balance realism, accuracy, and scalability [33, 39].

Preprint, Under Review.

* Equal Contribution. † Corresponding Authors.



Fig. 1. **Overview of Follow-Your-Instruction.** We introduce Follow-Your-Instruction, an advanced MLLM-driven agent framework that synthesizes high-quality world data across 2D, 3D, and 4D levels, benefiting various downstream applications.

To address these limitations, we introduce our **Follow-Your-Instruction**, an efficient MLLM-based data-synthetic agent framework designed to generate realistic and diverse world data for a wide range of AIGC tasks. More importantly, to the best of our knowledge, our benchmark is the first data generation system that supports both 2D, 3D, and 4D generative tasks. As illustrated in Fig. 1, our agent encompasses seven representative applications, including 2D object removal, 3D restoration, inpainting, and 4D multi-view generation. In detail, by leveraging the extensive real-world understanding and interactive capabilities of MLLMs, we incorporate strong MLLMs into our agent and introduce four key components, including *MLLM-Collector*, *MLLM-Generator*, *MLLM-Optimizer*, and *MLLM-Planner*, to assist in the design and validation of our benchmark.

We mainly evaluate the performance of our proposed **Follow-Your-Instruction** in two scopes: (1) Evaluating the MLLM-Driven synthetic data quality: To benchmark the ability of MLLM-Driven synthesis, we perform the experiments on 8 MLLMs, including both commercial tools and research methods, on 4 metrics. (2) Evaluation on several Downstream Applications: To further evaluate the effectiveness of synthetic data, we finetune 3 various downstream tasks using our synthetic dataset, such as the 2D object removal task, 3D reconstruction, and 4D video generation. The results show substantial improvements in task-

specific performance, highlighting the practical benefits of our framework.

In summary, our contributions are as follows:

- We propose an efficient MLLM-based data-synthetic agent framework, Follow-Your-Instruction, which synthesizes realistic world data for diverse AIGC tasks.
- To achieve high-quality and efficient data generation, we introduce a comprehensive benchmark to evaluate MLLM-based data-synthetic agents at 2D, 3D, and 4D levels. Additionally, we develop various forms of MLLM-assisted data generation, including in-context and long-term guidance.
- To validate the practical performance of our proposed agent, we finetune 3 recent baseline models across representative 2D, 3D, and 4D tasks. Experimental results show that incorporating our data significantly enhances the performance of these models on their respective downstream applications.

2. Related Work

Multi-modal Large Language Models. Multi-modal Large Language Models (MLLMs) are advancing by integrating text, vision, and 3D modalities. In content restoration, RestoreAgent [7] shows strong performance in 2D tasks, while RL-Restore [65] focuses on progressive recovery for blur and noise. Clarity ChatGPT [54] combines con-

versation with restoration but remains narrow in scope. For spatial modeling, Text2World [14] and Spatial-MLLM [55] focus on symbolic structure generation and dual-encoder-based reasoning, respectively. VSI-Bench [62] benchmarks spatial reasoning tasks like counting and navigation. In embodied interaction, models like GEA show strong performance on VisualAgentBench [26], while Embodied-Bench [63] reveals limitations in long-term planning for models such as GPT-4V. Despite these advances, challenges like the lack of unified multimodal evaluation and training data remain.

Diffusion-based Generative Applications. Diffusion models are widely applied in generative tasks across 2D, 3D, and 4D domains. For 2D tasks like object removal and relighting, prior works [18, 20, 28, 37] rely on manually curated datasets and segmentation pipelines. In 3D, LiDAR Diffusion Models [45] reconstruct depth/point clouds using dedicated datasets, while MV-Adapter [16] ensures multi-view consistency via a plug-and-play module. For 4D, methods like ReCamMaster [4] and TrajectoryCrafter [66] leverage 3D structures to ensure cross-camera consistency in video generation. Follow-Your-Creation [34] explores 4D video editing frameworks. However, these methods require large-scale datasets, which are costly to obtain. Our Follow-Your-Instruction addresses this by using MLLMs to generate high-quality synthetic data, reducing real-world data dependency and enhancing adaptability.

3. Method

In this section, we introduce our proposed agent, a MLLM-based, comprehensive benchmark across 2D, 3D, and 4D levels. The pipeline of our agent is shown in Fig. 2, which is built upon the advanced multimodal Large Language Models (e.g., GPT-4o [40], QWEN3 [61]). From Sec. 3.1 to Sec. 3.3, we present the details of our proposed agent.

3.1. Assets Collection via Multimodal Inputs

Given a conditional input (e.g., an image I , a text T , or an action A), we aim to create high-quality scenes and keep both spatial and temporal consistency. Recent work like SceneCraft [15] designs the LLM-decomposer to generate a list of assets and the description of each sub-scene, then this information can be used for scene generation. Although this method provides a structured decomposition pipeline, it is inherently constrained by the limitations of the input. In particular, complex visual concepts and styles are often difficult to fully express through language alone, which in turn restricts the user’s ability to customize the generative scene.

Our proposed agent incorporates a multimodal asset retrieval mechanism that leverages MLLMs to incorporate both textual and visual information during the asset discovery process. As shown in Fig. 2, in addition to prompting with natural language, users can supply specific modalities

such as reference images or specific objects. These inputs offer greater flexibility and control, allowing users to specify creative intent in diverse ways. Specifically, we first utilize the MLLMs to transform the inputs into the assets list:

$$(d_1, A_1), \dots, (d_k, A_k) \leftarrow \text{MLLM-Collector}(I), \quad (1)$$

where the I is inputs, d_i is the sub-scene description, A_i is the list of assets. For text input, we follow the settings from the SceneCraft [15] and apply a top- k retrieval strategy based on relevance scores from our asset repository, selecting the most semantically matched assets for further composition. As for visual input, our agent skips the selection process and directly integrates the assets into the scene construction pipeline. This design not only improves the grounding and fidelity of the generated scenes but also greatly enhances controllability, allowing for more precise and expressive scene authoring. Therefore, our framework significantly enhances the customizable and user-friendly controllability of the scene creation process, making it more suitable for real-world content generation tasks.

3.2. Global Scenes Construction and Optimization

3D Layout Generation. After acquiring the assets from the multimodal inputs, the next step is to integrate all assets and create the entire scene. This process is managed by our *MLLM-Generator*, which performs object generation, spatial placement, and coordinate transformation. Specifically, given an asset A_i and its description D_i , it first generates the objects and bounding box in 3D space:

$$B_i = \text{MLLM-}Creator(A_i, D_i), \quad (2)$$

where $B_i \in R^6$ represents the 3D bounding box parameters of object i (e.g., center position, width, height, depth). These parameters are adapted to ensure consistency between appearance and semantics.

Then each object is assigned a 3D location p_i^{world} through the *MLLM-Locator*, which is in the global layout matrix L_{world} . In detail, there are two placement strategies, one of which is **Human-Instruction-Guided Placement**, which places the object through the specific location in the input instruction as follows:

$$p_i^{world} = p_i^{target}, \quad (3)$$

The other is the **default strategy** that the object is placed by aligning its bottom-center point c_i to a suitable unoccupied region in the world matrix:

$$p_i^{world} = \text{FindFreeRegion}(L_{world}, c_i), \quad (4)$$

Then, we apply a transformation matrix to embed the object into the global layout:

$$L_{world}[i] = \text{ComposeTransform}(p_i^{world}, \mathbf{R}_i, \mathbf{s}_i), \quad (5)$$

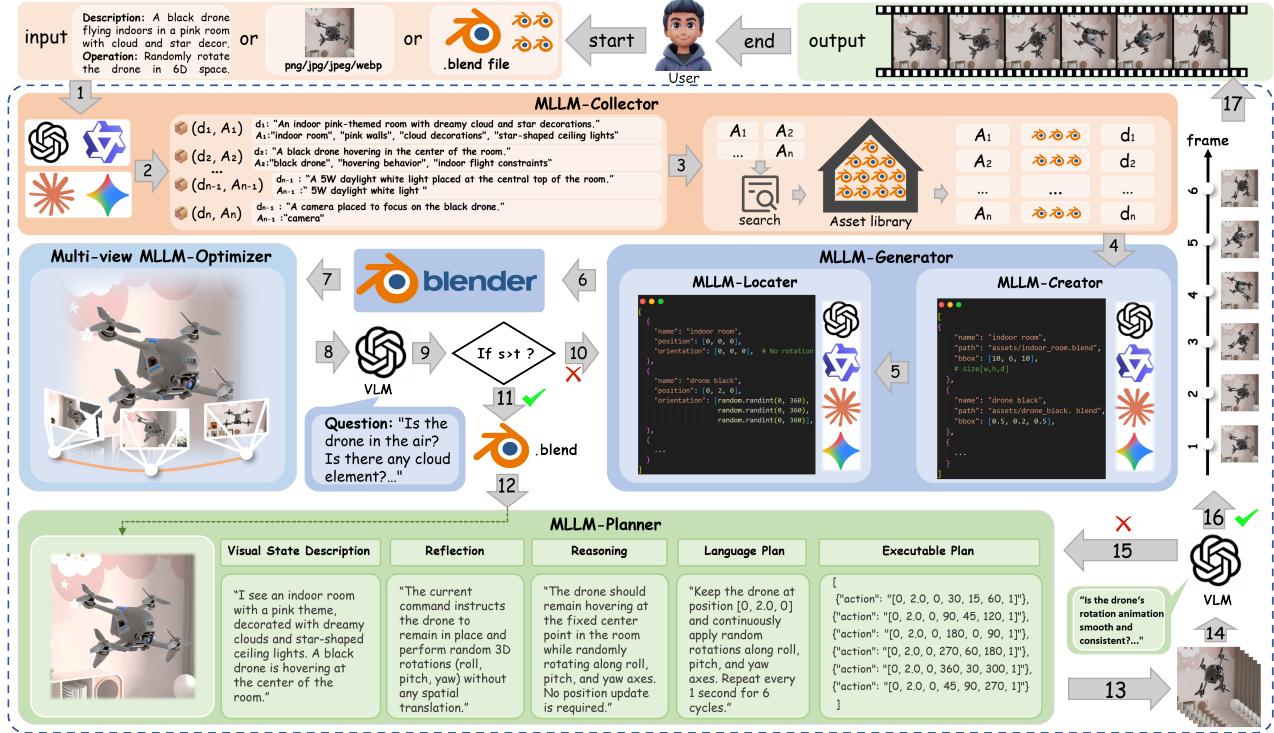


Fig. 2. **The Pipeline of Follow-Your-Instruction.** Given multimodal inputs, Follow-Your-Instruction first collects the assets and their descriptions via *MLLM-Collector*. Then the *MLLM-Generator* creates the 3D layout scene and optimizes the scene via multi-view *MLLM-Optimizer* with a powerful VLM. Based on the scene, *MLLM-Planner* formulates a clear plan to generate the high-quality output video.

where \mathbf{R}_i and \mathbf{s}_i are the estimated rotation and scale matrices respectively.

Finally, the *MLLM-Locator* projects the 3D layout into the 2D image plane by the given intrinsic matrix \mathbf{K} and extrinsic pose \mathbf{E} of calibrated camera:

$$u_i = \pi(\mathbf{K}, \mathbf{E}, p_i^{world}), \quad (6)$$

where $\pi(\cdot)$ denotes the perspective projection function, and $u_i \in R^2$ is the image coordinate of object i .

Multi-View Optimization. Despite constructing the whole scenes with the multimodal inputs, there still exists some mismatch in the whole layout. In the prior work [15], they adopt an iterative visual feedback loop to refine scene layouts by leveraging an MLLM. However, such an approach based on a single rendered view is often insufficient, especially when dealing with physical interactions between objects. For instance, as shown in Fig. 3, the input text condition is "Place two cups on a table.". If we only optimize the generated scene from a single viewpoint, it might only adjust the pink cup to a correct position in this view, while from another angle, the pink cup could still be floating above the table (as shown in Fig. 3 (a)). This discrepancy arises because the MLLM is unaware of depth inconsistencies hidden from the current view.

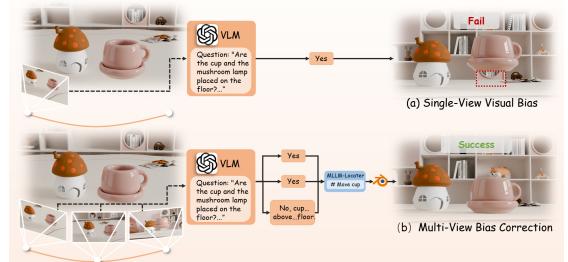


Fig. 3. **Motivation for Multi-view Optimization.** (a): Optimizing the constructed scene from a single view may yield satisfactory results in that specific view, but object placements often exhibit semantic misalignments when observed from other perspectives. (b): Our proposed multi-view optimization effectively mitigates such inconsistencies by improving semantic correctness across multiple viewpoints, leading to globally coherent scene layouts.

To ensure more reliable and physically grounded scene layouts, it is crucial to incorporate multi-view renderings during the feedback process. Multiple views enable the model to better verify spatial relationships, reducing visual illusions caused by limited viewpoints and producing more robust layouts. In our agent, we introduce a multi-view feedback optimization strategy, guided by a powerful

Vision-Language Model (VLM). Our agent renders the current scene L from multiple views $N = \{v_1, v_2, \dots, v_n\}$, and interacts with the VLM (e.g. *"Is the pink cup placed on the table?"*) to verify the spatial relations for each view:

$$S_{VLM} = \frac{1}{n} \sum_{i=1}^n s_i, \quad (7)$$

where the S is the confidence score feedback by VLM, if the scores exceed the threshold t , our agent determines that the current scene optimization is successful. Otherwise, the system creates a new location of object refinement via the *MLLM-Locator*. The effectiveness of this strategy is illustrated in Fig. 3 (b), where a pink cup is initially misaligned from a side view, but after VLM-guided correction, it is accurately placed across all perspectives.

3.3. MLLM-Guided Task Planner

Even though the 2D images dataset generated by the *MLLM-Optimizer* is sufficient for some simple tasks, such as 2D object removal, relighting, and inpainting. We aim to synthesize a high-quality video dataset for various practical applications. Equipped with the powerful ability of MLLM, e.g., in-context learning, long-term learning, we introduce the *MLLM-Planner* for video generation.

As shown in Fig. 2, our *MLLM-Planner* receives both the human instructions and the generated scenes as the input information. It first understands the visual scene and creates the visual state descriptions, locating the main object of the current frame. Then it reflects the human instructions and the feedback from the VLM-guided optimizer to refine the actions, facilitating reasoning of the accuracy target objectives. Based on the reasoning results, it formulates the language plan and then converts it to an executable plan for generating the subsequence frames.

However, a common issue remains that of temporal inconsistency across consecutive frames. This stems from the *MLLM-Planner*, which focuses on discrete action execution without ensuring smooth transitions. As a result, generated sequences may suffer from abrupt changes, unnatural motion, or missing intermediate states. To address this, we introduce a VLM-guided frame prediction module (step 14 in Fig. 2), which leverages VLM’s visual reasoning to assess motion, object states, and scene dynamics across frames. Upon detecting inconsistencies, the module provides feedback to the *MLLM-Planner*, prompting it to refine actions or insert intermediate steps. This iterative process enhances both temporal coherence and video quality.

4. Experiments

4.1. Evaluate the Quality of Generative Scenes

Experimental Setup. Most existing Multimodal Large Language Models (MLLMs) have shown promising capa-

bilities in both vision and language understanding. Following the recent work [63], the primary baselines we compare against are state-of-the-art MLLMs, which can be categorized into closed-source proprietary models and open-sourced models, as they represent the current frontier of multimodal reasoning and decision-making.

Closed-source models include GPT-4o and GPT-4o-mini[40, 41], Claude-3.5-Sonnet and Claude-4-Sonnet [1, 2], Gemini-2.5-Pro, Gemini-2.0-flash [10, 11], and Qwen-VL-Max [3]. These models are known for their strong performance in general multimodal tasks, with advanced reasoning abilities and extensive training on diverse internet-scale data. Open-sourced models, such as Llama-3.2 Vision Instruct [38], InternVL2.5, InternVL3 [9, 71], Qwen3, Qwen2.5-VL [5, 61], Gemma-3 [50], and Ovis2 [27], cover a range of model sizes (7B to 90B parameters) and provide accessible alternatives for research, allowing for deeper analysis of architectural design and scaling effects.

Experimental Results. Tab. 1 demonstrates a quantitative comparison of different MLLMs applied to our data-synthetic agents. We use the aesthetic score[49] to assess perceptual quality, and measure scene consistency follows the VBench [17] in terms of subject appearance and background stability. Text alignment is evaluated via CLIP similarity [44]. Results demonstrate the vital role of MLLM guidance in Follow-Your-Instruction, with GPT-4o achieving the best performance across all metrics, highlighting its superior cross-modal reasoning and alignment capabilities. Claude-4-Sonnet and Claude-3.7-Sonnet follow closely in aesthetics and consistency but lag in alignment. Among open-sourced models, InternVL3-78B and Qwen3-235B-A22B-Ins perform best overall, though a notable gap remains compared to GPT-4o. Noteably, this experiment is designed to highlight the generality of Follow-Your-Instruction’s core MLLM-driven capability across diverse AIGC tasks and MLLM structures, rather than focusing on achieving peak performance with any single MLLM. User study is provided in the supplementary materials.

Applications. As shown in Fig. 4, we illustrate several representative tasks along with the corresponding ground-truth annotations generated by our agent. These examples highlight the agent’s capacity to generalize across environments and task objectives. Our proposed agent’s applications across a diverse set of tasks spanning 2D (object removal and re-lighting), 3D (reconstruction, rotation, and embodied intelligence), and 4D environments (4D inpainting and reconstruction). These tasks reflect the **Follow-Your-Instruction**’s potential for content creation in emerging research areas.

4.2. Evaluation on Downstream Application

Baselines. To further evaluate the quality of the synthetic data, we fine-tune several baseline models on both

Method	Scene Quality		Scene Consistency		Text Align
	Aesthetic \uparrow	Subject Consis. \uparrow	Background Consis. \uparrow	CLIP Sim \uparrow	
<i>Proprietary MLLMs</i>					
Claude-4-Sonnet [2]	6.574	0.9135	0.9078	68.03	
Claude-3.7-Sonnet [1]	6.572	0.9122	0.9079	67.33	
Gemini-2.0-flash [11]	6.472	0.9033	0.8925	65.48	
Gemini-2.5-Pro [10]	6.514	0.9138	0.9073	67.52	
Qwen-VL-Max [3]	6.408	0.9131	0.9065	67.89	
GPT-4o [40]	6.592	0.9144	0.9087	68.75	
GPT-4o-mini [41]	6.137	0.9038	0.8977	63.49	
<i>Open-Source MLLMs</i>					
Llama-3.2-90B-Vision-Ins [38]	6.357	0.9124	0.9057	66.35	
Llama-3.2-11B-Vision-Ins [38]	6.324	0.9107	0.9031	65.17	
InternVL2.5-78B [9]	6.209	0.9115	0.9062	65.23	
InternVL3-78B [71]	6.213	0.9127	0.9075	67.18	
Qwen2.5-VL-72B-Ins [5]	6.312	0.9087	0.9011	65.47	
Qwen3-235B-A22B-Ins [61]	6.379	0.9122	0.9043	66.12	
Ovis2-34B [27]	6.255	0.9081	0.9002	64.35	
Ovis2-16B [27]	6.231	0.9075	0.8971	62.15	
gemma-3-27b-it [50]	6.204	0.9053	0.8953	60.35	
gemma-3-12b-it [50]	6.197	0.9042	0.8891	59.67	

Tab. 1. **Comparison with different MLLMs methods.** We generate 50 videos guided by each MLLM, and evaluate the performance of this data. The best result is in **Red**, and the second best result is in **Blue**.

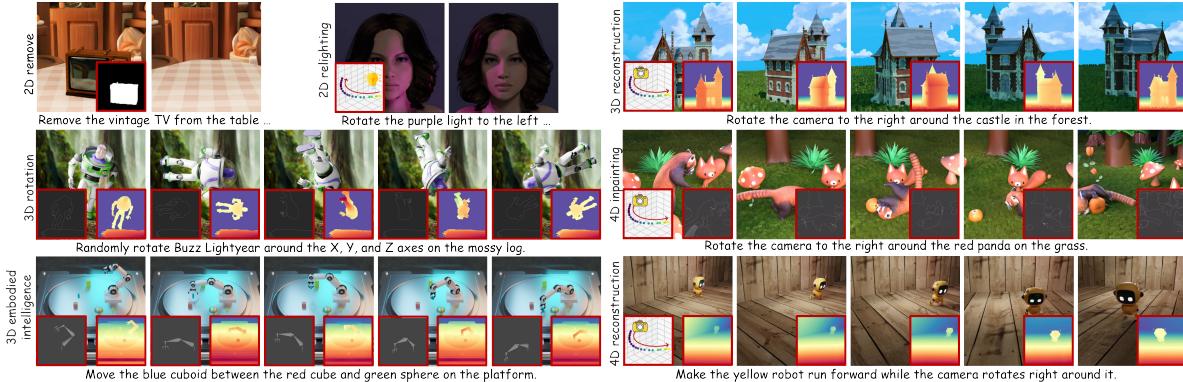


Fig. 4. **Diverse downstream applications supported by Follow-Your-Instruction.** Each task is accompanied by tailored annotations, such as background masks for object removal, camera trajectories for relighting, 3D and 4D reconstruction, as well as depth maps and object poses for 3D embodied intelligence.

2D/3D/4D AIGC applications, including object removal, 3D reconstruction, and 4D video generation. For the 2D object removal task, we adopt RoRem [21] as the baseline and assess the improvement after fine-tuning with our data. For the 3D reconstruction task, we utilize MV-Adapter [16], a recent multi-view reconstruction framework, and evaluate performance improvements in terms of geometry accuracy and consistency. For the 4D video generation task, we em-

ploy ReCamMaster [4] to measure temporal coherence and fidelity in dynamic scene synthesis. These baselines allow us to systematically quantify the impact of our synthetic data on different aspects of AIGC models across various dimensions.

Qualitative Results. The visual comparison for the 2D, 3D, and 4D applications is shown in Fig. 5. We can see that before fine-tuning with our generated data, the performance

Method	Visual Quality				Camera Accuracy		View Synchronization		
	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	RotErr ↓	TransErr ↓	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
ReCamMaster [4]	62.48	160.72	34.97	96.23	1.45	5.22	630.51	151.28	88.59
+ Our data	60.32	155.71	35.88	96.66	1.35	4.69	682.57	135.24	89.92

Tab. 2. **Quantitative results for 4D generation tasks.** We assess visual quality, camera accuracy, and view synchronization. The best results are in bold.

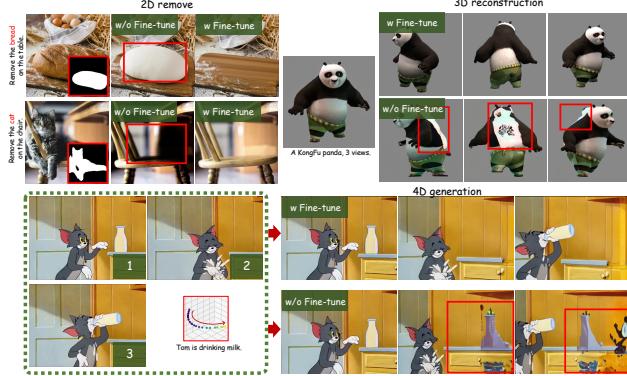


Fig. 5. **Qualitative results for 2D object removal, 3D reconstruction, and 4D generation applications.** The results show that our generated data exhibits better effectiveness for improving the performance of existing models.

of object removal is struggled with the semantic completion (as shown in the first row for the 2D task in Fig. 5, the model generate a strange white object rather than inpaint the cutting board), and generates some artifacts after removing (the second row for the 2D task in Fig. 5). In contrast, after fine-tuning with our generated data, these problems have been alleviated significantly. On the other hand, for the 3D task, without fine-tuning with our data, although the model can generate front views well, it struggles to generate high-quality and consistent back views, and the hallucination issue can be fixed after the fine-tuning process.

Additionally, the 4D generation task is an emerging paradigm of controllable video synthesis under the guidance of camera trajectories. As shown in Fig. 5, while the ReCamMaster achieves better pose accuracy and smooth camera movement, there are still some inconsistencies and artifacts in the background, and our generated data has improved its performance.

Quantitative Results. We also perform the quantitative experiments for the three applications. The quantitative results of 2D object removal and 3D reconstruction can be found in the supplementary material, and the results of 4D generation are shown in Tab. 2. Following the ReCamMaster [4], we evaluate the visual quality, camera accuracy, and the

view synchronization. Specifically, we calculate the rotation and translation errors to evaluate the camera trajectory accuracy, and compute the CLIP-V and FVD-V to assess the view synchronization between the different viewpoints in the same scene. The results show that the performance of the baseline model can be improved after fine-tuning.

4.3. Ablation Study

Effectiveness of the Multi-view optimization. As shown in Fig. 6, we evaluate the performance of different numbers of frames in the Multi-view optimization strategy. When we use only one view to optimize the generated scene, although in the current view the object has been located in the correct position, it often fails to align properly in other unseen perspectives. This limitation can be mitigated by incorporating more views during optimization. Furthermore, we also conduct a quantitative ablation study to assess the proper number of views we need to provide in this process, and the results are shown in the Tab. 3. We can observe that increasing the number of views leads to longer generation times, while offering only marginal improvements in optimization success rate. Based on this analysis, we select two views as the optimal setting, balancing both efficiency and performance.

Num of Views	CLIP Sim ↑	Success Rate ↑	Time (s)
1	53.24	0.2415	380
2	68.75	0.9987	386
3	68.63	0.9994	398

Tab. 3. **Quantitative ablation results for Multi-view optimization (step 7 in Fig. 2).** The results demonstrate that with the increase of views, the quality of scenes exhibits a trend of first improving and then keeping the same level, while an increase in time has raised the cost. A view count of 2 offers the best balance.

Effectiveness of VLM-guided frame prediction. We also evaluate the contribution of the VLM-guided frame prediction in Fig. 7 and Tab. 4. In the second row of Fig. 7, it can be observed that without the VLM-guided frame prediction strategy, the resulting video often suffers from temporal inconsistency, where the motion between adjacent frames is abrupt and unsmooth. Specifically, the rotation angle of

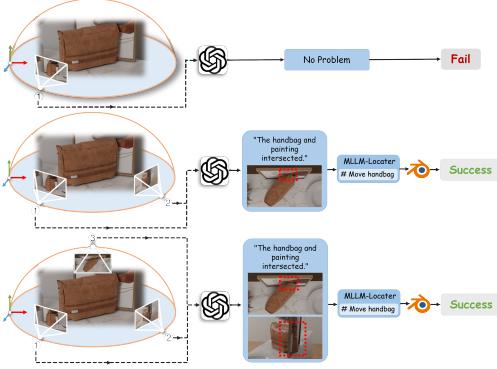


Fig. 6. **Ablation study about the Multi-view optimization.** The global scene still has misalignment, only optimizing with a single view. (first row). And this issue can be addressed by increasing the number of reference views (second and third rows).

Method	CLIP Sim \uparrow	Temporal Consis. \uparrow
w/o VLM	50.76	0.6524
ours	68.75	0.9128

Tab. 4. **Ablation results for VLM-guided frame prediction (step 14 in Fig. 2).** The best results are in bold.

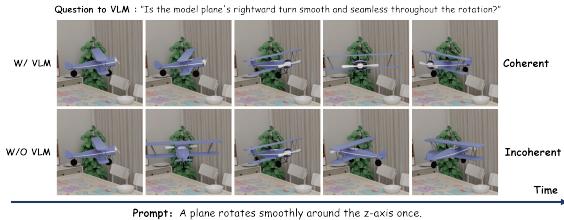


Fig. 7. **Ablation study about the VLM-guided frame prediction.** The second row shows that without the VLM optimization, the plane exhibits abrupt, unsMOOTH motion with rapid successive turns.

the airplane is excessively large between two consecutive frames, making it appear as if the plane turns twice within a short duration. This indicates that the planned actions lack continuity, resulting in suboptimal visual quality and temporal inconsistency.

5. Conclusion and Discussion

Conclusion. This paper has proposed the **Follow-Your-Instruction**, an efficient MLLM-based data-synthetic agent framework, which generates the realistic scenes across 2D, 3D, and 4D levels from the multimodal inputs (e.g., text, image, or the blend file). **Follow-Your-Instruction** is built on the Multimodal Large Language Models, combining with four main components: *MLLM-Collector*, *MLLM-Generator*, *MLLM-Optimizer*, and *MLLM-Planner*. First,

the *MLLM-Collector* transforms the text input into the assets or integrates the assets from the visual inputs, enhancing the user-oriented scene creation. Then the *MLLM-Generator* creates the 3D layout for the scene and optimizes it via *MLLM-Optimizer*. Finally, the *MLLM-Planner* creates the subsequent frames and refines them by the VLM-guided frame prediction module. Experimental results demonstrate that our agent leverages the capability of the MLLMs in our data synthesis process, facilitating several downstream AIGC applications.

Limitations. There are three limitations of our method: (1) the performance relies on the capabilities of its underlying proprietary MLLM; (2) the lack of validation of generated data in improving generalization to other real-world benchmarks; (3) the scalability is constrained by its reliance on pre-existing asset libraries.

References

- [1] Anthropic. Claude 3.5 sonnet, 2024. [1](#), [5](#), [6](#)
- [2] Anthropic. Claude 4 sonnet, 2025. [5](#), [6](#)
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. [5](#), [6](#)
- [4] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. [3](#), [6](#), [7](#)
- [5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. [5](#), [6](#)
- [6] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. [1](#)
- [7] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 110643–110666, 2024. [2](#)
- [8] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024. [1](#)
- [9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. [5](#), [6](#)

- [10] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blissein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5, 6
- [11] Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, 2024. 5, 6
- [12] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 1
- [13] Blender Foundation. Blender. <https://www.blender.org/>, 2024. 1
- [14] Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Qiwei Liang, Yao Mu, Hongyuan Zhang, Wenqi Shao, et al. Text2world: Benchmarking large language models for symbolic world model generation. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*. 3
- [15] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An LLM agent for synthesizing 3d scenes as blender code. In *Forty-first International Conference on Machine Learning*, 2024. 3, 4
- [16] Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024. 3, 6
- [17] Ziqi Huang, Yinan He, Jiahuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5
- [18] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24452–24462, 2025. 3
- [19] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 1
- [20] Jingzhi Li, Zongwei Wu, Eduard Zamfir, and Radu Timofte. Recap: Better gaussian relighting with cross-environment captures. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21307–21316, 2025. 3
- [21] Ruibin Li, Tao Yang, Song Guo, and Lei Zhang. Rorem: Training a robust object remover with human-in-the-loop. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14024–14035, 2025. 6
- [22] Zhichao Liao, Fengyuan Piao, Di Huang, Xinghui Li, Yue Ma, Pingfa Feng, Heming Fang, and Long Zeng. Freehand sketch generation from mechanical components. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 6755–6764, 2024. 1
- [23] Thomas Lips et al. Evaluating text-to-image diffusion models for texturing synthetic data. *arXiv preprint arXiv:2411.10164*, 2024. 1
- [24] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 1
- [25] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuna Yang, et al. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*, . 1
- [26] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Song XiXuan, Yifan Xu, Shudan Zhang, Hanyu Lai, Jiadai Sun, Xinyue Yang, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. In *The Thirteenth International Conference on Learning Representations*, . 3
- [27] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 5, 6
- [28] Li Lun, Kunyu Feng, Qinglong Ni, Ling Liang, Yuan Wang, Ying Li, Dunshan Yu, and Xiaoxin Cui. Towards effective and sparse adversarial attack on spiking neural networks via breaking invisible surrogate gradients. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3540–3551, 2025. 3
- [29] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4132–4141, 2022. 1
- [30] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 1
- [31] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 1
- [32] Yue Ma, Xiaodong Cun, Sen Liang, Jinbo Xing, Yingqing He, Chenyang Qi, Siran Chen, and Qifeng Chen. Magic-stick: Controllable video editing via control handle transformations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9385–9395. IEEE, 2025. 1
- [33] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025. 1
- [34] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhua Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025. 3

- [35] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025. 1
- [36] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 1
- [37] Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levant Taşel, Ning Yu, et al. Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5510–5522, 2025. 3
- [38] Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, 2024. 5, 6
- [39] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerar. A survey of synthetic data augmentation methods in machine vision. *Machine Intelligence Research*, 21(5):831–869, 2024. 1
- [40] OpenAI. Hello gpt-4o, 2024. 1, 3, 5, 6
- [41] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. 5, 6
- [42] 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. 1
- [43] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multi-modal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025. 1
- [44] 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. 5
- [45] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14738–14748, 2024. 3
- [46] Parth Rawal, Mrunal Sompura, and Wolfgang Hintze. Synthetic data generation for bridging sim2real gap in a production environment. *arXiv preprint arXiv:2311.11039*, 2023. 1
- [47] 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
- [48] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1
- [49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 5
- [50] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 5, 6
- [51] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1
- [52] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. In *Forty-second International Conference on Machine Learning*. 1
- [53] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *Advances in Neural Information Processing Systems*, 37:96541–96565, 2024. 1
- [54] Yanyan Wei, Zhao Zhang, Jiahuan Ren, Xiaogang Xu, Richang Hong, Yi Yang, Shuicheng Yan, and Meng Wang. Clarity chatgpt: An interactive and adaptive processing system for image restoration and enhancement. *arXiv preprint arXiv:2311.11695*, 2023. 2
- [55] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 3
- [56] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18709–18719, 2024. 1
- [57] 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. In *The Thirteenth International Conference on Learning Representations*. 1
- [58] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Towards multiple character image animation through enhancing implicit decoupling. In *The Thirteenth International Conference on Learning Representations*. 1
- [59] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Follow-your-pose

- v2: Multiple-condition guided character image animation for stable pose control. *arXiv e-prints*, pages arXiv–2406, 2024. 1
- [60] Zexuan Yan, Yue Ma, Chang Zou, Wenteng Chen, Qifeng Chen, and Linfeng Zhang. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. *arXiv preprint arXiv:2503.10270*, 2025. 1
- [61] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Bin Yuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengu Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yingqi Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. 3, 5, 6
- [62] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 3
- [63] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Forty-second International Conference on Machine Learning*. 3, 5
- [64] Ryota Yoshihashi, Yuya Otsuka, Tomohiro Tanaka, Hirokatsu Kataoka, et al. Exploring limits of diffusion-synthetic training with weakly supervised semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 2300–2318, 2024. 1
- [65] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2443–2452, 2018. 2
- [66] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 3
- [67] Junpeng Yue, Xinrun Xu, Börje F Karlsson, and Zongqing Lu. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. In *The Thirteenth International Conference on Learning Representations*. 1
- [68] Yinhan Zhang, Yue Ma, Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Magiccolor: Multi-instance sketch colorization. *arXiv preprint arXiv:2503.16948*, 2025. 1
- [69] Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. Instantswap: Fast customized concept swapping across sharp shape differences. In *The Thirteenth International Conference on Learning Representations*. 1
- [70] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10923–10931, 2025. 1
- [71] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 5, 6