GENMAC: Compositional Text-to-Video Generation with Multi-Agent Collaboration

Kaiyi Huang[1], Yukun Huang[1], Xuefei Ning[2], Zinan Lin[3], Yu Wang[2], Xihui Liu[1]*

[1]The University of Hong Kong
[2]Tsinghua University
[3]Microsoft Research
huangky@connect.hku.hk

Abstract

Text-to-video generation models have shown significant progress in recent years. However, they still struggle with compositional text prompts, such as attribute binding for multiple objects, temporal dynamics associated with different objects, and interactions between objects. Inspired by effective human creative workflow, we propose GENMAC, a multi-agent collaboration framework that enables compositional text-to-video generation. The framework incorporates a three-stage collaborative workflow: DESIGN, GENERATION, and REDESIGN, with an iterative loop between the latter two stages to progressively verify and refine the generated videos. In the DESIGN stage, a large language model (Design Agent) plans objects with layouts, and then a video generation model synthesizes videos in the GENERATION stage. The REDESIGN stage is the most challenging stage that aims to verify the generated videos, suggest corrections, and redesign the text prompts, frame-wise layouts, and guidance scales for the next iteration of generation. To avoid hallucination of single-agent and naive multi-agent frameworks, we apply a division-of-labor strategy in this stage by introducing a sequence of specialized agents, executed by MLLMs (multimodal large language models): Verification Agent, Suggestion Agent, Correction Agent, and Output Structuring Agent. Furthermore, to tackle diverse scenarios of compositional text-to-video generation, we design a self-routing mechanism to adaptively select the proper correction agent from a suite of correction agents, each specialized for one scenario. Extensive experiments demonstrate the effectiveness of GENMAC by generating videos based on long compositional text prompts and achieving state-of-the-art in the compositional text-to-video generation benchmark.

Introduction

With the rapid development of diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021; Sohl-Dickstein et al. 2015), text-to-video generation (Blattmann et al. 2023b,a; Ho et al. 2022; Singer et al. 2022; Wu et al. 2021, 2022; Hong et al. 2022; Villegas et al. 2022; Zhou et al. 2022; Khachatryan et al. 2023; Luo et al. 2023; He et al. 2022; Wang et al. 2023b) has achieved impressive advancements in creating compelling visual content. However, a critical challenge remains in **compositional T2V generation** (Sun et al. 2024a): accurately generating videos that follow intricate instructions, such as binding attributes to multiple objects, portraying specific interactions, and maintaining temporal consistency or dynamic changes, is a significant hurdle for current models. This limitation hinders their practical application in scenarios that demand precise control and faithful adherence to user intent.

Unfortunately, existing techniques fall short in meeting the demands of compositional T2V generation. On one hand, *single-pass* approaches (Tian et al. 2024; Yang and Wang 2024) attempt to synthesize the entire video in one go. This single-pass paradigm struggles with complex prompts, often failing to capture critical contextual details, as it lacks a mechanism for refinement. On the other hand, iterative refinement has been explored in text-to-image generation using a *single-agent* approach, where a Multimodal Large Language Model (MLLM) plans or corrects the output (Wang et al. 2024b; Wu et al. 2023c; Yang et al. 2024a). However, directly extending this single-agent model to compositional video generation is problematic for two key reasons. First, the spatio-temporal complexity inherent in compositional video overwhelms a single MLLM, which must simultaneously verify content with multiple objects, reason about dynamics, and plan corrections across multiple frames, often leading to severe hallucination (Figure 5). Second, a single agent lacks the specialized expertise required for diverse compositional tasks; for instance, ensuring temporal consistency ("a red ball on the left") versus managing dynamic motion ("a red ball moving to the left") are distinct challenges that a single agent cannot effectively handle (Table 2).

The limitations of single-pass and single-agent systems naturally suggest an iterative multi-agent solution. However, we argue that a *naive multi-agent* approach, simply employing multiple agents without a structured collaboration strategy, is also insufficient. As our ablation studies demonstrate (Table 2 and Figure 5), an unstructured group of agents is prone to hallucination and often fails to converge on an effective correction. They struggle to manage the complex interplay of object attributes, spatial relationships, and temporal dynamics. True progress requires not just more agents, but smarter collaboration.

To this end, we draw inspiration from effective human creative workflows. Humans tackle complex projects not

*Corresponding author.

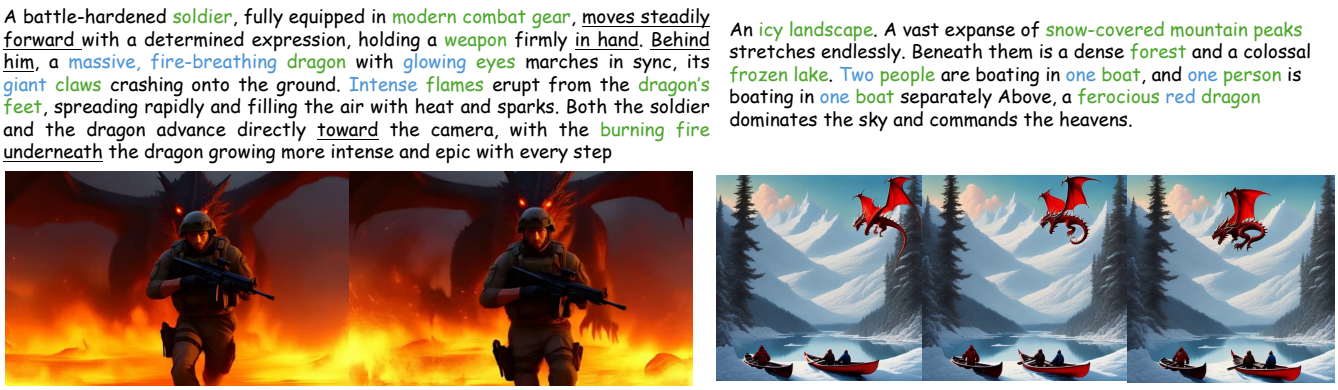

Figure 1: Videos generated by GENMAC based on **complex compositional prompts** involving **multiple objects, attribute binding, quantity, and dynamic motion binding**. Left: Hunyuan backbone. Right: VideoCrafter2 backbone.

with an unstructured mob, but with a team of specialists operating within a structured, iterative process. Each member has a distinct role, and the workflow ensures that their contributions build upon one another logically. Inspired by this principle of structured task decomposition and role specialization, we propose GENMAC, a novel multi-agent collaboration framework designed specifically for the challenges of compositional T2V generation.

Our framework follows these principles through a cohesive, hierarchical structure, as shown in Figure 2. (1) **Collaborative Workflow**: We decompose the complex generation task into an iterative loop of three stages: DESIGN, GENERATION, and REDESIGN. This structured process provides the iterative refinement lacking in single-pass methods. (2) **Division-of-Labor Strategy** for REDESIGN Stage: The REDESIGN stage itself is a complex reasoning process, which requires accurate understanding of video contents, semantic reasoning of spatial-temporal dynamics, and planning for the correction and refinement in the next generation iteration. To prevent the hallucination and chaos of single-agent or naive multi-agent systems, we further apply a division-of-labor strategy (Strauss 1985): we introduce a sequence of specialized agents to ensure participants may specialize in their tasks, *i.e.* a Verification Agent checks for misalignments, a Suggestion Agent proposes corrections, a Correction Agent implements them, and an Output Structuring Agent for structured output. This division-of-labor strategy ensures each agent handles a focused task, leading to reliable and coherent outcomes. (3) **Adaptive self-routing** for Correction Agents: to handle the complex scenarios of generated videos, text prompts, and refinement needs, we design a suite of specialized correction agents for correcting the designs from the perspectives of consistency, temporal dynamics, and spatial dynamics, respectively, which are critical parts in compositional T2V generation (Sun et al. 2024a). A self-routing mechanism is introduced to adaptively select the suitable agent for the current scenario.

In our multi-agent collaboration framework, the GENERATION stage is implemented by an off-the-shelf video generation model, compatible with both UNet- and Transformer-based diffusion models, while other agents are implemented by off-the-shelf LLM/MLLMs. The specific function of each agent is defined by its role and position within our structured workflow, rather than by fine-tuning the underlying models.

To the best of our knowledge, we are the first to tackle compositional text-to-video generation through a structured multi-agent framework. Our contributions are as follows:

- We propose GENMAC, a novel multi-agent collaborative framework that operationalizes a human-like iterative workflow, demonstrating its superiority over single-pass generation methods.
- We introduce a division-of-labor strategy that decomposes complex compositional T2V redesign process into sequential tasks. This structured approach effectively mitigates issues like hallucination, which plague both single-agent and naive multi-agent systems.
- We design a self-routing mechanism that adaptively selects expert agents for targeted corrections, includes consistency, temporal dynamics, and spatial dynamics, enabling a level of precision and flexibility that single-agent systems cannot achieve.
- Extensive experiments on the T2V-CompBench benchmark show that GENMAC achieves state-of-the-art performance on compositional T2V generation.

Related Work

Text-to-Video Generation Models. Text-to-video generation (Ho et al. 2022; Singer et al. 2022; Zhou et al. 2022; Khachatryan et al. 2023; Luo et al. 2023; Blattmann et al. 2023b; He et al. 2022; Wang et al. 2023b; Guo et al. 2023; Wan et al. 2025) has seen advancements with the development of diffusion models (Ho, Jain, and Abbeel 2020). More recently, language model-based methods (Villegas et al. 2022; Chang et al. 2023; Kondratyuk et al. 2023; Yu et al. 2023a,b; Chang et al. 2022) have enabled large-scale training, leading to significant improvements in generating high-quality videos.

Compositional Text-to-Video Generation. There have been studies on compositional text-to-image generation (Liu et al. 2022; Feng et al. 2023; Li et al. 2022; Wu et al. 2023b; Huang et al. 2024b; Patel et al. 2023; Liu et al. 2024; Chefer

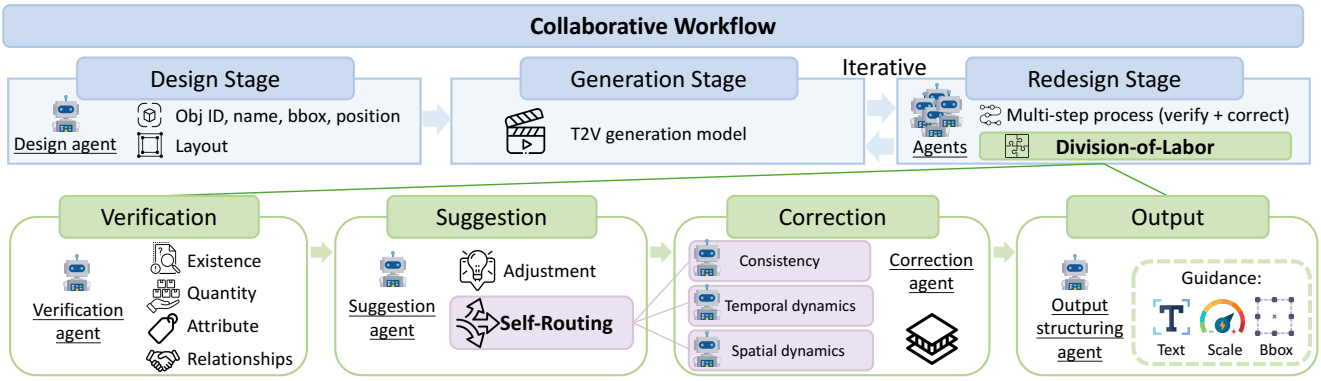

Figure 2: **Multi-Agent Collaboration Framework** of GENMAC. Collaborative workflow includes DESIGN, GENERATION, and REDESIGN stages with an iterative loop. Division-of-labor decomposes the REDESIGN stage into four sequential tasks, handled by four agents: Verification, Suggestion, Correction, and Output Structuring Agent. Self-routing mechanism allows for adaptive selection of suitable Correction Agent to address the diverse requirements for compositional text-to-video generation.

et al. 2023; Park et al. 2021; Lian et al. 2023; Chen, Laina, and Vedaldi 2024; Wu et al. 2023b; Wang et al. 2023c; Meral et al. 2023; Kim et al. 2023; Rassin et al. 2024; Gani et al. 2023; Li et al. 2023; Taghipour et al. 2024; Wang et al. 2024c; Chen et al. 2023; Yang et al. 2024a; Park et al. 2024). T2I-CompBench(++) (Huang et al. 2024b, 2025) introduces the first comprehensive benchmark in evaluating compositionality in text-to-image generation models, with attribute binding, relationships, and complex compositions. T2V-CompBench (Sun et al. 2024a) extends the compositional evaluation to text-to-video generation with the consideration of temporal dimensions. VideoTetris (Tian et al. 2024) proposes a framework of spatio-temporal compositional diffusion that enables compositional T2V generation. Vico (Yang and Wang 2024) builds a spatial-temporal attention graph to update the noise latent. There exist works that employ an LLM for planning layouts, such as RPG (Yang et al. 2024a) for text-to-image generation, and LVD (Lian et al. 2023), VideoDirectorGPT (Lin et al. 2024) and BlobGEN-Vid (Feng et al. 2025) for video generation. However, the existing works focus on generation in one go, failing to meet complex compositional requirements. Our work introduces a collaborative workflow with iterative loop that allows for precise alignment with compositional prompts, progressively refining key elements to achieve greater coherence across spatial and temporal dimensions.

LLM-based Agents. Recent advancements in (M)LLMs have boosted the development of highly capable AI agents, applied across various domains, such as software development (Wang et al. 2024a; Qian et al. 2024), robotics (Driess et al. 2023), scientific research (Tang et al. 2024), society simulation (Park et al. 2023), and beyond. A rapidly growing research focuses on automating interactions with computer environments to solve tasks, such as web manipulation (Yao et al. 2023; Deng et al. 2023), gaming (Wang et al. 2023a), command-line coding (Sun et al. 2024b), and text-to-image generation (Wang et al. 2024b). Various approaches (Park et al. 2023; Sun et al. 2024c; Wu et al. 2023a; Hong et al. 2024; Yuan et al. 2024) have been proposed to enable collaboration and communication among multi-agent to overcome hallucinations. While these methods have shown promising results in areas such as automated coding, they often rely on homogeneous agents, limiting the diversity and specialization required for more complex tasks as compositional text-to-video generation. Our work introduces a heterogeneous and hierarchical multi-agent system designed to handle various aspects of compositional requirements in text-to-video generation, expanding the range and effectiveness of multi-agent collaboration in this domain.

Methodology

Following the principle of task decomposition and role specialization, we introduce GENMAC, a multi-agent collaboration framework for compositional text-to-video generation.

Collaborative Workflow

Inspired by the human artistic workflow, our framework adopts a DESIGN → GENERATION → REDESIGN collaborative workflow, as shown in Figure 2.

Stage I: DESIGN. Our DESIGN stage translates the text prompt into a structured layout, which outlines the key instances, spatial relationships, and temporal dynamics required for compositional video generation. We leverage an LLM as the design agent to generate layouts, *i.e.*, structured bounding boxes (which include object IDs, names, box sizes, and positions) for each frame and each instance based on the given text prompt as LVD (Lian et al. 2023). This stage provides dynamic layout and semantic information in the GENERATION stage.

Stage II: GENERATION. We synthesize the video conditioned on the layout provided by the DESIGN stage based on a pre-trained text-to-video diffusion model. To enforce the spatial and temporal constraints specified by the dynamic layouts, we adopt an attention-guidance mechanism from LVD (Lian et al. 2023). Specifically, our goal is to increase the cross-attention scores for each object within its designated bounding box during the denoising process.

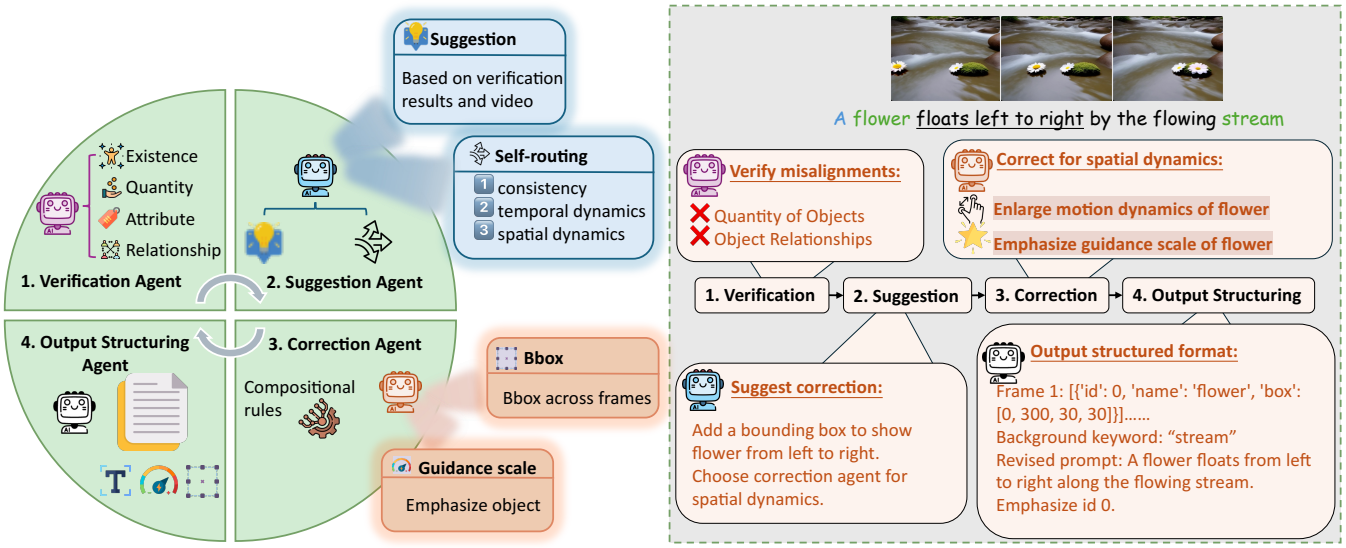

Figure 3: **Illustration Example** of the division-of-labor strategy for the REDESIGN stage: Verification, Suggestion, Correction, and Output Structuring Agent within a sequential division-of-labor strategy, highlighting the clear functional role of each agent.

Given a text prompt $\mathcal{P}$, we first extract the object tokens $\mathcal{O} = \{o_1, \ldots, o_k\}$ that correspond to the objects specified in the layout constraints. For each object $o \in \mathcal{O}$, energy function $\mathcal{L}_o$ is defined based on its cross-attention map A_t^o and target bounding box mask M_t^o as shown in Equation (1). By minimizing this energy function, we encourage high attention values inside the designated bounding box (the first term) while suppressing attention values outside the box (the second term).

$$\mathcal{L} = \sum_{o \in \mathcal{O}} \mathcal{L}_o$$
$$\mathcal{L}_o = -\beta \cdot \text{Top}k(A_t^o \odot M_t^o) + \text{Top}k\big(A_t^o \odot (1 - M_t^o)\big), \tag{1}$$

where A_t^o denotes the cross-attention map from the latent layers to the object token o at denoising timestep t, M_t^o denotes a binary mask for object o's bounding box at timestep t, $\odot$ denotes element-wise multiplication, and $\text{Top}k(\cdot)$ calculates the average of the top-k values. The guidance scale β controls the relative importance between encouraging attention inside the box versus suppressing attention outside.

We then update the noisy latent z_t by descending along the gradient of this energy function:

$$z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}, \tag{2}$$

where α_t is the step size. This updated latent z_t' is then used for the next denoising step, effectively steering the video generation to align with the desired layout.

Stage III: REDESIGN. The REDESIGN stage is the core stage of our framework. It aims to detect misalignment between the generated video and the compositional prompt, and adjust the design accordingly for re-generation. Unlike LVD (Lian et al. 2023), which uses a fixed text prompt, predefined guidance scale and single-pass bounding box predictions, our framework leverages the multi-agent system through division-of-labor and self-routing mechanism (detailed in the next subsection) in the REDESIGN stage. Three inputs are dynamically adjusted for Equation (1): $\mathcal{P}$, where the agents can refine or rewrite the prompt to improve object fidelity or style; β, where the agents can increase or decrease β to control how strictly the generation must adhere to the bounding boxes, based on the evaluation of previously generated frames; M_t^o, where the agents can adjust the spatio-temporal coordinates for each object at each frame, ensuring both spatial coherence and motion directions.

Iterative Loop. For complex compositions, a single-pass through the workflow may not address all issues in the generated video. Therefore, we introduce an iterative refinement loop between the GENERATION and REDESIGN stages, allowing progressive correction to meet compositional requirements like attribute binding, spatial relationships, and object counts. With correction guidance from the REDESIGN stage, including $\mathcal{P}$, β, and M_t^o, the GENERATION stage iteratively improves the video generation results. The loop terminates once the generated video aligned with $\mathcal{P}$ (verified by Verification Agent in the next subsection) or a maximum iteration number is reached.

Division-of-Labor Strategy in the REDESIGN Stage

The REDESIGN stage requires a sophisticated process of understanding generated videos, reasoning about video-text misalignments, and planning corrections. We found this task too complex for a single agent, often leading to hallucinations (Table 2 and Figure 5). Motivated by the principle of division-of-labor, we decompose the REDESIGN stage into a sequential pipeline of four specialized MLLM-based agents: a **Verification Agent** (π_{veri}), a **Suggestion Agent** (π_{sugg}), a **Correction Agent** (π_{corr}), and an **Output Structuring Agent** (π_{output}). The specific function of each agent is defined by its role and position within our structured workflow.

We formalize the iterative refinement loop at iteration i as follows. The process involves two types of outputs: free-form natural language for reasoning (ρ) and structured data for execution (ϵ). Let $\mathcal{V}_i$ be the video generated at iteration i and $\mathcal{P}$ be the initial text prompt. The information flows sequentially through the agents:

- **Verification Agent (π_{veri})**: This agent first checks how well the video $\mathcal{V}_i$ aligns with the prompt $\mathcal{P}$. It focuses on key aspects like object existence, quantity, attributes, and relationships, producing a reasoning report ρ_i detailing any misalignments. For instance, it might output: "There are two flowers in the video, while the text prompt indicates one flower". If alignment is verified, π_{veri} outputs a terminal flag.
$$\rho_i = \pi_{\text{veri}}(\mathcal{V}_i, \mathcal{P}) \tag{3}$$
- **Suggestion Agent (π_{sugg})**: Based on the verification report ρ_i, this agent proposes corrective actions in natural language (ρ_i') and, crucially, performs self-routing by selecting the most appropriate correction agent (detailed in the next subsection). The execution output ϵ_i^c is the identifier for the chosen correction agent.
$$(\rho_i', \epsilon_i^c) = \pi_{\text{sugg}}(\mathcal{V}_i, \rho_i) \tag{4}$$
- **Correction Agent (π_{corr})**: The selected correction agent π_{corr} receives the suggestion ρ_i' and the previous iteration's structured design ϵ_{i-1}^s (which contains the layout, guidance scale, and prompt). It formulates the concrete corrections, such as new bounding box coordinates or an adjusted guidance scale, as a reasoning output ρ_i''.
$$\rho_i'' = \pi_{\text{corr}}^{(\epsilon_i^c)}(\mathcal{V}_i, \rho_i', \epsilon_{i-1}^s) \tag{5}$$
- **Output Structuring Agent (π_{output})**: Finally, this agent translates the correction plan ρ_i'' into a structured JSON format, ϵ_i^s. This output, containing the updated bounding boxes, guidance scale, and potentially a revised prompt, is then fed into the GENERATION stage for the next iteration.
$$\epsilon_i^s = \pi_{\text{output}}(\mathcal{V}_i, \rho_i'') \tag{6}$$

This structured, sequential process ensures a detailed refinement of the video generation.

Self-Routing Mechanism for Correction

A single, general-purpose correction agent struggles with the diverse errors in compositional video generation (Table 2). We therefore introduce a suite of specialist **Correction Agents**, each an expert in a specific domain. The Suggestion Agent π_{sugg} analyzes the verification report ρ_i and routes the task to the most appropriate Correction Agent π_{corr} based on the nature of the detected misalignment. Our primary specialists are:

- **Consistency Agent**: Handles errors where attributes or spatial layouts should remain static but fail to do so across frames.
- **Temporal Dynamics Agent**: Corrects issues related to attribute changes or actions evolving over time (*e.g.*, a color changing).
- **Spatial Dynamics Agent**: Specializes in complex motion paths and changing spatial relationships between objects. For example, to correct a failed "move from left to right" instruction, the Suggestion Agent would route the task to this agent, which is prompted to propose corrections with a larger range of movement (example shown in Figure 3 right part).

This self-routing mechanism allows our system to adaptively apply the most suitable expertise to each specific problem, significantly improving correction effectiveness. Examples of the full process are detailed in (Huang et al. 2024a).

Experiments

Experimental Setups

Implementation Details. We apply our GENMAC on VideoCrafter2 (Chen et al. 2024) and HunyuanVideo (Kong et al. 2024) as the backbones. We use GPT-4o (OpenAI 2024) as (M)LLM agents. See more details in appendix.

Evaluated Models. We compare our approach with 22 text-to-video generation models, including 20 open-source models and 2 commercial models (details in appendix).

Benchmark and Evaluation Metrics. We use T2V-CompBench (Sun et al. 2024a) as the benchmark to evaluate the quality of compositional text-to-video generation from seven aspects: consistent and dynamic attribute binding, spatial relationships, motion binding, action binding, object interactions, and generative numeracy.

Quantitative Comparisons

We quantitatively compare our GENMAC with text-to-video generation models, evaluating seven crucial compositional aspects in Table 1 (detailed in appendix). (1) Compared to VideoCrafter2 (Chen et al. 2024) and Hunyuan (Kong et al. 2024) baselines, our method shows improvements across all seven categories, with an average of 35.70% and 23.43%, respectively. (2) Compared to other baselines, our GENMAC achieves superior performance. Among the baselines, the models such as Open-Sora-Plan (Lab and etc. 2024), Wan (Wan et al. 2025), CogVideoX (Yang et al. 2024b), the commercial Gen-3 (Runway AI 2024), and the methods specifically designed for compositionality like VideoTetris (Tian et al. 2024) and Vico (Yang and Wang 2024), can achieve higher quality. (3) HunyuanVideo baseline (Kong et al. 2024) performs well in attribute binding, action and interaction, but does not perform well in challenging compositional benchmarks, particularly in dynamic attribute binding, motion binding, and numeracy. Our method brings improvements, with an average increase of 39.23% in these challenging categories.

Qualitative Comparisons

Comparison with Existing Methods. We show visual comparisons of our proposed GENMAC with VideoCrafter2 (Chen et al. 2024) and HunyuanVideo (Kong et al. 2024) in Figure 4. We can observe that existing models struggle to meet compositional requirements, while our GENMAC generates videos that adhere to complex compositional scenarios: In the left example,

Model	Consist-attr	Dynamic-attr	Spatial	Motion	Action	Interaction	Numeracy
Metric	Grid-LLaVA ↑	D-LLaVA ↑	G-Dino ↑	DOT ↑	Grid-LLaVA ↑	Grid-LLaVA ↑	G-Dino ↑
Open-Sora 1.2 (hpcaitech 2024)	0.6600	0.1714	0.5406	0.2388	0.5717	0.7400	0.2556
Open-Sora-Plan v1.1.0 (Lab and etc. 2024)	0.7413	0.1770	0.5587	0.2187	0.6780	0.7275	0.2928
T2V-Turbo-v2 (Li et al. 2024)	0.7275	0.2280	0.5590	0.2686	0.6500	0.7900	0.2828
CogVideoX-5B (Yang et al. 2024b)	0.7220	0.2334	0.5461	0.2943	0.5960	0.7950	0.2603
Wan 2.1 (Wan et al. 2025)	0.7025	0.2158	0.5490	0.2942	0.6580	0.7375	0.3993
VideoTetris (Tian et al. 2024)	0.7125	0.2066	0.5148	0.2204	0.5280	0.7600	0.2609
Vico (Yang and Wang 2024)	0.7025	0.2376	0.4952	0.2225	0.5480	0.7775	0.2116
LVD (Lian et al. 2023)	0.5595	0.1499	0.5469	0.2699	0.4960	0.6100	0.0991
Pika (Pika 2023) (Commercial)	0.6513	0.1744	0.5043	0.2221	0.5380	0.6625	0.2613
Gen-3 (Runway AI 2024) (Commercial)	0.7045	0.2078	0.5533	0.3111	0.6280	0.7900	0.2169
VideoCrafter2 (Chen et al. 2024)	0.6663	0.2308	0.5106	0.2178	0.5640	0.8125	0.2869
+ GENMAC (Ours)	**0.7875**	**0.2498**	**0.7461**	**0.3623**	0.7273	0.8250	**0.5166**
HunyuanVideo (Kong et al. 2024)	0.7550	0.2106	0.6327	0.2067	0.6600	0.7700	0.1094
+ GENMAC (Ours)	**0.7950**	0.2362	0.7001	0.2331	**0.7880**	**0.8625**	0.2109

Table 1: **Quantitative Comparison** on T2V-CompBench (Sun et al. 2024a). Compared with existing text-to-video generation models and compositional methods, GENMAC demonstrates exceptional performances in consistent attribute binding, dynamic attribute binding, spatial relationships, motion binding, action binding, object interactions, and generative numeracy, indicating our method achieves superior compositional generation ability. The baseline data are sourced from (Sun et al. 2024a).

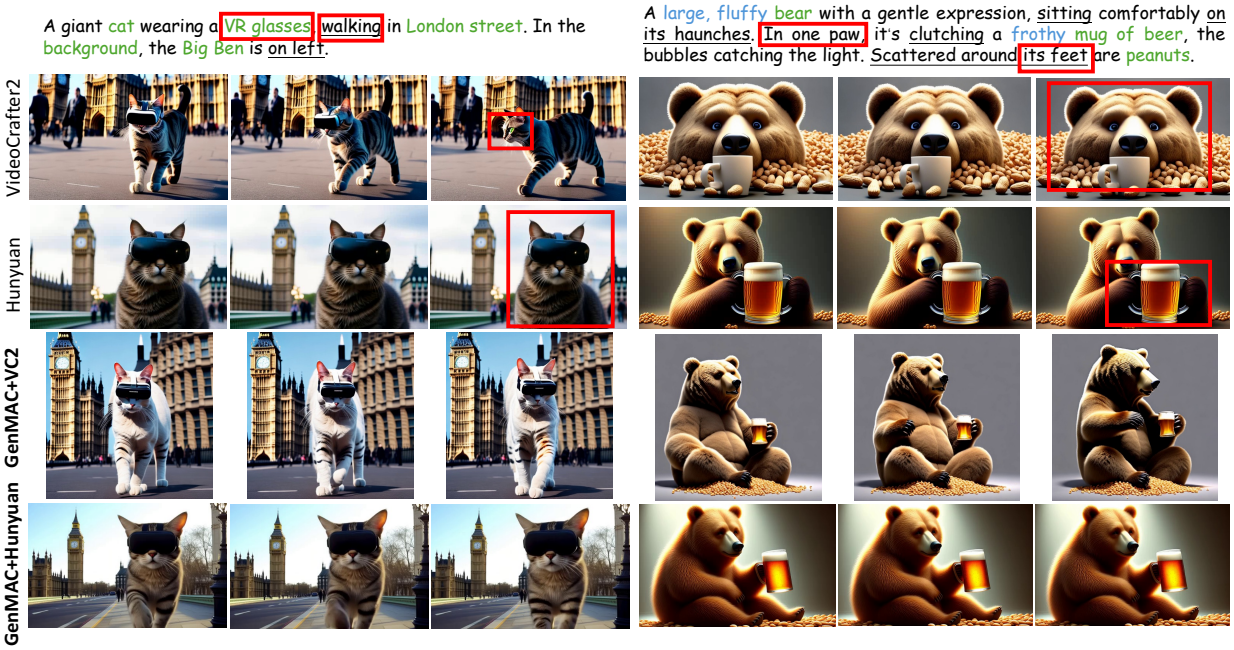

Figure 4: The errors are marked in red boxes. Our proposed GENMAC generates videos that accurately adhere to complex compositional scenarios in comparision with text-to-video models.

Metric	Multiple stages					Division-of-Labor						Self-routing		Ours
	Gen	+Red	+Iter	D+G	+Red	SA	+Iter	V+C	+Iter	V+S+C	+Iter	w/o	+Iter	
Average ↑	0.4684	0.5154	0.5559	0.5496	0.5790	0.5525	0.5508	0.5399	0.5484	0.5649	0.5971	0.5704	0.5800	**0.6021**

Table 2: **Ablation Study** of GENMAC+VC2. The complete framework (ours) achieves the highest scores. Gen: Generation, Red: Redesign, Iter: Iterative, D+G: Design+Generation, SA: Single-agent, V+C: Verification+Correction agent, V+S+C: Verification+Suggestion+Correction agent. Average: average score of 7 categories in T2V-CompBench.

VideoCrafter2 (Chen et al. 2024) omits the VR glasses, and HunyuanVideo (Kong et al. 2024) only shows a

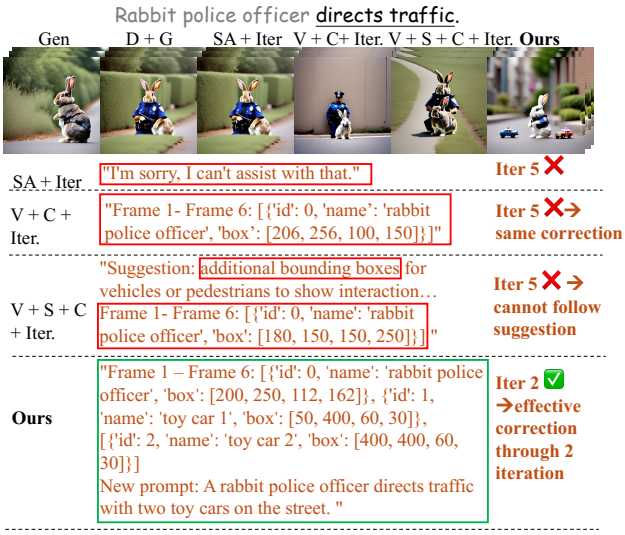

Figure 5: **Ablation study** of multiple stages, division-of-labor in the REDESIGN stage. Our method shows the effective correction and compositional prompt following ability.

standing cat instead of walking. In the right example, VideoCrafter2 (Chen et al. 2024) and HunyuanVideo (Kong et al. 2024) only show partial views of the bear without "feet" in the text prompt. See more examples in appendix.

More Qualitative Examples. The results in Figure 1 show that GENMAC demonstrates better performances in compositionality. See more examples in appendix.

Analysis on Iterative Generation

We calculate the cumulative corrected ratio within each compositional category at every iteration, which is the ratio of prompts that have completed the refinement and exited the GENMAC loop to the total size of the subset. As shown in Figure 6, the corrected ratio gradually increases with iterations across all compositional aspects, demonstrating the necessity of iterative refinement. Among the seven compositional aspects, we can observe that dynamic attribute binding presents the greatest challenge, consistently showing the lowest corrected ratios across iterations. In contrast, consistent attribute binding and spatial relationships begin with higher corrected ratios. We show more analysis (iteration process, efficiency, *etc.*) in appendix.

Ablation Study

We perform ablation studies in Table 2 and Figure 5.

Effect of Multiple Stages and Iterative Refinement. Table 2 shows that with only a GENERATION stage yields the lowest generation quality. Introducing a DESIGN stage ("DESIGN+GENERATION") improves the quality. Adding one ("+REDESIGN") or multiple iterative REDESIGN stages ("+iterative") further enhances the alignment.

Effect of Division-of-Labor in the REDESIGN Stage. Table 2 shows that the proposed method can bring notable improvements over single-agent or naive multi-agent framework. Removing the Output Structuring Agent and Suggestion Agent from the REDESIGN stage leads to significant degradation in quality. Qualitative results show that the single agent fails to understand the compositional requirements, and is unable to respond as the task exceeds its capability. Refinement without Suggestion Agent (V+C+Iter) produces the same correction in each iteration without making effective progress. Refinement without Output Structuring Agent (V+S+C+Iter) fails to follow suggestion correctly, and is unable to make necessary corrections. Baselines reach the maximum number of iterations (5) while our method completes the effective correction within 2 iterations. We show the full illustration in appendix.

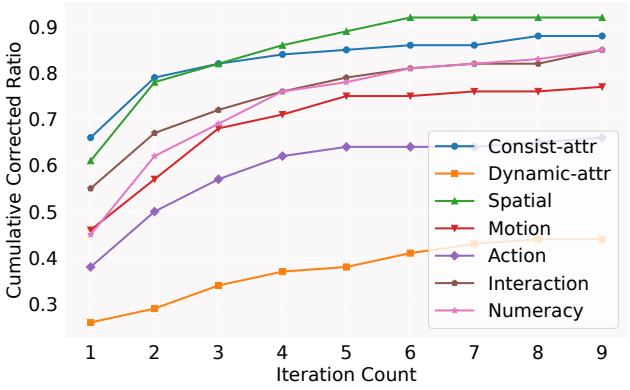

Figure 6: Cumulative Corrected Ratio. Dynamic attribute binding remains challenging, while generative numeracy, spatial relationships, and motion binding show substantial improvements.

Effect of Self-Routing for the Correction Agent. We compare our method with the version without the self-routing mechanism for the Correction Agent in Table 2. The results (0.6021 *v.s.* 0.5800) highlight the advantage of the self-routing mechanism.

Conclusion

In this paper, we address the challenges faced by state-of-the-art video generation models in producing complex compositional video content. Specifically, we introduce a multi-agent collaboration framework that enables high-quality compositional generation. Our framework incorporates three-stage collaborative workflow: DESIGN, GENERATION, and REDESIGN. We further decompose the core REDESIGN stage into four sequential tasks executed by specialized agents by a division-of-labor strategy: verification, suggestion, correction, and output structuring. Finally, we design a self-routing mechanism that adaptively selects from a suite of correction agents, enabling better handling of diverse compositional aspects. These agents in our framework leverage off-the-shelf (M)LLMs, each fulfills a specific role within the framework. Extensive experimental results confirm the effectiveness and superiority of our method in generating compositional text-to-video generation.

Acknowledgements

This work is supported in part by the NSFC-RGC Joint Research Scheme through the Research Grant Council of Hong Kong under grant N_HKU76925, and in part by National Nature Science Foundation of China (No. 62561160156).

References

Blattmann, A.; Dockhorn, T.; and Kulal, S. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; and Ling, H. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Chang, H.; Zhang, H.; and Barber, J. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chang, H.; Zhang, H.; and Jiang, L. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11315–11325.

Chefer, H.; Alaluf, Y.; and Vinker, Y. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *ACM Trans. Graph.*

Chen, H.; Zhang, Y.; and Cun, X. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*.

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-free layout control with cross-attention guidance. In *WACV*.

Chen, X.; Liu, Y.; and Yang, Y. 2023. Reason out your layout: Evoking the layout master from large language models for text-to-image synthesis. *arXiv preprint arXiv:2311.17126*.

Deng, X.; Gu, Y.; and Zheng, B. 2023. Mind2Web: Towards a Generalist Agent for the Web. arXiv:2306.06070.

Driess, D.; Xia, F.; and Sajjadi, M. S. M. 2023. PaLM-E: An Embodied Multimodal Language Model. arXiv:2303.03378.

Feng, W.; He, X.; and Fu, T.-J. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *ICLR*.

Feng, W.; Liu, C.; and Liu, S. 2025. BlobGEN-Vid: Compositional Text-to-Video Generation with Blob Video Representations.

Gani, H.; Bhat, S. F.; and Naseer, M. 2023. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. *arXiv preprint arXiv:2310.10640*.

Guo, Y.; Yang, C.; and Rao, A. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

He, Y.; Yang, T.; and Zhang, Y. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*.

Ho, J.; Chan, W.; and Saharia, C. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Hong, S.; Zhuge, M.; and Chen, J. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv:2308.00352.

Hong, W.; Ding, M.; and Zheng, W. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

hpcaitech. 2024. Open-sora: Democratizing efficient video production for all.

Huang, K.; Duan, C.; and Sun, K. 2025. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01): 1–17.

Huang, K.; Huang, Y.; and Ning, X. 2024a. Genmac: compositional text-to-video generation with multi-agent collaboration. *arXiv preprint arXiv:2412.04440*.

Huang, K.; Sun, K.; and Xie, E. 2024b. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*.

Khachatryan, L.; Movsisyan, A.; and Tadevosyan, V. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15954–15964.

Kim, Y.; Lee, J.; and Kim, J.-H. 2023. Dense text-to-image generation with attention modulation. In *ICCV*.

Kondratyuk, D.; Yu, L.; and Gu, X. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

Kong, W.; Tian, Q.; and Zhang, Z. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. arXiv:2412.03603.

Lab, P.-Y.; and etc., T. A. 2024. Open-Sora-Plan.

Li, J.; Long, Q.; and Zheng, J. 2024. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*.

Li, Y.; Liu, H.; and Wu, Q. 2023. Gligen: Open-set grounded text-to-image generation. In *ICCV*.

Li, Z.; Min, M. R.; and Li, K. 2022. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *CVPR*.

Lian, L.; Shi, B.; and Yala, A. 2023. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*.

Lin, H.; Zala, A.; and Cho, J. 2024. VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning. In *COLM*.

Liu, N.; Li, S.; and Du, Y. 2022. Compositional visual generation with composable diffusion models. In *ECCV*.

Liu, X.; Hu, T.; and Wang, W. 2024. Referee Can Play: An Alternative Approach to Conditional Generation via Model Inversion. *arXiv preprint arXiv:2402.16305*.

Luo, Z.; Chen, D.; and Zhang, Y. 2023. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10209–10218.

Meral, T. H. S.; Simsar, E.; and Tombari, F. 2023. CONFORM: Contrast is All You Need For High-Fidelity Text-to-Image Diffusion Models. *arXiv preprint arXiv:2312.06059*.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-11-14.

Park, D.; Kim, S.; and Moon, T. 2024. Rare-to-Frequent: Unlocking Compositional Generation Power of Diffusion Models on Rare Concepts with LLM Guidance. *arXiv preprint arXiv:2410.22376*.

Park, D. H.; Azadi, S.; and Liu, X. 2021. Benchmark for compositional text-to-image synthesis. In *NeurIPS*.

Park, J. S.; O'Brien, J. C.; and Cai, C. J. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442.

Patel, M.; Kim, C.; and Cheng, S. 2023. Eclipse: A resource-efficient text-to-image prior for image generations. *arXiv preprint arXiv:2312.04655*.

Pika. 2023. Pika Art. https://pika.art/. Accessed: 2024-11-14.

Qian, C.; Liu, W.; and Liu, H. 2024. ChatDev: Communicative Agents for Software Development. arXiv:2307.07924.

Rassin, R.; Hirsch, E.; and Glickman, D. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *NeurIPS*.

Runway AI. 2024. Gen-3. https://runwayml.com/blog/introducing-gen-3-alpha/. Accessed: 2024-11-14.

Singer, U.; Polyak, A.; and Hayes, T. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Sohl-Dickstein, J.; Weiss, E. A.; and Maheswaranathan, N. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:1503.03585.

Song, Y.; Sohl-Dickstein, J.; and Kingma, D. P. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. arXiv:2011.13456.

Strauss, A. 1985. Work and the division of labor. *Sociological quarterly*, 26(1): 1–19.

Sun, K.; Huang, K.; and Liu, X. 2024a. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. arXiv:2407.14505.

Sun, Q.; Chen, Z.; and Xu, F. 2024b. A Survey of Neural Code Intelligence: Paradigms, Advances and Beyond. arXiv:2403.14734.

Sun, Q.; Yin, Z.; and Li, X. 2024c. Corex: Pushing the Boundaries of Complex Reasoning through Multi-Model Collaboration. arXiv:2310.00280.

Taghipour, A.; Ghahremani, M.; and Bennamoun, M. 2024. Box It to Bind It: Unified Layout Control and Attribute Binding in T2I Diffusion Models. *arXiv preprint arXiv:2402.17910*.

Tang, X.; Jin, Q.; and Zhu, K. 2024. Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science. arXiv:2402.04247.

Tian, Y.; Yang, L.; and Yang, H. 2024. VideoTetris: Towards Compositional Text-to-Video Generation. arXiv:2406.04277.

Villegas, R.; Babaeizadeh, M.; and Kindermans, P.-J. 2022. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*.

Wan, T.; Wang, A.; and Ai, B. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wang, G.; Xie, Y.; and Jiang, Y. 2023a. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.

Wang, J.; Yuan, H.; and Chen, D. 2023b. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, R.; Chen, Z.; and Chen, C. 2023c. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*.

Wang, X.; Li, B.; and Song, Y. 2024a. OpenHands: An Open Platform for AI Software Developers as Generalist Agents. arXiv:2407.16741.

Wang, Z.; Li, A.; and Li, Z. 2024b. GenArtist: Multimodal LLM as an Agent for Unified Image Generation and Editing. arXiv:2407.05600.

Wang, Z.; Xie, E.; and Li, A. 2024c. Divide and Conquer: Language Models can Plan and Self-Correct for Compositional Text-to-Image Generation. *arXiv preprint arXiv:2401.15688*.

Wu, C.; Huang, L.; and Zhang, Q. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*.

Wu, C.; Liang, J.; and Ji, L. 2022. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, 720–736. Springer.

Wu, Q.; Bansal, G.; and Zhang, J. 2023a. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155.

Wu, Q.; Liu, Y.; and Zhao, H. 2023b. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *ICCV*.

Wu, T.-H.; Lian, L.; and Gonzalez, J. E. 2023c. Self-correcting LLM-controlled Diffusion Models. arXiv:2311.16090.

Yang, L.; Yu, Z.; and Meng, C. 2024a. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*.

Yang, X.; and Wang, X. 2024. Compositional Video Generation as Flow Equalization. arXiv:2407.06182.

Yang, Z.; Teng, J.; and Zheng, W. 2024b. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yao, S.; Chen, H.; and Yang, J. 2023. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. arXiv:2207.01206.

Yu, L.; Cheng, Y.; and Sohn, K. 2023a. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10459–10469.

Yu, L.; Lezama, J.; and Gundavarapu, N. B. 2023b. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. *arXiv preprint arXiv:2310.05737*.

Yuan, Z.; Liu, Y.; and Cao, Y. 2024. Mora: Enabling Generalist Video Generation via A Multi-Agent Framework. arXiv:2403.13248.

Zhou, D.; Wang, W.; and Yan, H. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.