MLLM-Guided Semantic Correction for Text-to-Video Generation

Junhao Chen, Zheqi Lv, Keting Yin, Shengyu Zhang,
Zhou Zhao, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Fei Wu

arXiv:2608.16513v1 [cs.CV] 17 Aug 2026

***Abstract*—Recent advances in diffusion models and Transformer architectures have led to significant progress in text-to-video generation. However, these models often suffer from semantic errors such as missing objects, incorrect attributes, or mismatched actions. Although some semantic correction methods perform optimization before sampling or refinement after sampling, how to detect and correct semantic deviations during the video generation process remains underexplored. In this paper, we introduce a training-free, interpretable mid-generation correction framework that integrates multimodal large language model (MLLM) feedback directly into the diffusion sampling loop. Our framework achieves diffusion trajectory correction by injecting semantic evaluation signals during video synthesis, enabling the model to optimize the generated content through continuous self-reflection. We propose two key modules: a Semantic Assessment Supervisor that generates intermediate preview frames for semantic evaluations and deviation diagnostics, and a Semantic Modification Assistant that corrects semantic drift during inference via a controllable latent trajectory intervention. Our method improves semantic alignment, visual fidelity, and temporal consistency without modifying model parameters. We validate the effectiveness of our approach through extensive experiments across multiple benchmarks.**

***Index Terms*—Text-to-video generation, Multimodal large language model.**

I. Introduction

Recently, rapid advances in diffusion models [1], [2] and transformer architectures [3] have resulted in substantial progress in the field of text-to-video generation. Both proprietary business models (e.g., Sora [4]) and open-source frameworks (e.g., HunyuanVideo [5], CogVideoX [6], [7]) exhibit a strong ability to adhere to textual prompts throughout the generation process, substantially improving overall video quality. However, conventional text-to-video diffusion models typically lack mechanisms for understanding the semantics encoded in the intermediate latent representations. During the generation process, the lack of semantic awareness prevents the model from detecting and correcting semantic deviations, such as missing objects, incorrect attributes, or mismatched actions (see Fig. 3).

To address these challenges, existing approaches can be broadly categorized into three groups (see Fig. 1). (1) *Non–self-correcting methods*: Classical techniques such as Classifier-Free Guidance (CFG) [8] enhance prompt adherence by scaling the guidance strength. However, excessively high and static guidance scales often compromise visual fidelity and motion diversity while amplifying accumulated errors along the diffusion trajectory. (2) *Starting-point correction methods*: Methods (e.g., Free-Bloom [9], GPT4Motion [10], FreeInit [11]) improve alignment by optimizing prompts or the initial noise distribution before sampling. These techniques mitigate early-stage drift but are inherently unable to react to semantic deviations that occur later in the trajectory. (3) *End-point correction methods*: Approaches (e.g., VideoRepair [12], NeuS-E [13]) perform refinement after the generation is completed or during the late inference stage by employing external scorers or feedback evaluators to assess and correct semantic inconsistencies. Although these refinement strategies improve semantic alignment, they are inherently reactive, correcting errors post-generation instead of preventing them during sampling. In summary, existing methods lack the capacity to adaptively correct dynamic semantic deviations within the diffusion sampling process.

In this paper, we introduce a *training-free, interpretable, midgeneration correction* framework that integrates multimodal large language model (MLLM) feedback directly into the diffusion sampling loop. The framework injects semantic evaluation signals during video synthesis, enabling online trajectory correction. Our key insight is to enable the diffusion model to operate analogously to a painter who inspects and adjusts the evolving canvas, thereby functioning as an active semantic reasoner that performs continuous self-reflection and refinement during generation. Specifically, an MLLM is employed to analyze intermediate generations and guide the diffusion process to correct semantic drift, introducing an external supervisory signal that prevents error accumulation during sampling.

Our insights guide our progressive pipeline, which consists of two modules: *Semantic Assessment Supervisor* and *Semantic Modification Assistant*. (i) *Semantic Assessment Supervisor*. A naive strategy of providing raw, noisy latents to an MLLM for evaluation proves ineffective, as intermediate representations lack coherent semantics. Therefore, at selected sampling steps, the Semantic Assessment Supervisor makes the model generate intermediate preview frames that are processed by the MLLM to produce semantic evaluations and bias diagnostics. (ii) *Semantic Modification Assistant*. Once these assessments are available, a central technical question remains: How to incorporate feedback into the diffusion trajectory in a controllable and temporally consistent manner without violating model priors? To solve this, we introduce a Semantic Modification Assistant. First, we perform an unconditional reverse diffusion to roll the current latent back to an uncondi-

Junhao Chen and Keting Yin are with the School of Software Technology, Zhejiang University, Hangzhou 310027, China (e-mail: chenjunhao100@zju.edu.cn; yinkt@zju.edu.cn).

Zheqi Lv, Shengyu Zhang, Zhou Zhao, and Fei Wu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China(e-mail: zheqilv@zju.edu.cn; sy_zhang@zju.edu.cn; zhaozhou@zju.edu.cn; wufei@zju.edu.cn).

Feiyang Chen, Xinyu Duan, and Baoxing Huai are with the AI System Innovation Lab, Huawei Cloud Computing Technology Co., Ltd., Hangzhou 310005, China(e-mail: chenfeiyang2@huawei.com; duanxinyu@huawei.com; huaibaoxing@huawei.com).

tional anchor state, thereby removing accumulated conditional errors. Then, using MLLM-derived enhancement and suppression cues, a conditional forward diffusion step reintroduces corrected semantic constraints into the generation process. Together, the *Semantic Assessment Supervisor* converts the diffusion model from a passive generator into an active self-monitor, while the *Semantic Modification Assistant* enables dynamic, inference-time semantic correction without modifying generator parameters, which we term understanding-aware generation.

The contributions of this paper can be summarized as:

- We propose a *training-free, interpretable, mid-generation correction* framework that embeds MLLM feedback directly into the text-to-video diffusion sampling loop, improving semantic alignment and visual quality without modifying model parameters.
- We design two key modules, *Semantic Assessment Supervisor* and *Semantic Modification Assistant*, that respectively enable reliable interpretation of intermediate states and controllable corrective interventions within the ongoing sampling process, achieving inference-time semantic correction.
- We demonstrate through extensive experiments that our method significantly enhances semantic consistency, visual fidelity, and temporal coherence across multiple text-to-video benchmarks, validating the efficacy of MLLM-guided semantic feedback during diffusion sampling.

II. Related Work

A. Text-to-Video Generation

Text-to-video (T2V) generation aims at synthesizing coherent videos conditioned on textual descriptions. Recent advances in diffusion models have established two main research paradigms: training-based [14]–[28] and training-free approaches [9], [11], [29], [30]. Training-based methods typically train large-scale video diffusion models directly on paired video–text datasets to capture spatial and temporal correlations. For example, Video Diffusion Models (VDM) [19] extend image diffusion UNets into 3D UNets for joint spatio-temporal learning, while Make-A-Video [20] learns visual–textual alignment from image–text pairs and acquires motion understanding from unlabeled video data. These methods achieve high visual fidelity but require substantial computational resources and large video–text datasets for training. In contrast, training-free methods leverage pre-trained text-to-image diffusion models to synthesize videos without additional training. Approaches such as Text2Video-Zero [29], FreeBloom [9], and FreeInit [11] adapt image diffusion backbones by introducing temporal attention, cross-frame consistency modules, or motion priors to produce temporally coherent frames. Despite their efficiency, these models often suffer from semantic deviations, generating scenes or actions that drift from the input text due to the lack of explicit semantic reasoning. To address this issue, we propose a training-free MLLM-guided inference framework that integrates textual and visual understanding into the diffusion process.

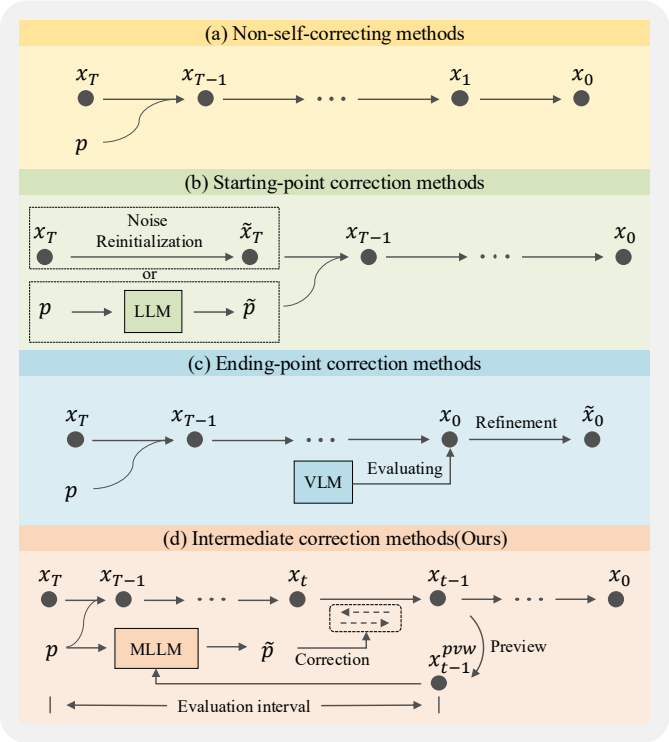

Fig. 1. **Overview of semantic correction strategies for text-to-video generation.** (a) *Non–self-correcting*: fixed-scale guidance tends to amplify artifacts and motion errors. (b) *Starting-point correction*: prompt/noise optimization before sampling cannot prevent drift during generation. (c) *End-point correction*: post-generation refinement reacts to errors but cannot prevent them. (d) *Intermediate correction(Ours)*: a training-free, MLLM-guided mid-generation correction that injects semantic feedback into the diffusion loop to maintain coherence during synthesis.

B. LLM-assisted Video Generation

Recent advances in Large Multimodal Models (LMMs) [31]–[33] have significantly influenced video generation. Several works employ Large Language Models (LLMs) as high-level planners that interpret textual prompts into structured scene descriptions or multiscene scripts before generation [9], [10], [12], [13], [34]–[38]. For example, LVD [35] transforms text into detailed scene layouts to guide diffusion-based video synthesis, while VideoDrafter [37] employs LLMs to produce multiscene scripts and maintain entity consistency via reference images. Similarly, Mora [38] introduces an LLM-driven multi-agent framework for generalist video generation and editing. While these methods apply LLM reasoning as static guidance before or after diffusion, our mid-generation MLLM mechanism provides dynamic semantic feedback, which relates to recent efforts that leverage semantic representations for perceptual evaluation and content understanding [39]–[41].

III. Method

In this section, we present our proposed framework for interpretable mid-generation correction in text-to-video diffusion models. We first formalize the diffusion-based generation process and introduce key notation (§III-A). Then, we describe the two major components of our framework: the *Semantic*

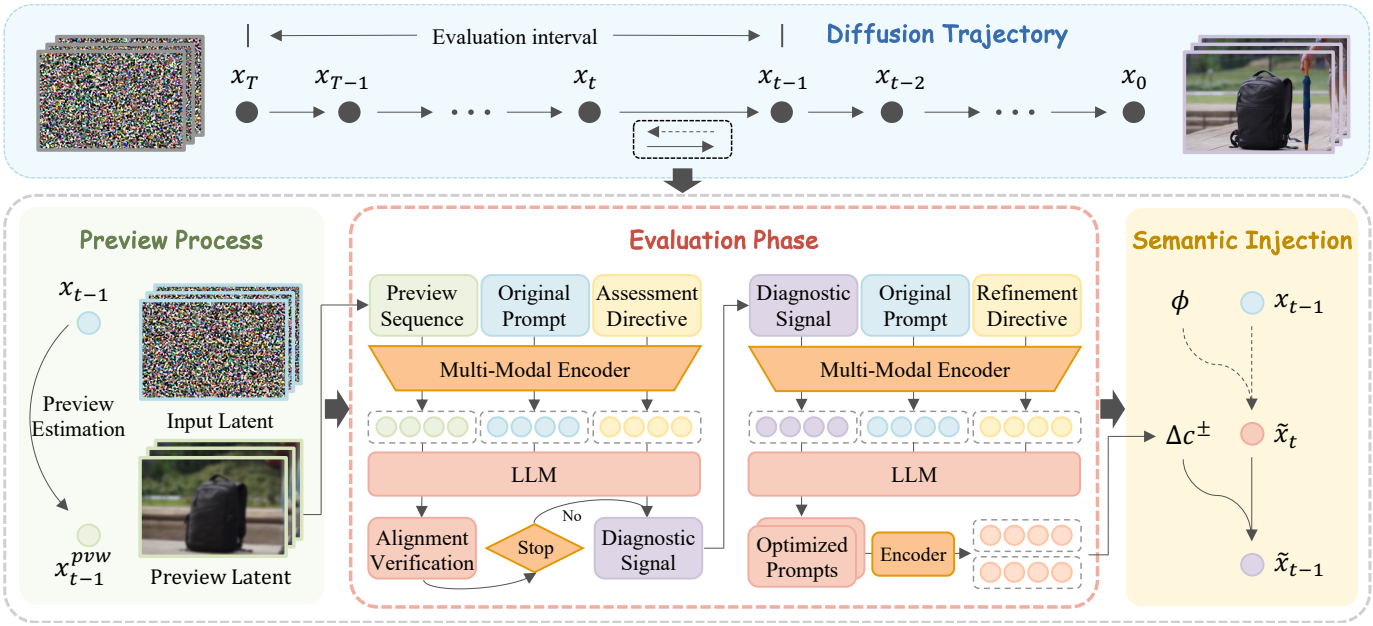

Fig. 2. **Overview of the proposed mid-generation semantic correction framework (SASMA).** The framework, termed SASMA (Semantic Assessment Supervisor and Modification Assistant), integrates MLLM feedback into the diffusion sampling loop for dynamic semantic correction during text-to-video generation. During sampling, intermediate latents are periodically decoded into preview frames for semantic evaluation (left and middle). The MLLM provides diagnostic feedback and corrective prompts, which are encoded into semantic deltas $\Delta c_t^{\pm}$ and injected back into the diffusion process (right) to guide subsequent denoising steps toward semantically consistent video synthesis.

Assessment Supervisor (§III-B) and the *Semantic Modification Assistant* (§III-C), which jointly enable correction based on dynamic reasoning during inference.

A. Background and Preliminaries

1) Diffusion model: Let a text-to-video diffusion model be parameterized by θ, designed to generate a sequence of video frames conditioned on a textual prompt p. We denote the conditioning embedding as $c = E(p)$, where $E(\cdot)$ is a pretrained text encoder that maps the input prompt to a semantic feature space.

The model operates in the latent space, denoted by $\{x_t\}_{t=0}^{T}$, where $x_T \sim \mathcal{N}(0, I)$ represents Gaussian noise, and x_0 denotes the final clean latent to be decoded into a video sequence. Following the deterministic DDIM formulation [42], the sampling process is expressed as:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{x}_0(x_t, t, c; \theta) + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(x_t, t, c), \quad (1)$$

where ϵ_θ is the denoising network that estimates the noise component at timestep t, and $\hat{x}_0$ is the model's prediction of the corresponding clean latent.

After the final denoising step, the latent x_0 is decoded into a sequence of video frames by a video decoder $D(\cdot)$:

$$\mathbf{v} = D(x_0) = \{v_1, v_2, \ldots, v_N\}, \quad (2)$$

where N denotes the number of generated frames.

2) Incorporating multimodal reasoning: To enable interpretable and adaptive correction during sampling, we integrate a MLLM denoted as $M(\cdot)$, which performs high-level semantic reasoning across both visual and textual modalities. At an arbitrary timestep t, we partially decode the latent x_t through $D(\cdot)$ to obtain a coarse visual representation $\mathbf{v}_t = D(x_t)$, and feed it into the MLLM together with the textual prompt p:

$$S_t = M(\mathbf{v}_t, p), \quad (3)$$

where S_t denotes the semantic reasoning feedback that encapsulates multimodal understanding signals and potential corrective signals.

B. Semantic Assessment Supervisor

A fundamental challenge of mid-generation correction lies in the nature of intermediate diffusion states. At timestep t, the latent x_t is dominated by stochastic noise and resides far from any interpretable visual manifold. Consequently, direct reasoning over x_t offers little semantic insight, as its structure does not correspond to perceivable content. This limitation requires a mechanism that can expose the evolving semantics of the model without disrupting its generative trajectory.

We draw inspiration from the diffusion process itself. Each denoising step implicitly carries an internal estimate of the underlying clean signal, formalized in DDIM as:

$$\hat{x}_0(x_t, t, c; \theta) = \frac{x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x_t, t, c)}{\sqrt{\alpha_t}}. \quad (4)$$

Rather than treating $\hat{x}_0$ merely as an auxiliary computation, we reinterpret it as the model's instantaneous hypothesis of the latent clean sample.

Let t_{start} and t_{end} denote the evaluation interval boundaries and Δ the interval step. The set of timesteps at which the MLLM evaluation is performed is defined as:

$$\mathcal{T} = \left\{ t \,\middle|\, t = t_{\text{start}} + k \cdot \Delta,\ k = 0, 1, \ldots, \left\lfloor \tfrac{t_{\text{end}} - t_{\text{start}}}{\Delta} \right\rfloor \right\}. \quad (5)$$

For each $t \in \mathcal{T}$, we decode an intermediate preview:

$$\mathbf{v}_t^{pvw} = D(\hat{x}_0(x_t, t, c; \theta)), \quad (6)$$

which yields a low-fidelity, yet semantically aligned, representation that reflects the evolving generation state of the model.

The intermediate visualization $\mathbf{v}_t^{pvw}$ and the textual prompt p are then fed into the MLLM $M(\cdot)$ to produce diagnostic feedback:

$$S_t = (f_t, p_t^+, p_t^-) = M(\mathbf{v}_t^{pvw}, p), \quad \Delta c_t^{\pm} = E(p_t^{\pm}), \tag{7}$$

where f_t denotes the structured diagnostic feedback describing semantic inconsistencies (e.g., object incompleteness or attribute mismatch), (p_t^+, p_t^-) represents the positive and negative corrective prompts generated according to the diagnosis, and $E(\cdot)$ encodes these prompts into the conditioning embedding space. The resulting corrective embeddings $\Delta c_t^{\pm}$ are forwarded to the injection module (§III-C) for integration into subsequent diffusion updates.

In particular, before generating feedback f_t, our framework first evaluates the visual-semantic consistency between $\mathbf{v}_t^{pvw}$ and p. If no inconsistency is detected, the correction process is terminated early to avoid unnecessary semantic injection and reduce computational overhead.

Algorithm 1 MLLM-Guided Semantic Correction for Text-to-Video Generation

INITIALIZE$(p, E, \epsilon_\theta, M, D)$
$c \leftarrow E(p)$
$x_T \sim \mathcal{N}(0, I)$
$\mathcal{T} \leftarrow \{\, t_{\text{start}} + k\Delta \mid k \in \mathbb{N},\ t \le t_{\text{end}} \,\}$

DIFFUSION SAMPLING
for $t = T, \dots, 1$ **do**
$\hat{x}_0 \leftarrow (x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x_t, t, c))/\sqrt{\alpha_t}$
$x_{t-1} \leftarrow \sqrt{\alpha_{t-1}}\,\hat{x}_0 + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(x_t, t, c)$
if $t \in \mathcal{T}$ **then**
SEMANTIC ASSESSMENT
$\mathbf{v}_t^{\text{pvw}} \leftarrow D(\hat{x}_0)$
$(f_t, p_t^+, p_t^-) \leftarrow M(\mathbf{v}_t^{\text{pvw}}, p)$
$\Delta c_t^+ \leftarrow E(p_t^+), \quad \Delta c_t^- \leftarrow E(p_t^-)$
SEMANTIC CORRECTION
$\lambda_t \leftarrow \sqrt{1-\alpha_t} - \sqrt{\alpha_t/\alpha_{t-1}}\sqrt{1-\alpha_{t-1}}$
$\tilde{x}_t \leftarrow \sqrt{\alpha_t/\alpha_{t-1}}\,x_{t-1} + \lambda_t\,\epsilon_\theta(x_{t-1}, t-1)$
$\hat{x}_0 \leftarrow (x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}))/\sqrt{\alpha_t}$
$x_{t-1} \leftarrow \sqrt{\alpha_{t-1}}\,\hat{x}_0 + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm})$
end if
end for

OUTPUT
$\mathbf{v} \leftarrow D(x_0)$
return $\mathbf{v}$

C. Semantic Modification Assistant

Once the semantic feedback is obtained from the Semantic Assessment Supervisor, the key question becomes how to effectively inject this corrective information into the diffusion trajectory while maintaining temporal coherence and preserving model priors. To achieve this, we introduce a **Semantic Injection** process that performs controllable bidirectional refinement within the diffusion trajectory. This mechanism allows the model to partially reverse the accumulated semantic bias and then reintroduce corrected guidance under updated conditions. Putting the two components together, our full algorithm is presented in algorithm 1.

Generally, our semantic injection process consists of three sequential steps.

1) Semantic Dilution: Given the latent x_{t-1} at step $t-1$, we first compute an intermediate latent $\tilde{x}_t$ by applying a one-step back transition that weakens the previously accumulated conditional influence:

$$\begin{aligned} \tilde{x}_t &= \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\,x_{t-1} + \lambda_t\,\epsilon_\theta(x_{t-1}, t-1, \phi), \\ \lambda_t &= \sqrt{1-\alpha_t} - \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\sqrt{1-\alpha_{t-1}}, \end{aligned} \tag{8}$$

where ϵ_θ denotes the noise prediction network parameterized by θ, and ϕ represents the unconditional case. This step effectively dilutes the condition-induced semantic information, producing a semantically neutral latent $\tilde{x}_t$ that serves as a clean basis for subsequent feedback-guided correction.

2) Semantic Injection: We apply the precomputed corrective embeddings $\Delta c_t^{\pm}$ as modified conditioning signals and perform a feedback-guided denoising step:

$$\begin{aligned} \tilde{x}_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{x}_0(\tilde{x}_t, t, \Delta c_t^{\pm}; \theta) \\ + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}), \end{aligned} \tag{9}$$

where $\hat{x}_0(\tilde{x}_t, t, \Delta c_t^{\pm}; \theta)$ denotes the denoised estimate under the modified semantic guidance. This step injects the corrected semantic information into the diffusion trajectory, yielding a refined latent $\tilde{x}_{t-1}$ that incorporates high-level feedback from the preview representation.

3) : Trajectory Resumption The corrected latent re-enters the normal diffusion trajectory and continues with normal denoising using the original conditioning:

$$\begin{aligned} x_{t-2} = \sqrt{\alpha_{t-2}}\,\hat{x}_0(\tilde{x}_{t-1}, t-1, c; \theta) \\ + \sqrt{1-\alpha_{t-2}}\,\epsilon_\theta(\tilde{x}_{t-1}, t-1, c), \end{aligned} \tag{10}$$

where c is the initial condition. This step ensures that the generation process is resumed on the corrected trajectory.

In conclusion, this process enables the diffusion model to integrate high-level semantic feedback in a stable and interpretable manner.

IV. THEORETICAL FOUNDATIONS

In this section, we provide a rigorous theoretical analysis of the proposed Semantic Assessment–Semantic Modification Assistant (SASMA). Our objective is to formally characterize how semantic feedback injection alters the diffusion trajectory and to establish conditions under which such modification yields a strictly smaller denoising error than standard DDIM inference.

We begin by explicitly formulating the three sequential operations that constitute one semantic correction cycle in SASMA. Based on these operations, we derive an exact four-term decomposition of the resulting update and subsequently analyze its denoising error relative to the baseline diffusion process.

A. Semantic Correction as a Three-Step Diffusion Operator

At a given timestep t, SASMA modifies the standard diffusion trajectory through the following three operations.

a) Semantic Dilution.: The first step performs a controlled forward diffusion from x_{t-1} to an intermediate noisy state $\tilde{x}_t$:

$$\tilde{x}_t = \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\, x_{t-1} + \left(\sqrt{1-\alpha_t} - \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\sqrt{1-\alpha_{t-1}}\right)\epsilon_\theta(x_{t-1}, t-1, \phi). \tag{11}$$

This step intentionally dilutes the current latent state, creating a re-noised representation that allows semantic feedback to exert non-local influence on the diffusion trajectory.

b) Semantic Injection.: Given the intermediate state $\tilde{x}_t$, we perform a backward diffusion conditioned on the refined semantic guidance $\Delta c_t^{\pm}$:

$$\tilde{x}_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{x}_0(\tilde{x}_t, t, \Delta c_t^{\pm}; \theta) + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}). \tag{12}$$

This step injects semantic corrections directly into the denoising process, yielding a refined latent $\tilde{x}_{t-1}$.

c) Trajectory Resumption.: Finally, the diffusion trajectory is resumed under the original prompt condition c:

$$x_{t-2} = \sqrt{\alpha_{t-2}}\,\hat{x}_0(\tilde{x}_{t-1}, t-1, c; \theta) + \sqrt{1-\alpha_{t-2}}\,\epsilon_\theta(\tilde{x}_{t-1}, t-1, c). \tag{13}$$

B. Exact Algebraic Reformulation

We first rewrite Eq. (12) by substituting the definition of $\hat{x}_0$:

$$\tilde{x}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\,\tilde{x}_t + \left(\sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}}\right)\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}). \tag{14}$$

Similarly, Eq. (13) can be reformulated as:

$$x_{t-2} = \sqrt{\frac{\alpha_{t-2}}{\alpha_{t-1}}}\,\tilde{x}_{t-1} + \left(\sqrt{1-\alpha_{t-2}} - \sqrt{\frac{\alpha_{t-2}(1-\alpha_{t-1})}{\alpha_{t-1}}}\right)\epsilon_\theta(\tilde{x}_{t-1}, t-1, c). \tag{15}$$

Substituting Eq. (14) into Eq. (15) and expanding terms yields:

$$x_{t-2} = \sqrt{\frac{\alpha_{t-2}}{\alpha_t}}\,\tilde{x}_t + \sqrt{\frac{\alpha_{t-2}}{\alpha_{t-1}}}\left(\sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}}\right)\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}) + \left(\sqrt{1-\alpha_{t-2}} - \sqrt{\frac{\alpha_{t-2}(1-\alpha_{t-1})}{\alpha_{t-1}}}\right)\epsilon_\theta(\tilde{x}_{t-1}, t-1, c). \tag{16}$$

Substituting $\tilde{x}_t$ using Eq. (11) and collecting terms, we arrive at the following four-term decomposition:

$$x_{t-2} = \eta_1 x_{t-1} + \eta_2\epsilon_\theta(x_{t-1}, t-1, \phi) + \eta_3\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}) + \eta_4\epsilon_\theta(\tilde{x}_{t-1}, t-1, c), \tag{17}$$

where the coefficients η_1–η_4 depend solely on the noise schedule:

$$\eta_1 = \sqrt{\frac{\alpha_{t-2}}{\alpha_{t-1}}}, \tag{18}$$

$$\eta_2 = \sqrt{\frac{\alpha_{t-2}(1-\alpha_t)}{\alpha_t}} - \sqrt{\frac{\alpha_{t-2}(1-\alpha_{t-1})}{\alpha_{t-1}}}, \tag{19}$$

$$\eta_3 = \sqrt{\frac{\alpha_{t-2}(1-\alpha_{t-1})}{\alpha_{t-1}}} - \sqrt{\frac{\alpha_{t-2}(1-\alpha_t)}{\alpha_t}}, \tag{20}$$

$$\eta_4 = \sqrt{1-\alpha_{t-2}} - \sqrt{\frac{\alpha_{t-2}(1-\alpha_{t-1})}{\alpha_{t-1}}}. \tag{21}$$

A key observation is that $\eta_2 + \eta_3 = 0$, which allows us to rewrite Eq. (17) as:

$$x_{t-2} = \eta_1 x_{t-1} + \eta_4\epsilon_\theta(\tilde{x}_{t-1}, t-1, c) + \eta_3\big[\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}) - \epsilon_\theta(x_{t-1}, t-1, \phi)\big]. \tag{22}$$

C. Comparison with Standard DDIM

For reference, the standard DDIM update from $t-1$ to $t-2$ is given by:

$$x_{t-2} = \eta_1 x_{t-1} + \eta_4\epsilon_\theta(x_{t-1}, t-1, c). \tag{23}$$

Comparing Eqs. (22) and (1), we observe that SASMA introduces an additional semantic correction term that explicitly compensates for the discrepancy between the original and refined noise predictions.

D. Denoising Error Analysis

Let ϵ denote the true noise at timestep $t-1$. The denoising error of the standard DDIM update is:

$$\delta_{\text{DDIM}} = |\eta_4| \cdot \|\epsilon_\theta(x_{t-1}, t-1, c) - \epsilon\|. \tag{24}$$

For SASMA, the denoising error becomes:

$$\delta_{\text{SASMA}} = \big\|\eta_4[\epsilon_\theta(\tilde{x}_{t-1}, t-1, c) - \epsilon] + \eta_3[\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}) - \epsilon_\theta(x_{t-1}, t-1, \phi)]\big\|. \tag{25}$$

Define the state correction gain:

$$\Delta_{\text{state}} = \|\epsilon_\theta(x_{t-1}, t-1, c) - \epsilon\| - \|\epsilon_\theta(\tilde{x}_{t-1}, t-1, c) - \epsilon\|, \tag{26}$$

and the semantic correction magnitude:

$$C_{\text{sem}} = \|\epsilon_\theta(\tilde{x}_t, t, \Delta c_t^{\pm}) - \epsilon_\theta(x_{t-1}, t-1, \phi)\|. \tag{27}$$

Applying the triangle inequality yields:

$$\delta_{\text{SASMA}} \le |\eta_4|(\|\epsilon_\theta(x_{t-1}, t-1, c) - \epsilon\| - \Delta_{\text{state}}) + |\eta_3| C_{\text{sem}}. \tag{28}$$

Therefore, when the enhanced semantic guidance satisfies:

$$|\eta_3| C_{\text{sem}} < |\eta_4| \Delta_{\text{state}}, \tag{29}$$

TABLE I
QUANTITATIVE COMPARISON PER DIMENSION ON VBENCH.

Model	Method	Subject Cons. ↑	Aesthetic ↑	Imaging ↑	Human Act. ↑	Spatial Rel. ↑	Scene ↑	Overall Cons. ↑
CogVideoX1.5	Standard	0.9088	0.5435	0.5624	0.8200	0.4328	0.3343	0.2483
	Ours (SASMA)	**0.9410**	**0.5633**	**0.5857**	**0.8440**	**0.4738**	**0.3894**	**0.2536**
HunyuanVideo	Standard	**0.9625**	0.6012	0.6179	0.8840	0.5307	0.3581	0.2619
	Ours (SASMA)	0.9604	**0.6038**	**0.6323**	**0.9020**	**0.5803**	**0.3728**	**0.2643**
AnimateDiff	Standard	0.9524	0.5938	0.6161	0.9240	0.5196	0.4987	**0.2720**
	Ours (SASMA)	**0.9569**	**0.6066**	**0.6463**	**0.9460**	**0.5355**	**0.5259**	0.2713

TABLE II
QUANTITATIVE COMPARISON OF OVERALL PERFORMANCE ON VBENCH.

Model	Method	Quality ↑	Semantic ↑	Total ↑
CogVideoX1.5	Standard	0.7717	0.6419	0.7457
	Ours (SASMA)	**0.7982**	**0.6689**	**0.7723**
HunyuanVideo	Standard	0.8171	0.6975	0.7932
	Ours (SASMA)	**0.8201**	**0.7048**	**0.7970**
AnimateDiff	Standard	0.8087	0.7195	0.7909
	Ours (SASMA)	**0.8136**	**0.7277**	**0.7963**

we obtain:

$$\delta_{\text{SASMA}} < \delta_{\text{DDIM}}. \tag{30}$$

This establishes that SASMA achieves a strictly smaller denoising error than standard DDIM under mild and practically satisfied assumptions, completing the theoretical justification.

V. EXPERIMENTS

A. Experimental Setup

1) Benchmarks and Evaluation Metrics: We perform comprehensive experiments on two widely adopted text-to-video generation benchmarks: VBench [43] and ChronoMagic-Bench [44]. VBench provides diverse and fine-grained evaluation dimensions, such as subject consistency, temporal coherence, and text-video alignment, while ChronoMagic-Bench focuses on evaluating generative temporal reasoning and long-horizon consistency.

2) Comparison Baselines: We compare our method with several state-of-the-art video diffusion frameworks, including CogVideoX1.5 [7], HunyuanVideo [5] and AnimateDiff [45]. All baselines are evaluated according to their default generation settings with prompts from the benchmark datasets. For each prompt, we generate multiple samples to measure the stability and consistency of the semantic correction performance.

3) Implementation Details: All experiments were performed on NVIDIA RTX 3090 GPUs. We adopt a DDIM-based sampling schedule with $T = 50$ steps. The semantic injection process begins with $t_s = 0.1T$ and ends at $t_e = 0.9T$, with an interval $\Delta = 5$ between the evaluation boundaries. For multimodal semantic reasoning, we employ VideoLLaMA3-7B [46] as MLLM. The model operates in a training-free inference setting, and all compared methods share identical random seeds and diffusion hyperparameters to ensure a fair comparison.

4) Backbone-Specific Configurations: All experiments adopt three representative text-to-video diffusion backbones. For each model, we denote the number of diffusion steps as T, the classifier-free guidance scale as γ, the number of generated frames as N, the frame rate as FPS, and the resulting video duration as $D = N/\text{FPS}$. For CogVideoX1.5, we set $(T, \gamma, N, \text{FPS}) = (50, 6.0, 41, 8)$, yielding $D = 5\text{s}$. For HunyuanVideo, we use $(T, \gamma, N, \text{FPS}) = (30, 6.0, 41, 8)$, also producing $D = 5\text{s}$. For AnimateDiff, due to its shorter temporal horizon, we adopt $(T, \gamma, N, \text{FPS}) = (50, 7.5, 16, 8)$, resulting in $D = 2\text{s}$. Our method operates in a training-free, plug-and-play manner and is applied consistently to all models without modifying any parameters or architectures.

B. Main Results

1) Quantitative Results: Table I and Table II display the VBench results for three representative text-to-video backbones (CogVideoX1.5, HunyuanVideo and AnimateDiff). We present ten metrics in total: the first seven columns are per-dimension measures (Subject Consistency, Aesthetic Quality, Imaging Quality, Human Action, Spatial Relationship, Scene, and Overall Consistency), and the last three columns (Quality Score, Semantic Score, and Total Score) are aggregated scores computed from all per-dimension metrics.

The quantitative trends in Table II demonstrate that our proposed SASMA improves overall performance across three architectures, with improvements spanning both semantic alignment and perceptual quality dimensions. In particular, SASMA exhibits adaptive improvement behavior in different baseline strengths. For weaker models such as CogVideoX1.5, the framework yields substantial gains in both structural and aesthetic aspects, while for already strong baselines like HunyuanVideo, it provides fine-grained refinements on semantic dimensions, especially in human action and spatial relationship understanding. Moreover, the improvements are well balanced, as SASMA simultaneously increases both the Quality Score and Semantic Score across all architectures. This indicates that a higher semantic correctness is achieved without compromising perceptual fidelity. Although individual metrics for each dimension may exhibit minor variations, the aggregated Total Score consistently improves for the three models, confirming the robustness and generalizability of SASMA in diverse diffusion architectures.

2) Qualitative Results: As shown in Figure 3, our proposed SASMA enhances both temporal consistency and

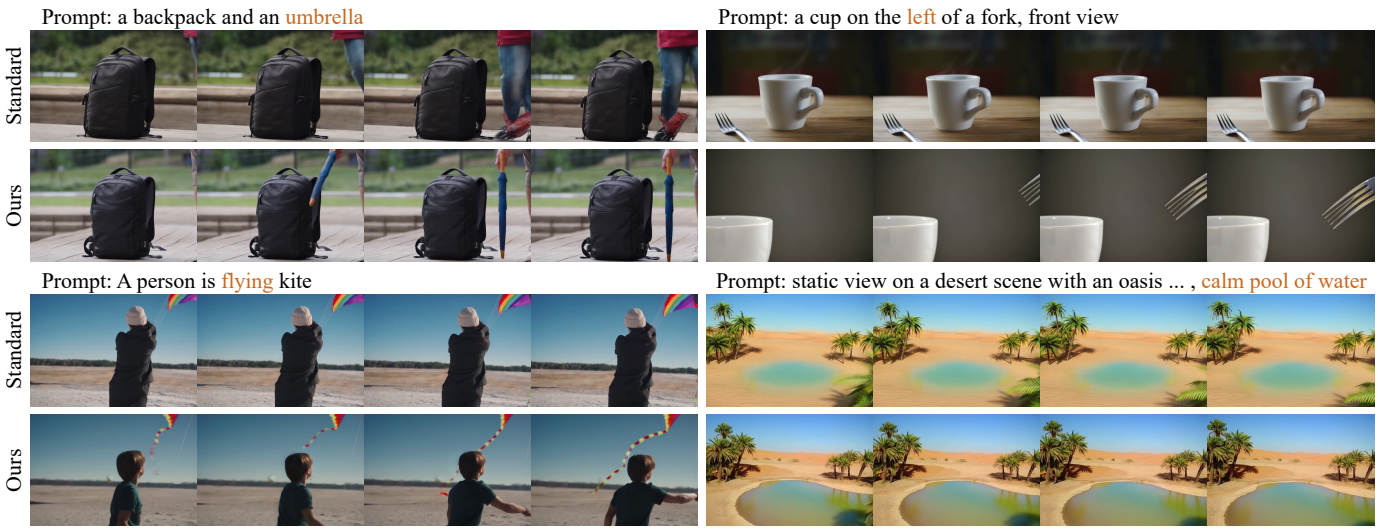

Fig. 3. **Qualitative comparison on text-to-video generation.** Our method introduces an MLLM-guided feedback mechanism into the diffusion process, allowing mid-generation semantic correction without retraining. This integration enhances both temporal consistency and visual-semantic alignment in challenging cases such as multi-object composition, spatial relations, actions, and scenes.

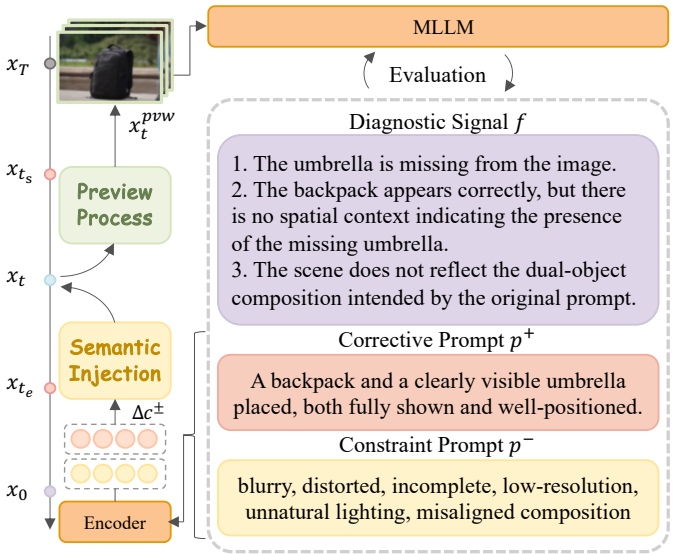

Fig. 4. **Case of interpretable evaluation and correction.** Along the denoising trajectory from x_T to x_0, our framework performs semantic assessment at discrete intervals on intermediate latents. For each evaluated timestep, the MLLM produces three interpretable signals: a structured *diagnostic signal* f identifying semantic deviations, a *corrective prompt* p^+ providing refined positive guidance, and a *constraint prompt* p^- specifying exclusion constraints.

visual-semantic alignment in text-to-video generation scenarios. In addition, Figure 4 presents a representative case of interpretable evaluation and correction, demonstrating how SASMA provides transparent and traceable feedback during the diffusion process. The qualitative evaluation covers a variety of scenes, including multi-object arrangements, spatial relationships, dynamic human activities, and natural environments with multiple interacting elements, highlighting the robustness of SASMA across diverse content.

TABLE III
ABLATION STUDY OF SASMA MODULES ON CHRONOMAGIC-BENCH-150 WITH COGVIDEOX1.5.

Method	UMT-FVD ↓	UMTScore ↑	MTScore ↑	CHScore ↑
Standard	216.68	2.8485	0.3420	45.566
+ Semantic Injection	213.26	2.8619	0.3476	59.187
+ Evaluation Module	**212.50**	**2.8771**	0.3452	60.578
+ Preview Module	213.18	2.8653	**0.3486**	**61.784**

C. Method Analysis and Ablation

1) Ablations on the component of SASMA: We conduct ablation studies on Chronomagic-Bench-150 using CogVideoX1.5 to assess the contribution of each module in the proposed SASMA framework. As shown in Table III, all modules consistently improve semantic and perceptual quality over the standard baseline. The *Semantic Injection* module enhances semantic coherence by embedding corrective signals into the diffusion trajectory. This process reduces semantic drift and improves frame-level consistency, demonstrating that even simple semantic steering can benefit mid-generation dynamics. The *Evaluation Module* further refines semantic accuracy through adaptive feedback based on MLLM assessment, improving overall alignment and structural correctness. The *Preview Module* contributes most to temporal stability. By generating cleaner intermediate previews, it enables the MLLM to reason over both spatial and temporal cues rather than noisy latents. This leads to more reliable feedback, especially for motion-related inconsistencies, resulting in smoother trajectories and stronger temporal coherence.

2) Effect of Multi-round Evaluation Strategies: We study how to schedule semantic feedback during inference, as the temporal ordering of corrective operations affects both efficacy and stability. We compare three configurations:

- **One-round.** All operations (assessment, prompt polishing, and negative prompts) are applied in a single feed-

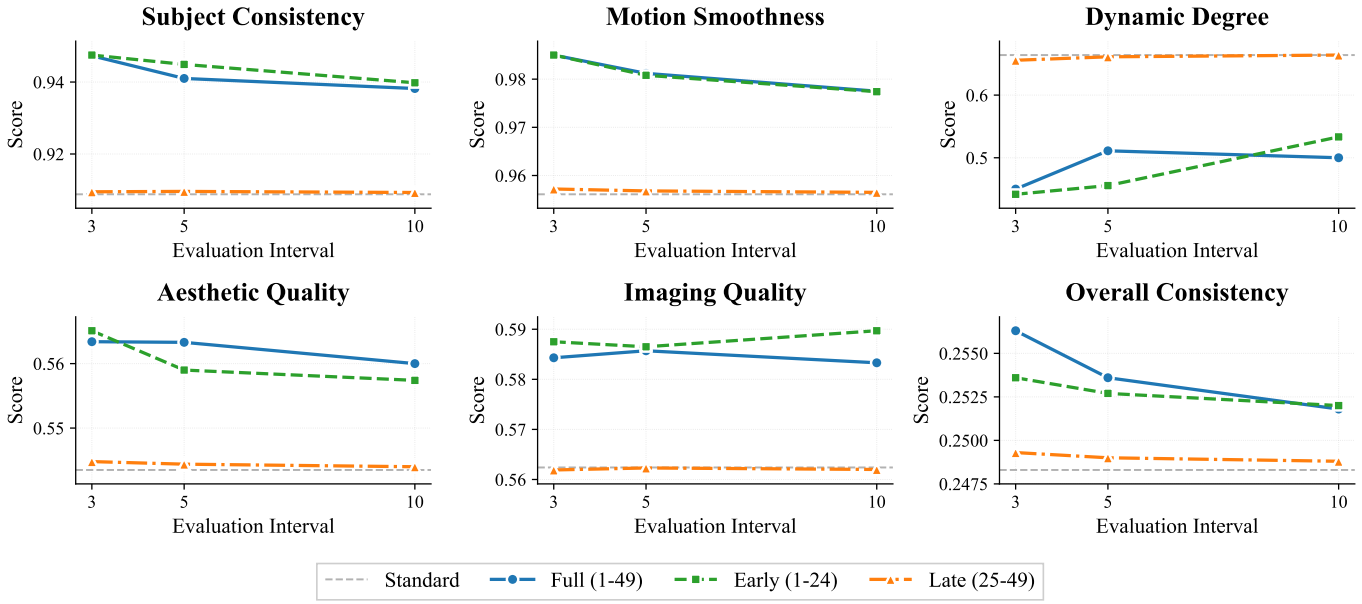

Fig. 5. **Effect of evaluation interval and stage scheduling.** We compare three stage configurations (Early: 1-24, Late: 25-49, Full: 1-49) with three evaluation intervals (3, 5, 10) on CogVideoX1.5. Early-stage high-frequency evaluation improves subject consistency and motion smoothness, while late-stage performance remains stable across frequencies. Full-process mid-frequency scheduling (interval=5) achieves optimal balance. The gray dashed line indicates baseline performance without SASMA.

TABLE IV
ABLATION ON DIFFERENT NUMBERS OF EVALUATION–FEEDBACK ROUNDS WITHIN THE SASMA FRAMEWORK ON COGVIDEOX1.5.

Method / Configuration	Motion Smooth. ↑	Dynamic Deg. ↓	Subject Cons. ↑	Aesthetic ↑	Imaging ↑	Overall Cons. ↑
Standard	0.9561	**0.6639**	0.9088	0.5435	0.5624	0.2483
Ours (1 round)	0.9812	0.4944	0.9396	0.5530	0.5745	0.2511
Ours (2 rounds)	**0.9817**	0.4167	**0.9485**	**0.5668**	**0.6004**	0.2531
Ours (3 rounds)	0.9812	0.5111	0.9410	0.5633	0.5857	**0.2536**

back round.

- **Two-rounds.** Round 1 performs semantic assessment, and Round 2 applies corrective actions (refined prompt and negative prompts).
- **Three-rounds.** Round 1 performs assessment, Round 2 applies prompt polishing, and Round 3 applies negative-prompt refinements.

Table IV demonstrates the results of CogVideoX1.5. The *two-rounds* configuration achieves the best subject consistency and perceptual scores, while the *three-rounds* approach yields higher overall consistency and more balanced improvements across temporal and semantic metrics. The *one-round* scheme delivers substantial gains over the baseline, but is less effective in refining semantic fidelity. These observations suggest that distributing corrective operations across multiple rounds enables progressive refinement, where each round builds upon the previous feedback, reducing abrupt over-corrections and cumulative errors.

Specifically, the *three-rounds* strategy separates prompt enhancement from negative prompt refinement, allowing the model to first correct positive semantic guidance before applying negative constraints. This staged approach prevents conflicts between complementary operations. Although the *two-round* strategy achieves marginal gains in specific metrics, the *three-round* strategy provides more consistent and balanced improvements with better generalization properties. We adopt the *three-round* configuration as our default based on its superior overall consistency and robustness. Notably, an early stopping mechanism is employed to terminate the refinement process once satisfactory semantic alignment is detected, allowing many samples to bypass additional correction rounds.

3) Effect of Evaluation Interval and Stage Scheduling: We investigate how the frequency of evaluation and the temporal scheduling affect the accuracy of the correction on the diffusion sampling trajectory. Specifically, we systematically vary both the correction stage and the evaluation frequency throughout the 50-step sampling process. We compare three stage configurations (early: steps 1-24, late: steps 25-49, full: steps 1-49) with three evaluation intervals (3, 5, and 10), corresponding to high-frequency, mid-frequency, and low-frequency scheduling strategies, respectively.

Figure 5 presents the quantitative results in six VBench metrics on CogVideoX1.5. The results reveal distinct stage-dependent patterns. For the early stage (1-24), high-frequency evaluation (interval=3) consistently achieves the best performance in subject consistency, motion smoothness, and aesthetic quality, as this phase corresponds to the formation of semantic structures where frequent intervention helps stabilize scene layout and subject appearance. In contrast, the late stage (25-49) exhibits minimal performance variation across

TABLE V
EFFECT OF MLLM SCALE ON SEMANTIC CORRECTION QUALITY.

MLLM	Motion Smooth.	Dynamic Deg.	Subject Cons.	Aesthetic	Imaging	Overall Cons.	VRAM (GB)
Qwen2.5-VL-3B	0.9587	**0.5018**	0.9266	0.5345	0.5684	0.2467	9,290
Qwen2.5-VL-7B	0.9808	0.4831	0.9424	**0.5669**	**0.5945**	**0.2528**	17,142
Qwen2.5-VL-72B	**0.9815**	0.4228	**0.9478**	0.5656	0.5912	0.2526	155,354

different frequencies, indicating that the generation process has stabilized and excessive evaluation may introduce unnecessary perturbations. For full-process evaluation (1-49), mid-frequency scheduling (interval=5) strikes an optimal balance between correction capability and computational efficiency, with performance approaching or even surpassing stage-specific strategies.

These observations suggest that, while a dense-to-sparse scheduling mechanism theoretically aligns with the progressive refinement nature of diffusion models, a fixed mid-frequency strategy across the entire sampling process achieves comparable effectiveness with simpler implementation. The mid-frequency configuration maintains stable performance across all metrics, demonstrating robustness in balancing semantic correction and temporal consistency. Based on these findings, we adopt interval=5 as our default configuration for subsequent experiments.

4) Effect of MLLM Scale: We investigate how the scale of the MLLM affects semantic correction quality by comparing Qwen2.5-VL at 3B, 7B, and 72B parameter scales. Table V presents the results on CogVideoX1.5.

Larger models generally improve temporal and semantic consistency metrics. The 72B model achieves the highest motion smoothness and subject consistency, reflecting its superior capability in detecting fine-grained semantic misalignments. However, the 7B model outperforms 72B in perceptual quality metrics including aesthetic quality, imaging quality, and overall consistency.

Interestingly, the 3B model achieves the highest dynamic degree. We attribute this to the correction intensity: larger models tend to provide more detailed and stringent feedback, which can over-constrain the generation process and suppress motion dynamics. Smaller models apply more conservative corrections that preserve dynamic elements while still improving semantic alignment.

The 72B model requires approximately 9$\times$ the VRAM of the 7B model (155GB vs. 17GB), yet yields only marginal improvements in a subset of metrics. The 7B model strikes an effective balance between semantic understanding capability and correction intensity, achieving the best overall consistency while maintaining computational efficiency.

D. Component-level Analysis

1) Effect of the Intermediate Preview Mechanism: To illustrate the effect of our preview mechanism, Fig. 6 visualizes a sampling trajectory from timestep t_s to t_e, where a preview is triggered at every Δ step. At each selected timestep t, we compare the raw decoded video $D(x_t)$ with the preview video obtained from $D(x_t^{pvw})$. Without the preview mechanism, the MLLM is forced to evaluate sequences whose frames are still dominated by diffusion noise, making it difficult to reliably judge object presence, attributes, or actions. In contrast, the preview video provides a much cleaner and more semantically meaningful approximation of the final output, enabling the MLLM to more accurately identify semantic errors and produce actionable feedback for subsequent correction.

2) Quantitative Validation of Preview Effectiveness: To evaluate whether intermediate samples support reliable semantic assessment, we analyze MLLM judgments at different diffusion stages. Experiments are conducted at 10 timesteps within a 50-step sampling process under two evaluation settings: (1) *Raw Decoding*, where noisy latents are directly decoded by the VAE, and (2) *Preview*, where the predicted clean latent $\hat{x}_0$ is decoded instead of the noisy x_t.

For each setting, we report two metrics: (1) the *agreement rate* between Qwen2.5-VL-7B and Qwen2.5-VL-72B, measuring judgment consistency, and (2) the *MATCH ratio*, defined as the proportion of samples judged as semantically aligned with the prompt.

Figure 7(a) shows that both settings reach high agreement ($>$85%) at later timesteps. However, high agreement alone does not indicate reliable semantic evaluation. As shown in Figure 7(b), raw decoding yields nearly zero MATCH predictions until $t/T \approx 0.5$, suggesting that intermediate samples remain semantically unrecognizable.

In contrast, preview-based evaluation produces valid semantic responses at much earlier stages. As illustrated in Figure 7(c), the MATCH ratio reaches 65% at $t/T = 0.14$ and exceeds 90% after $t/T = 0.24$. These results indicate that preview significantly improves the visibility of semantic structure in intermediate samples, enabling earlier and more stable MLLM assessment.

3) MLLM Feedback Quality Analysis: To further assess the reliability of the textual signals produced by our method, we evaluate the quality of the generated *Diagnostic Signal*, *Corrective Prompt*, and *Constraint Prompt* using an external verifier. Specifically, we generate these three forms of feedback using three different 7B or 8B MLLM models (Qwen2.5-VL-7B [47], VideoLLaMA3-7B [46], and InternVL2.5-8B [48]) and subsequently ask the more capable Qwen2.5-VL-72B [47] model to judge their consistency and correctness. The evaluation is conducted on 165 prompts selected from the *subject consistency* and *overall consistency* dimensions of VBench, which provide diverse scenarios to test the quality of textual feedback across different visual-text alignment challenges.

Figure 8 presents a comprehensive comparison of MLLM feedback quality in the three models. Across all models evaluated, the diagnostic signal and the corrective prompt

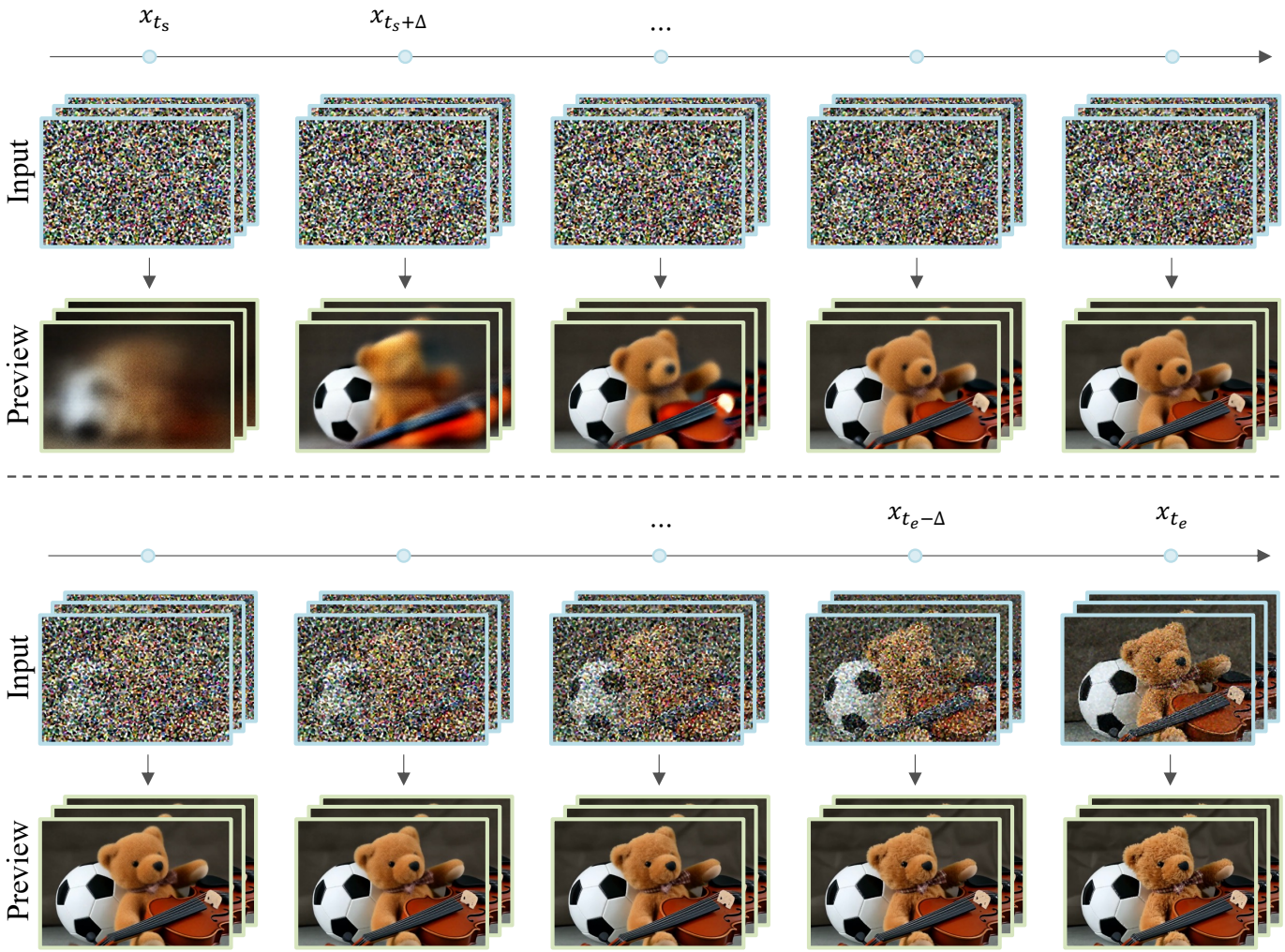

Fig. 6. **Effect of Intermediate Previewing in Mid-Generation Evaluation.** From timestep t_s to t_e, we insert preview operations at an interval of Δ steps along the diffusion trajectory. For each selected timestep t, we compare the noisy intermediate decoding $D(x_t)$ with the preview video obtained from $D(x_t^{pvw})$. The prompt is "A teddy bear sitting between a football and a violin."

consistently achieve high agreement rates with the Qwen2.5-VL-72B verifier, indicating that both the semantic descriptions of the mismatches and the proposed corrective guidance are generally reliable and well grounded in visual content. In contrast, the constraint prompt receives lower agreement scores, which suggests that describing what should be explicitly suppressed is inherently more ambiguous and model dependent. This observation aligns with the limitations discussed in Section VII, where we note that the effectiveness of the method may be partially contingent on the reliability of the MLLM feedback, especially in challenging cases or low-fidelity previews. Overall, the results confirm that the textual components driving our mid-generation correction framework are largely trustworthy, especially for diagnostic interpretation and positive guidance refinement.

TABLE VI
INFERENCE COST COMPARISON UNDER DIFFERENT CONFIGURATIONS ON COGVIDEOX1.5-5B WITH A SINGLE RTX 3090 GPU.

Configuration	Overall Cons.	Time (s)	MLLM Calls	VRAM (GB)
Standard	0.2483	254.43	0	13.7
Ours (1-round)	0.2511	579.28	4.631	30.9
Ours (2-rounds)	<u>0.2531</u>	588.95	3.289	30.9
Ours (3-rounds)	**0.2536**	590.04	3.168	30.9

TABLE VII
EARLY STOPPING STATISTICS ON VBENCH USING COGVIDEOX1.5-5B.

Dataset	Stop $\leq$1	Stop $\leq$2	Stop $\leq$3	Avg. Calls
VBench	34.96%	66.95%	78.79%	3.168

E. Efficiency and Inference Time Analysis

1) Inference Cost and Early Stopping Analysis: Table VI reports the computational cost under different multi-round configurations. Compared with the standard diffusion process, introducing semantic feedback increases the runtime due to the additional preview generation, MLLM inference, and semantic injection operations. The peak memory usage rises to 30.9,GB because the MLLM is loaded during inference, while the memory footprint remains constant across different configurations.

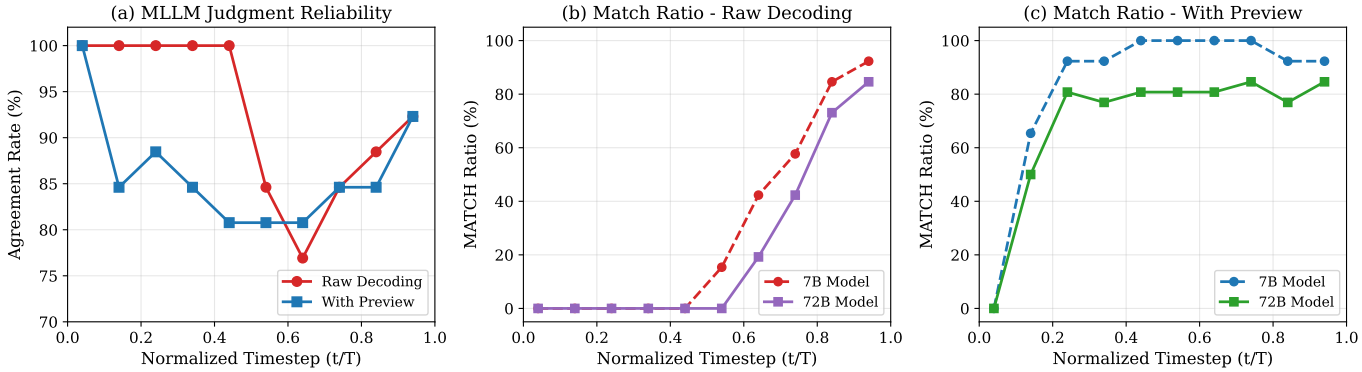

Fig. 7. Quantitative analysis of mid-generation semantic evaluation. (a) Agreement rate between Qwen2.5-VL-7B and Qwen2.5-VL-72B. (b) MATCH ratio under raw decoding. (c) MATCH ratio with preview. Raw decoding produces near-zero MATCH predictions before $t/T \approx 0.5$, whereas preview enables earlier semantic recognition.

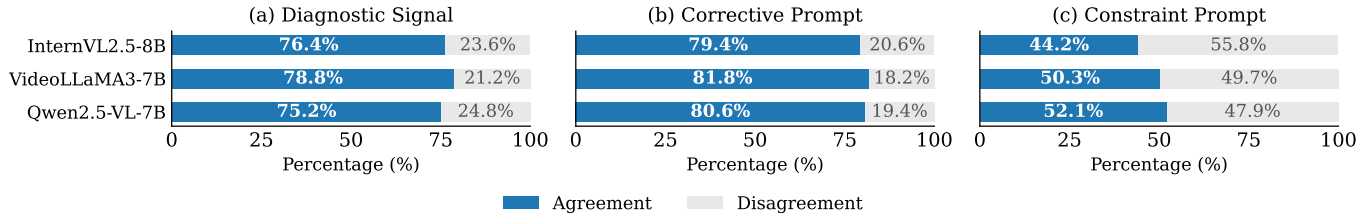

Fig. 8. Evaluation of textual feedback quality across three models (Qwen2.5-VL-7B, VideoLLaMA3-7B, and InternVL2.5-8B). Each subplot shows the agreement rates between the model-generated feedback and the Qwen2.5-VL-72B verifier for (a) Diagnostic Signal, (b) Corrective Prompts, and (c) Constraint Prompt. The results are based on 165 test prompts from VBench. Darker bars indicate agreement (Yes), while lighter bars indicate disagreement (No).

Although the two-round and three-round strategies decompose the correction process into multiple stages, their runtime is only slightly higher than that of the one-round setting. Meanwhile, the average number of MLLM calls decreases as more stages are introduced.

This behavior is explained by the early stopping mechanism. During sampling, semantic evaluation is performed at scheduled timesteps (every Δ steps). Before executing the full self-reflection pipeline, the MLLM is first asked whether the current video already matches the original prompt. If semantic alignment is confirmed, all subsequent correction operations at later timesteps are skipped and the generation continues with the standard diffusion process. Otherwise, the corresponding staged operations (assessment, prompt refinement, or negative prompt refinement) are executed according to the selected configuration.

Table VII further reports the early stopping distribution on VBench. Notably, 34.96% of samples terminate after a single MLLM evaluation, and 66.95% stop within two calls. Overall, 78.79% of samples converge within three calls, well below the maximum number of evaluations allowed by the sampling schedule. The average number of calls is only 3.168, indicating that most samples achieve semantic alignment at early stages.

As a result, the one-round configuration applies all corrections simultaneously at each evaluation step, which can lead to over-correction and introduce new semantic inconsistencies that require additional evaluations to resolve. In contrast, multi-round configurations apply corrections progressively, allowing finer-grained adjustments that are less likely to overshoot. This leads to faster convergence and fewer MLLM calls on average. The mechanism explains why the runtime difference between configurations is small despite their increased structural complexity. In large-scale generation scenarios, early stopping further reduces the amortized inference cost.

TABLE VIII
COMPONENT-WISE EFFICIENCY BREAKDOWN OF SASMA ON COGVIDEOX1.5-5B.

Component	Avg. Time (s)	Total Time (s)	Percentage
MLLM Inference	7.53	75.28	19.68%
Intermediate Preview	14.76	147.59	38.59%
Semantic Injection	15.96	159.58	41.73%
Total	—	382.45	100%

2) Component-wise Efficiency Breakdown: Table VIII reports the time distribution of the major components in SASMA, measured over 10 semantic evaluation steps on CogVideoX1.5-5B. The results show that the additional overhead is dominated by the latent-space operations associated with semantic injection (41.73%) and intermediate preview generation (38.59%), while MLLM inference accounts for a relatively smaller portion (19.68%). This observation indicates that the computational cost of SASMA is primarily determined by video decoding and diffusion-related processing rather than the multimodal reasoning itself. In practice, this suggests that further acceleration should focus on improving preview efficiency (e.g., low-resolution decoding or lightweight decoders) and reducing the frequency of semantic injection. Although

SASMA introduces additional computation, the overhead remains moderate compared with the overall diffusion cost. Combined with the performance gains reported in Table I, this trade-off demonstrates the practical feasibility of mid-generation semantic correction.

VI. Conclusion

We present a training-free framework for mid-generation semantic correction in text-to-video diffusion models. By decoupling semantic interpretation from trajectory manipulation, the Semantic Assessment Supervisor generates coherent intermediate previews and structured diagnostic signals, while the Semantic Modification Assistant performs trajectory corrections without modifying model parameters. This design enables precise semantic control and consistently improves semantic accuracy on diverse prompts. Our work demonstrates that semantic coherence can be enhanced using external multimodal reasoning signals during sampling, without modifying architectures or task-specific training, and we believe this perspective generalizes to other conditional synthesis tasks.

VII. Limitations

Our framework relies on the semantic assessment capability of the MLLM, which may exhibit prejudice when evaluating certain attributes or object categories in the low-fidelity intermediate previews. Although such limitations do not fundamentally undermine the effectiveness of the approach, they could lead to suboptimal corrections in edge cases where the diagnostic signals of MLLM are untrustworthy, suggesting that the robustness of the method is partially contingent upon the quality and generalization capability of the underlying MLLM.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS

[3] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[4] OpenAI, "Sora: Creating video from text." https://openai.com/sora, 2024.

[5] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[6] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[7] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[8] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022. [Online]. Available: https://arxiv.org/abs/2207.12598

[9] H. Huang, Y. Feng, C. Shi, L. Xu, J. Yu, and S. Yang, "Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator," *Advances in Neural Information Processing Systems*, vol. 36, pp. 26 135–26 158, 2023.

[10] J. Lv, Y. Huang, M. Yan, J. Huang, J. Liu, Y. Liu, Y. Wen, X. Chen, and S. Chen, "Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 1430–1440.

[11] T. Wu, C. Si, Y. Jiang, Z. Huang, and Z. Liu, "Freeinit: Bridging initialization gap in video diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 378–394.

[12] D. Lee, J. Yoon, J. Cho, and M. Bansal, "Videorepair: Improving text-to-video generation via misalignment evaluation and localized refinement," 2025. [Online]. Available: https://arxiv.org/abs/2411.15115

[13] M. Choi, S. P. Sharan, H. Goel, S. Shah, and S. Chinchali, "We'll fix it in post: Improving text-to-video generation with neuro-symbolic feedback," 2025. [Online]. Available: https://arxiv.org/abs/2504.17180

[14] J. An, S. Zhang, H. Yang, S. Gupta, J.-B. Huang, J. Luo, and X. Yin, "Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation," 2023. [Online]. Available: https://arxiv.org/abs/2304.08477

[15] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 563–22 575.

[16] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 7346–7356.

[17] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J.-B. Huang, M.-Y. Liu, and Y. Balaji, "Preserve your own correlation: A noise prior for video diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 930–22 941.

[18] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[19] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in neural information processing systems*, vol. 35, pp. 8633–8646, 2022.

[20] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[21] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[22] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang *et al.*, "Nuwa-xl: Diffusion over diffusion for extremely long video generation," *arXiv preprint arXiv:2303.12346*, 2023.

[23] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao, and M. Z. Shou, "Show-1: Marrying pixel and latent diffusion models for text-to-video generation," *International Journal of Computer Vision*, vol. 133, no. 4, pp. 1879–1893, 2025.

[24] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *arXiv preprint arXiv:2211.11018*, 2022.

[25] Q. Chen, Q. Wu, J. Chen, Q. Wu, A. van den Hengel, and M. Tan, "Scripted video generation with a bottom-up generative adversarial network," *IEEE Transactions on Image Processing*, vol. 29, pp. 7454–7467, 2020.

[26] K. Gao, J. Shi, H. Zhang, C. Wang, J. Xiao, and L. Chen, "Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing," in *ICML*. PMLR, 2025.

[27] H. Zhao, T. Lu, J. Gu, X. Zhang, Q. Zheng, Z. Wu, H. Xu, and Y.-G. Jiang, "Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 205–221.

[28] Q. Li, Z. Xing, R. Wang, H. Zhang, Q. Dai, and Z. Wu, "Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 12 112–12 123.

[29] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964.

[30] S. Hyun, J. Lew, J. Chung, E. Kim, and J.-P. Heo, "Frequency-based motion representation for video generative adversarial networks," *IEEE Transactions on Image Processing*, vol. 32, pp. 3949–3963, 2023.

[31] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[32] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[33] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[34] H. Lin, A. Zala, J. Cho, and M. Bansal, "Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning," *arXiv preprint arXiv:2309.15091*, 2023.

[35] L. Lian, B. Shi, A. Yala, T. Darrell, and B. Li, "Llm-grounded video diffusion models," *arXiv preprint arXiv:2309.17444*, 2023.

[36] S. Hong, J. Seo, H. Shin, S. Hong, and S. Kim, "Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation," *arXiv preprint arXiv:2305.14330*, 2023.

[37] F. Long, Z. Qiu, T. Yao, and T. Mei, "Videodrafter: Content-consistent multi-scene video generation with llm," *CoRR*, 2024.

[38] Z. Yuan, Y. Liu, Y. Cao, W. Sun, H. Jia, R. Chen, Z. Li, B. Lin, L. Yuan, L. He *et al.*, "Mora: Enabling generalist video generation via a multi-agent framework," *arXiv preprint arXiv:2403.13248*, 2024.

[39] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 2404–2418, 2024.

[40] C. Li, G. Lu, D. Feng, H. Wu, Z. Zhang, X. Liu, G. Zhai, W. Lin, and W. Zhang, "Misc: Ultra-low bitrate image semantic compression driven by large multimodal model," *IEEE Transactions on Image Processing*, vol. 34, pp. 335–349, 2025.

[41] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Task-specific normalization for continual learning of blind image quality models," *IEEE Transactions on Image Processing*, vol. 33, pp. 1898–1910, 2024.

[42] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[43] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, "VBench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[44] S. Yuan, J. Huang, Y. Xu, Y. Liu, S. Zhang, Y. Shi, R.-J. Zhu, X. Cheng, J. Luo, and L. Yuan, "Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 236–21 270, 2024.

[45] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *International Conference on Learning Representations*, 2024.

[46] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, and D. Zhao, "Videollama 3: Frontier multimodal foundation models for image and video understanding," 2025. [Online]. Available: https://arxiv.org/abs/2501.13106

[47] Q. Team, "Qwen2.5-vl," January 2025. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5-vl/

[48] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.