Uniworld-OSP2.0: A VLM-Enhanced Unified Framework for Image-to-Video Generation

Open-Sora Plan Team

Abstract

Existing conditional video generation (I2V/T2V) models often struggle with limited semantic understanding and artistic diversity. Building upon our previous Fourier-Guided Latent Shifting I2V (FlashI2V) framework, which effectively mitigates conditional image leakage and ensures high-fidelity motion, we further scale its capability for large-scale video synthesis. Specifically, we scale FlashI2V to 14B parameters and introduce a unified framework, Uniworld-OSP2.0, by integrating a 7B Visual Language Model (VLM). This design enables lossless inheritance of VLM semantic knowledge and replaces conventional shallow text encoders with multi-modal embeddings, substantially improving high-level semantic alignment and generation controllability. To further enrich artistic expressiveness, we additionally construct a stylized video dataset containing 12 distinct artistic styles, enabling a unified Image-to-Stylized-Video Generation framework that transforms an input image into a dynamic sequence rendered in a chosen style. Through large-scale training and architectural enhancements, Uniworld-OSP2.0 achieves consistent semantic fidelity and precise style control. Experiments demonstrate that our method enables controllable and semantically aligned stylized video synthesis and achieves superior performance, surpassing Wan2.1 across six key evaluation metrics.

1 Introduction

The field of conditional video generation has seen rapid advancements, with significant commercial and creative impact demonstrated in both Text-to-Video (T2V) and Image-to-Video (I2V) generation. While T2V offers unparalleled creative freedom, it often struggles with scene definition and maintaining fidelity to specific visual concepts. Conversely, I2V generation, which uses a conditional image to ensure pixel-level alignment for the first frame, has proven highly effective for initial scene control, accounting for a large portion of real-world usage calls in SOTA products like Kling (AI, 2025).

Our previous work introduced Fourier-Guided Latent Shifting I2V (FlashI2V) (Ge et al., 2025), which successfully addressed the critical issue of conditional image leakage plaguing existing I2V paradigms (e.g., SVD (Blattmann et al., 2023), Wan2.1 (Wan et al., 2025)). This leakage, caused by the denoiser taking a "shortcut" and over-relying on the concatenated image condition, resulted in performance degradation, including slow motion and color inconsistency (as analyzed in Fig. 1a. By employing Latent Shifting and Fourier Guidance, FlashI2V exhibited superior generalization and dynamic motion, effectively avoiding the overfitting observed in out-of-domain I2V tasks.

While FlashI2V framework successfully enhances motion fidelity and generalization, like many current I2V models, it faces challenges with deep semantic comprehension and precise text-driven control. This limitation often stems from relying on conventional, text-only encoders, whose features primarily capture surface-level lexical cues and lack the high-level, visually-grounded semantics necessary for robust alignment with complex visual dynamics. To address this, we introduce

See Contributions section for full author list.

(a) Performance Degradation (b) Overfitting

Figure 1: Conditional image leakage. (a) Leakage of conditioning signals leads to lower-quality generations, as observed in Wan2.1-I2V-14B-480P results on VBench-I2V. (b) The chunk-wise FVD grows on in-domain data but stays high for out-of-domain inputs, indicating poor generalization of conventional I2V models.

Uniworld-OSP2.0, a VLM-Enhanced Conditioning mechanism into the flow-matching pipeline of FlashI2V. Specifically, we leverage a powerful pretrained VLM to extract rich multi-modal features that comprehensively integrate both the linguistic input and the image conditioning. These robust multi-modal embeddings replace the shallow representations conventionally generated by a T5 encoder. To ensure optimal compatibility and prevent representational mismatch when feeding these new features into the generative core, the VLM output is first processed by a lightweight adapter module. The features transformed by this adapter are then input into the DiT backbone of FlashI2V as the sole conditional input. Crucially, the pretrained VLM remains frozen throughout the training of the generative model. This design choice guarantees the preservation of the VLM's superior cross-modal knowledge and stable zero-shot understanding, thereby providing high-quality, semantically enriched conditioning without compromising the efficiency or stability of the FlashI2V generation process. By integrating this VLM-augmented input via the adapter, we achieve a more effective bridge between high-level semantic comprehension and generative modeling, significantly enhancing controllability and context-aware video synthesis.

Furthermore, we extend our framework to stylized video generation, a relatively underexplored direction compared to image stylization. While prior works have achieved artistic rendering for static images, video stylization in a generative context remains constrained by limited datasets and weak temporal consistency. To this end, we construct a new stylized video dataset that encompasses 13 distinct artistic styles—ranging from watercolor and anime to oil painting and cyberpunk—providing an interesting dataset for downstream tasks. Leveraging this dataset, our framework realizes a unified Image-to-Stylized-Video (I2SV) generation paradigm, enabling users to transform an input image into a coherent, dynamic sequence rendered in the desired style.

In summary, our contributions are fourfold: **(1) Unified VLM-Enhanced Framework:** We integrate a pretrained Visual Language Model (VLM) into our framework, establishing a unified architecture that losslessly inherits the VLM's powerful semantic understanding while simultaneously leveraging this capability to significantly enhance video generation performance. **(2) Superior Image-to-Video Performance :** Our final model achieves superior generation quality, surpassing the advanced Wan2.1 model across multiple key evaluation metrics. **(3) Stylized Video Generation:** We extend the FlashI2V paradigm to support unified image-to-stylized-video generation, effectively achieving both strong semantic alignment and high artistic diversity. **(4) Novel Stylized Dataset Introduction:** We introduce a new Stylized Video Dataset comprising 12 distinct artistic styles, facilitating further research into controllable and expressive video generation.

2 Related Work

2.1 Text-to-Video Generation

Diffusion-based generative modeling (Ho et al., 2020; Song et al., 2020a,b) dominates recent advances in Text-to-Video (T2V) generation. Early systems extend strong image diffusion models with temporal modules atop spatial U-Net (Ronneberger et al., 2015) backbones, following a 2+1D design wherein temporal transformers are appended after spatial layers (Yuan et al., 2025c, 2024; Guo et al., 2023; Wang et al., 2025; Chen et al., 2023a, 2024a). After (Brooks et al., 2024), Diffusion Transformers (DiTs) (Peebles and Xie, 2023; Yao et al., 2025) have increasingly replaced U-Nets for video denoising (Zheng et al., 2024; Lin et al., 2024b; Ma et al., 2024; Xu et al., 2024); however,

many retain 2+1D separation, which can hamper long-range temporal reasoning due to asymmetric treatment of spatial and temporal dimensions.

Recent work therefore promotes unified 3D formulations that jointly model spatiotemporal tokens (Lin et al., 2024b). Combined with flow matching (Lipman et al., 2022; Liu, 2022) and stronger temporal attention, modern T2V models achieve high photorealism, smooth motion, and long-duration coherence. Nonetheless, semantic conditioning is often restricted by compact text encoders (e.g., T5/CLIP), which may under-ground complex or style-sensitive prompts—an issue we address by incorporating richer semantics from a pretrained Visual Language Model (VLM).

2.2 Image-to-Video Generation

Image-to-Video (I2V) augments T2V by providing a reference image to control the first frame while allowing temporal flexibility. Stable Video Diffusion (SVD) (Blattmann et al., 2023) established a widely adopted pipeline: concatenate conditional image latents with noisy latents and inject CLIP-based (Radford et al., 2021; Zhu et al., 2023; Lin et al., 2023, 2024c; Chen et al., 2024b) semantics into the denoiser. DynamiCrafter (Xing et al., 2025) enhances motion expressiveness via a query transformer (Li et al., 2023) that adaptively attends to CLIP tokens. CogVideoX (Yang et al., 2024) employs zero-padded latent concatenation for guidance, whereas SEINE (Chen et al., 2023b) casts I2V as a temporal inpainting problem. Open-Sora Plan v1.3 (Lin et al., 2024b; Li et al., 2025) unifies diverse video tasks under a progressive temporal training scheme, and Wan2.1 (Wan et al., 2025) strengthens conditioning through richer CLIP-derived semantics. These strategies involve concatenating the complete conditional image data directly into the denoiser. This technique ensures exceptional fidelity for the initial frame.

2.3 Conditional Image Leakage

Conditional image leakage (Zhao et al., 2024; Yuan et al., 2025a,b) arises when the denoiser exploits the conditional image as a shortcut, reconstructing it instead of treating it as soft guidance. This yields overfitting, muted or unnatural motion, and color drift, especially at higher noise levels. Mitigations include SVD's conditional perturbations, time-dependent noise scheduling and earlier-step initialization (Zhao et al., 2024), and training-free Adaptive Low-pass Guidance (ALG) (Choi et al., 2025), which feeds only low-frequency components early in the diffusion trajectory. Yet leakage remains persistent, limiting motion expressiveness and out-of-domain robustness. Our Fourier-Guided Latent Shifting I2V (FlashI2V) tackles the leakage at its root. Latent Shifting implicitly incorporates the image condition by modifying flow-matching distributions, reducing the incentive to copy the condition; Fourier Guidance regulates frequency content to stabilize learning and calibrate detail. FlashI2V thus suppresses leakage while maintaining fidelity and improving generalization to unseen domains.

2.4 VLM-Enhanced and Stylized Video Generation

A complementary line of research integrates large language or multimodal backbones to strengthen semantic controllability in generative models. HiDream-I1 (Cai et al., 2025) incorporates a Meta-Llama-3.1-8B backbone within a sparse DiT to improve text-conditioned image synthesis, while Qwen-Image (Wu et al., 2025) leverages Qwen2.5-VL as an auxiliary semantic encoder, aligning representations between a VLM and a diffusion transformer for better semantic consistency in image generation and editing. Although effective, such approaches often involve joint optimization or tight architectural coupling between understanding and generation modules, raising training cost and complexity. In contrast, our Uniworld-OSP2.0 uses a pretrained VLM purely as a semantic extractor. We fuse the extracted multi-modal embeddings and feed them into the DiT backbone within a flow-matching pipeline, preserving the original generative structure while substantially improving high-level semantic alignment—particularly for nuanced, style-related instructions. To expand artistic expressiveness, we construct a Stylized Video Dataset with 13 distinct artistic styles and formulate a unified Image-to-Stylized-Video framework. Conditioning on an input image and a chosen style, our method produces temporally consistent, semantically grounded stylized videos, while inheriting the strong fidelity and generalization of FlashI2V.

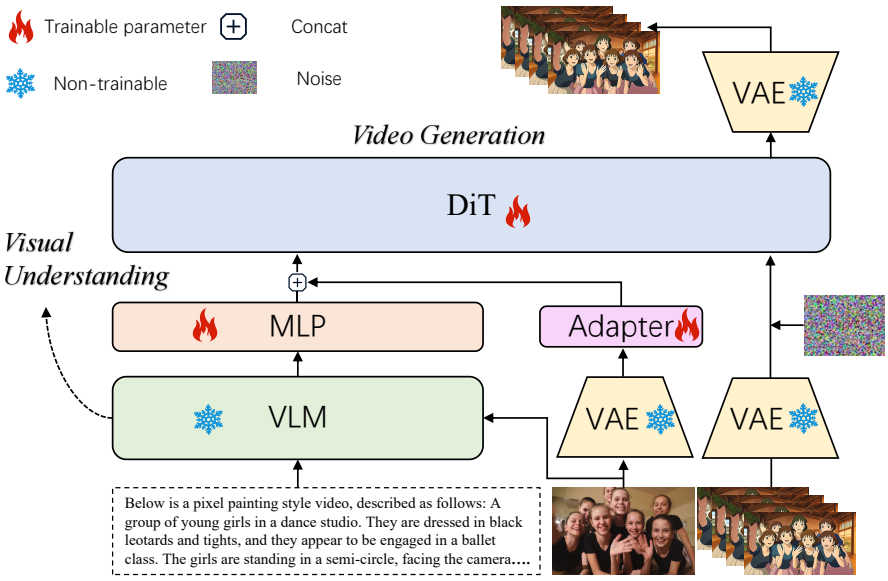


Figure 2: **Architecture of Uniworld-OSP2.0.** For effective multi-modal conditioning, we employ a frozen VLM coupled with two trainable MLPs to extract semantically rich features. Concurrently, a trainable adapter processes features from the frozen VAE encoder to ensure latent space compatibility. These enhanced multi-modal conditions are leveraged by the trainable DiT (Diffusion Transformer) to execute conditional denoising, leading to high-fidelity video synthesis in the VAE latent space.

3 Model Architecture

This section presents the overall architecture of our VLM-enhanced video generation framework, which extends the FlashI2V paradigm by incorporating robust multi-modal conditioning, as shown in Fig. 2. The proposed system fundamentally consists of three components: a Causal Variational Autoencoder (Causal VAE) for latent video representation, a VLM-Enhanced Multi-modal Conditioning Module that extracts features from a frozen VLM and passes them through a trainable Adapter, and a Diffusion Transformer (DiT) that uses these adapted features to perform conditional denoising and synthesize temporally coherent videos in the latent space.

3.1 Causal VAE

Following Open-Sora Plan (Lin et al., 2024a) and Wan2.1(Wan et al., 2025), we adopt a Causal Variational Autoencoder (Causal VAE) as the bidirectional mapping between the high-dimensional pixel space of the video and a compact, causally structured latent space. Given an input sequence

$$V \in \mathbb{R}^{F \times H \times W \times 3},$$

where $F = 1 + T$ denotes the total number of frames, the encoder compresses V into a latent representation:

$$\text{Encoder}(V) \to z \in \mathbb{R}^{(1+T/4) \times (H/8) \times (W/8) \times C},$$

with temporal downsampling factor 4 and spatial downsampling factor 8. The architecture is built upon causal 3D convolutions, ensuring that each temporal step depends only on the current and previous frames, strictly maintaining temporal causality. To improve stability and efficiency, we replace GroupNorm with RMSNorm. The first frame (conditioning frame) undergoes only spatial compression to preserve pixel-level consistency. This causal formulation supports an efficient feature Cache mechanism that accelerates decoding by reusing intermediate states across frames.

3.2 Multimodal Conditioning Module

To integrate both semantic and visual conditioning into the video generation process, we propose a multimodal conditioning module composed of three branches: a frozen visual language model

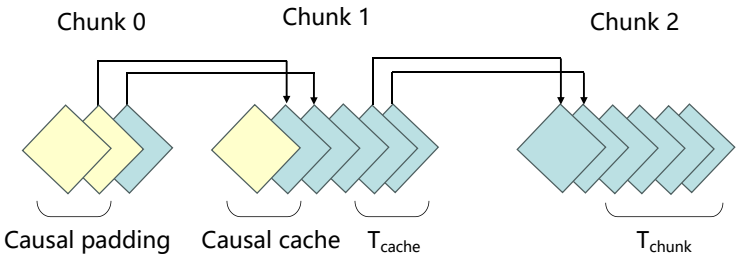


Figure 3: Illustration of Causal Cache.

for high-level semantics, a trainable text encoder for controllable linguistic guidance, and a visual encoder for spatial conditioning from the initial frame.

(1) Semantic Branch. We employ a pretrained visual-language model (Qwen2.5-VL) as a frozen multi-modal feature extractor, denoted as $E_{\mathrm{VL}}(\cdot)$. Given an input prompt T and image I, the VLM produces a semantically rich multi-modal representation:

$$F_{\mathrm{VL}} = E_{\mathrm{VL}}(T, I) \in \mathbb{R}^{L' \times D_{\mathrm{VL}}}, \tag{1}$$

where L' denotes the token length and D_{VL} the hidden dimension of the VLM. A lightweight projection head $f_{\mathrm{map}}(\cdot)$ is subsequently applied to align the extracted features with the diffusion transformer's input space:

$$F_{\mathrm{map}} = f_{\mathrm{map}}(F_{\mathrm{VL}}) \in \mathbb{R}^{L' \times D_{\mathrm{C}}}, \tag{2}$$

where D_{C} represents the hidden dimension of the diffusion transformer. The VLM remains frozen during training and serves as a stable semantic comprehension module that bridges multi-modal understanding and video generation.

(2) Visual Conditioning Branch. To provide spatial priors and appearance consistency, we introduce a visual conditioning branch based on a frozen VAE image encoder $E_{\mathrm{img}}(\cdot)$. Given the first-frame image I_0, the encoder extracts image features:

$$F_{\mathrm{img}} = E_{\mathrm{img}}(I_0) \in \mathbb{R}^{L_{\mathrm{img}} \times D_{\mathrm{img}}}, \tag{3}$$

which are subsequently projected into the diffusion transformer space through a lightweight MLP mapping:

$$F_{\mathrm{img}}^{\mathrm{map}} = f_{\mathrm{img}}(F_{\mathrm{img}}) \in \mathbb{R}^{L_{\mathrm{img}} \times D_{\mathrm{C}}}. \tag{4}$$

(3) Unified Conditioning Sequence. Finally, the multimodal conditioning sequence is constructed by concatenating the three feature sources along the token dimension:

$$c = [F_{\mathrm{map}}; F_{\mathrm{img}}^{\mathrm{map}}] \in \mathbb{R}^{(L' + L_{\mathrm{img}}) \times D_{\mathrm{C}}}, \tag{5}$$

which is then fed into the DiT to guide the denoising process with both semantic and visual cues.

3.3 Diffusion Transformer

The Diffusion Transformer primarily comprises three components: a patchifying module, the Transformer blocks, and an unpatchifying module. Inside each Transformer block (as depicted in Fig.4), the model focuses on effectively modeling spatio-temporal contextual relationships while simultaneously embedding both the text conditions and the time steps. We employ the Cross-Attention mechanism to embed the input text conditions, ensuring the model's ability to follow instructions even under long-context modeling scenarios. Additionally, the model utilizes an MLP, which includes a Linear layer and a SiLU (Elfwing et al., 2018) activation function, to process the input time embeddings and predict six individual modulation parameters. This specific MLP is shared across all Transformer blocks, with each block learning a distinct set of biases for specialized regulation.

4 Training Methods

We first introduce the fundamentals of flow matching in Sec. 4.1. Then, Sec. 4.2 presents the proposed Latent Shifting strategy, which implicitly integrates conditional information based on flow matching characteristics. Finally, Sec. 4.3 describes the Fourier Guidance module, which supplements the denoiser with high-frequency magnitude features extracted via the Fourier Transform.

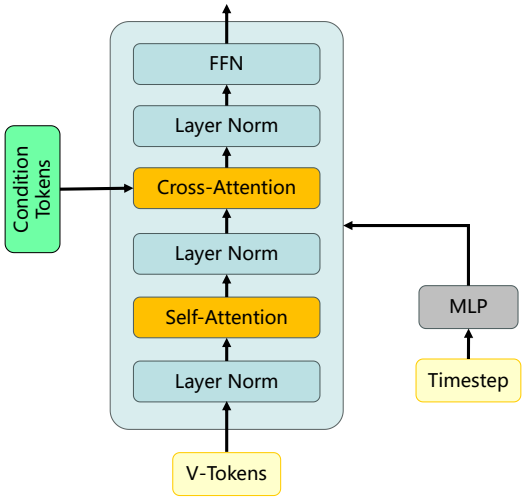


Figure 4: Transformer Block of DiT.

4.1 Preliminary for Flow Matching

Continuous Normalizing Flows (CNFs) (Chen, 2018) learn a transformation from a sample $\boldsymbol{z}_1$ drawn from a source distribution $q_1(\boldsymbol{z})$ to a target sample $\boldsymbol{z}_0$ from $q_0(\boldsymbol{z})$, where q_0 denotes the data distribution and q_1 is typically a known prior, such as a standard Gaussian. This transformation is modeled as an ordinary differential equation (ODE) over $t \in [0, 1]$:

$$\frac{d\boldsymbol{z}_t}{dt} = \boldsymbol{v}_t(\boldsymbol{z}_t, t), t \in [0, 1]. \tag{6}$$

Here, $\boldsymbol{v}_t(\boldsymbol{z}_t, t)$ defines the velocity field that governs how the distribution evolves over time.

Flow Matching (FM) (Lipman et al., 2022; Liu, 2022; Tong et al., 2023) learns this vector field $\boldsymbol{v}_t(\boldsymbol{z}_t, t)$ directly using a neural network $\boldsymbol{v}_\theta(\boldsymbol{z}_t, t)$. Given $\boldsymbol{z}_1 \sim q_1$ and $\boldsymbol{z}_0 \sim q_0$, their linear interpolation is:

$$\boldsymbol{z}_t = (1 - t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, t \in [0, 1]. \tag{7}$$

The corresponding velocity field is:

$$\boldsymbol{v}_t(\boldsymbol{z}_t, t) = \frac{d\boldsymbol{z}_t}{dt} = \boldsymbol{z}_1 - \boldsymbol{z}_0. \tag{8}$$

Thus, $\boldsymbol{v}_t(\boldsymbol{z}_t, t)$ depends only on $\boldsymbol{z}0$ and $\boldsymbol{z}1$, independent of t. FM trains $\boldsymbol{v}\theta(\boldsymbol{z}t, t)$ to approximate this velocity field via an MSE loss. Under a condition $\boldsymbol{y}$, the model becomes Conditional Flow Matching (CFM), with the objective:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\boldsymbol{z}_0\sim q_0,\boldsymbol{z}_1\sim q_1}\left[\|\boldsymbol{v}_\theta((1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, t, \boldsymbol{y}) - (\boldsymbol{z}_1 - \boldsymbol{z}_0)\|_2^2\right] \tag{9}$$

where $\mathcal{U}[0, 1]$ denotes the uniform distribution. During sampling, $\boldsymbol{z}_1 \sim q_1$ is drawn, and the ODE $\frac{d\boldsymbol{z}t}{dt} = \boldsymbol{v}\theta(\boldsymbol{z}_t, t, \boldsymbol{y})$ is solved from $t = 1$ to 0 to obtain $\boldsymbol{z}_0 \sim q_0$. Unlike DDPMs (Ho et al., 2020), FM imposes no constraint on the source distribution, enabling transformations between arbitrary distributions.

4.2 Latent Shifting

We implement image-to-video (I2V) generation without explicitly embedding the full conditional image into the denoiser. Let $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and define a conditional image $\boldsymbol{S} \in \mathbb{R}^{c\times h\times w}$ and its target video $\boldsymbol{X} \in \mathbb{R}^{c\times t\times h\times w}$, where $\boldsymbol{X}[:, 0] = \boldsymbol{S}$. Let $\mathcal{E}$ denote the VAE encoder (Kingma and Welling, 2013). For the T2V task, define $\boldsymbol{z}_1^T = \boldsymbol{\epsilon}$, $\boldsymbol{z}_0^T = \boldsymbol{x} = \mathcal{E}(\boldsymbol{X})$, and the intermediate state:

$$\boldsymbol{z}_t^T = (1 - t)\boldsymbol{z}_0^T + t\boldsymbol{z}_1^T = (1 - t)\boldsymbol{x} + t\boldsymbol{\epsilon}. \tag{10}$$

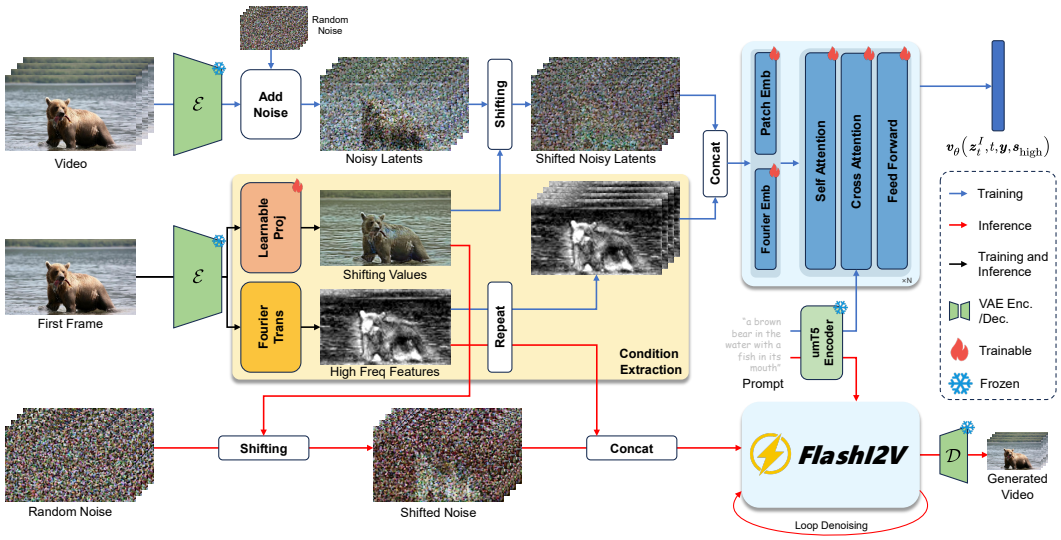


Figure 5: **FlashI2V.** Conditional image latents are first projected and shifted to form an intermediate representation that implicitly encodes the conditioning information. Meanwhile, the Fourier-transformed high-frequency magnitudes of the image are concatenated with noisy latents and fed into the DiT backbone. During inference, denoising starts from the shifted noise and follows the ODE trajectory until the final video is reconstructed.

Its velocity field is $\boldsymbol{v}_t^T(\boldsymbol{z}_t^T, t) = \frac{d\boldsymbol{z}_t^T}{dt} = \boldsymbol{\epsilon} - \boldsymbol{x}$. For I2V, since FM allows arbitrary source and target distributions, we redefine them to implicitly encode conditions while avoiding image leakage:

$$\boldsymbol{z}_1^I = \alpha \boldsymbol{s} + \beta \boldsymbol{\epsilon}, \tag{11}$$

$$\boldsymbol{z}_0^I = \gamma \boldsymbol{s} + \kappa \boldsymbol{x}, \tag{12}$$

where $\boldsymbol{s} = \mathcal{E}(\boldsymbol{S})$ and $\alpha, \beta, \gamma, \kappa$ are constants. The intermediate state and velocity field become:

$$\boldsymbol{z}_t^I = (1-t)\boldsymbol{z}_0^I + t\boldsymbol{z}_1^I = \kappa \boldsymbol{z}_t^T + [\gamma + (\alpha - \gamma)t]\boldsymbol{s} + (\beta - \kappa)t\boldsymbol{\epsilon}, \tag{13}$$

$$\boldsymbol{v}_t^I(\boldsymbol{z}_t^I, t) = \frac{d\boldsymbol{z}_t^I}{dt} = \kappa \boldsymbol{v}_t^T(\boldsymbol{z}_t^T, t) + (\alpha - \gamma)\boldsymbol{s} + (\beta - \kappa)\boldsymbol{\epsilon}. \tag{14}$$

When $\alpha = \gamma$ and $\beta = \kappa = 1$, we obtain $\boldsymbol{v}_t^I(\boldsymbol{z}_t^I, t) = \boldsymbol{v}_t^T(\boldsymbol{z}_t^T, t)$, making the I2V and T2V objectives identical. Then:

$$\boldsymbol{z}_t^I = \boldsymbol{z}_t^T + \gamma \boldsymbol{s}. \tag{15}$$

If $\gamma = 0$, I2V degenerates to T2V; if $\gamma \neq 0$, the conditional image influences the denoiser via latent shifting, equivalent to shifting by $-\gamma \boldsymbol{s}$.

To adaptively weight $\boldsymbol{s}$, we replace $-\gamma \boldsymbol{s}$ with a learnable projection $\boldsymbol{\phi}(\boldsymbol{s})$, initialized to zero to preserve the initial latent distribution. Since $\boldsymbol{\phi}(\cdot)$ is independent of t, we maintain $\boldsymbol{v}_t^I = \boldsymbol{v}_t^T$. The final I2V formulation is:

$$\boldsymbol{z}_1^I = \boldsymbol{\epsilon} - \boldsymbol{\phi}(\boldsymbol{s}), \tag{16}$$

$$\boldsymbol{z}_0^I = \boldsymbol{x} - \boldsymbol{\phi}(\boldsymbol{s}), \tag{17}$$

$$\boldsymbol{z}_t^I = (1-t)\boldsymbol{x} + t\boldsymbol{\epsilon} - \boldsymbol{\phi}(\boldsymbol{s}), \tag{18}$$

$$\boldsymbol{v}_t^I(\boldsymbol{z}_t^I, t) = \boldsymbol{\epsilon} - \boldsymbol{x}. \tag{19}$$

4.3 Fourier Guidance

In latent shifting, recovering $\boldsymbol{s}$ from the mixture of $\boldsymbol{\epsilon}$ and $\boldsymbol{\phi}(\boldsymbol{s})$ is challenging, especially for high-frequency details like edges and textures. To address this, we introduce additional high-frequency guidance. Since the VAE's latent space preserves the spectral structure of the image, we extract high-frequency magnitude components via a Fourier Transform:

$$\boldsymbol{s}_{\text{high}} = \boldsymbol{f}_{\text{high}}(\boldsymbol{s}), \tag{20}$$

where $\boldsymbol{f}_{\text{high}}$ denotes a high-frequency magnitude filter. $\boldsymbol{s}_{\text{high}}$ is concatenated with $\boldsymbol{z}_t^I$ and embedded as:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{W}^I & \boldsymbol{W}^F \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_t^I \\ \boldsymbol{s}_{\text{high}} \end{bmatrix}, \tag{21}$$

where $\boldsymbol{W}^I$ and $\boldsymbol{W}^F$ are the patch embedding layers for $\boldsymbol{z}_t^I$ and $\boldsymbol{s}_{\text{high}}$, respectively. $\boldsymbol{W}^F$ is zero-initialized to keep the hidden state distribution stable at training start.

The final training objective is:

$$\mathcal{L}_{\text{Flash}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),\boldsymbol{X}\sim q(\boldsymbol{X}),\boldsymbol{x}=\mathcal{E}(\boldsymbol{X}),\boldsymbol{s}=\mathcal{E}(\boldsymbol{X}[:,0])} \left[\left\| \boldsymbol{v}_\theta^I((1-t)\boldsymbol{x} + t\boldsymbol{\epsilon} - \boldsymbol{\phi}(\boldsymbol{s}), t, \boldsymbol{y}, \boldsymbol{s}_{\text{high}}) - (\boldsymbol{\epsilon} - \boldsymbol{x}) \right\|_2^2 \right], \tag{22}$$

where $\boldsymbol{y}$ denotes the text embedding and $\boldsymbol{v}_\theta^I$ is the denoiser excluding $\boldsymbol{\phi}$.

5 FlashI2V Experiment

In this section, we first describe the FlashI2V experimental setup in Sec. 5.1. Next, Sec. 5.2 presents both quantitative and qualitative comparisons with other image-to-video (I2V) approaches. In Sec. 5.3, we conduct ablation studies to validate the effectiveness of each component in FlashI2V. Finally, Sec. 5.4 provides an in-depth analysis of the roles and behaviors of the key modules within FlashI2V.

5.1 Experimental Setup

Training Setup. For fair comparison, we train our model for 40K steps on 2.5M internally collected high-quality videos, following the data collection and preprocessing pipeline described in Open-Sora Plan (Lin et al., 2024b). Each video clip contains 49 randomly sampled frames at a fixed frame rate of 16 fps and a resolution of 480×832. The model is initialized from Wan2.1-T2V-14B (Wan et al., 2025). The learnable projection module consists of two Conv3D (Tran et al., 2015) layers with SiLU (Elfwing et al., 2018) activations, while the Fourier embedding layer is implemented in the same way as the patch embedding and initialized to zero. During training, the first frame of each video is used as the conditional image. The cutoff frequency percentile for the Fourier Transform is uniformly sampled from $\mathcal{U}[0.05, 0.95]$. The text prompt is randomly dropped with a probability of 0.1. We use a batch size of 64, a learning rate of 2e-5, weight decay of 1e-2, and the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon =$ 1e-15. Exponential Moving Average (EMA) with a decay rate of 0.9999 is applied to stabilize training. For the ablation experiments, we train the models with initialization of Wan2.1-T2V-1.3B. We train on a 2M subset with 30K training steps, keeping all other settings unchanged.

Sampling Setup. For sampling, we employ the Discrete Euler Sampler with the sigma-shifting strategy introduced in HunyuanVideo (Kong et al., 2024), using a shifting coefficient of 7.0, classifier-free guidance scale of 5.0, 50 sampling steps, and a cutoff frequency percentile fixed at 0.1.

Evaluation. For evaluation, we generate 49-frame videos at the default resolution and report all VBench-I2V (Huang et al., 2024; Zheng et al., 2025) metrics except for Camera Motion. To obtain more reliable scores, we reception the short prompts from VBench-I2V using ChatGPT (Achiam et al., 2023). For the ablation study, we randomly sample 1,000 videos from the HD subset of OpenVid-1M (Nan et al., 2024) as the validation set and compute the chunk-wise FVD under each experimental setting.

Table 1: **VBench-I2V Results.** Quantitative comparison of various approaches on the VBench-I2V benchmark. All metrics are reported as percentages. † denotes evaluation using recaptioned image-text pairs.

Model	I2V Paradigm	Subject Consistency↑	Background Consistency↑	Motion Smoothness↑	Dynamic Degree↑	Aesthetic Quality↑	Imaging Quality↑	I2V Subject Consistency↑	I2V Background Consistency↑
SVD-XT-1.0 (1.5B)	Repeating Concat and Adding Noise	95.52	96.61	98.09	52.36	60.15	69.80	97.52	97.63
SVD-XT-1.1 (1.5B)	Repeating Concat and Adding Noise	95.42	96.77	98.12	43.17	60.23	70.23	97.51	97.62
SEINE-512x512 (1.8B)	Inpainting	95.28	97.12	97.12	27.07	64.55	**71.39**	97.15	96.94
CogVideoX-5B-I2V	Zero-padding Concat and Adding Noise	94.34	96.42	98.40	33.17	61.87	70.01	97.19	96.74
Wan2.1-I2V-14B-720P	Inpainting	94.86	97.07	97.90	51.38	64.75	70.44	96.95	96.44
CogVideoX1.5-5B-I2V†	Zero-padding Concat and Adding Noise	95.04	96.52	**98.47**	37.48	<u>62.68</u>	**70.99**	97.78	98.73
Wan2.1-I2V-14B-480P†	Inpainting	<u>95.68</u>	<u>97.44</u>	<u>98.46</u>	<u>45.20</u>	61.44	70.37	<u>97.83</u>	**99.08**
Uniworld-OSP2.0† (14B)	**FlashI2V**	**96.21**	**97.71**	**98.47**	**46.10**	**66.55**	<u>70.57</u>	**97.99**	<u>98.94</u>

Figure 6: **Method Comparison.** Quantitative results comparing FlashI2V with CogVideoX1.5-5B-I2V (Yang et al., 2024) and Wan2.1-I2V-14B-480P (Wan et al., 2025). Both CogVideoX1.5 and Wan2.1 show color inconsistency, and Wan2.1 further suffers from overly slow or static motion. Benefiting from the prevention of conditional image leakage, FlashI2V effectively eliminates these artifacts and achieves more stable video generation.

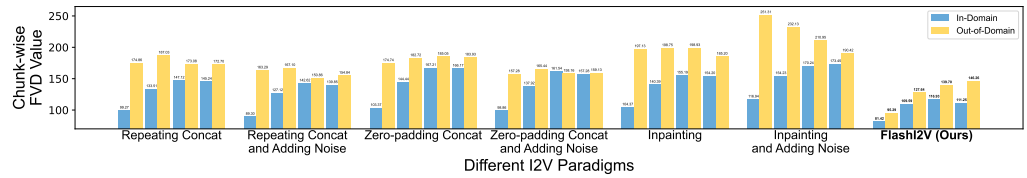
5.2 Main Results

Quantitative Results. Table 1 presents a quantitative comparison of different methods on VBench-I2V. By effectively preventing conditional image leakage, FlashI2V achieves a substantially higher dynamic degree score than all competing methods. For other metrics, FlashI2V performs comparably to CogVideoX1.5-5B-I2V and Wan2.1-I2V-14B-480P.

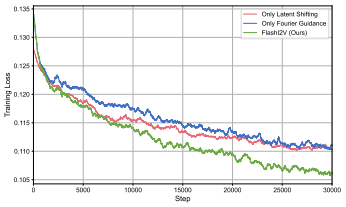
Qualitative Results. Figure 6 illustrates qualitative comparisons. Conditional image leakage causes CogVideoX1.5 and Wan2.1 to suffer from artifacts such as color inconsistency and slow motion, with Wan2.1 occasionally generating completely static videos. In contrast, FlashI2V produces videos with more pronounced motion while maintaining physically plausible dynamics.

5.3 Ablation Study

Generalization to Out-of-Domain Data. We analyze how different I2V paradigms generalize by examining chunk-wise FVD variations on both in-domain and out-of-domain datasets (Fig. 7a). Each generated video is divided into four temporal chunks of equal length. Unlike FlashI2V, most other methods feed the full conditional image into the noisy latents, and the "Adding Noise" variants inject small noise into the conditional image latents following CogVideoX (Yang et al., 2024). Our results show that only FlashI2V maintains consistent FVD patterns across both data domains, indicating robust generalization. Other paradigms display large discrepancies between in-domain and out-of-domain behavior, highlighting the impact of conditional image leakage. Additionally, FlashI2V



(a) Chunk-wise FVD Variation Patterns Across Various I2V Paradigms



(b) Training Loss

(c) Qualitative Results

Figure 7: **Ablation Study**. (a) Comparing the chunk-wise FVD variation patterns of different I2V paradigms on both the training and validation sets, it is observed that only FlashI2V exhibits the same time-increasing FVD variation pattern in both sets. This suggests that only FlashI2V is capable of applying the generation law learned from in-domain data to out-of-domain data. Additionally, FlashI2V has the lowest out-of-domain FVD, demonstrating its performance advantage. (b) From the training loss, we can observe that Fourier guidance accelerates the convergence of latent shifting. (c) Fourier guidance alone causes color deviation, while latent shifting alone leads to mismatched details. FlashI2V achieves consistency in both color and details.

achieves the lowest FVD on out-of-domain data, confirming its superior performance and stability among all tested paradigms.

Impact of Latent Shifting and Fourier Guidance. We further study the contributions of FlashI2V's core modules through ablation experiments. Quantitatively, incorporating Fourier guidance accelerates training convergence compared to using latent shifting alone (Fig. 7b), suggesting it effectively enhances latent optimization.

Qualitative analysis (Fig. 7c) reveals complementary roles of the two modules. Using only Fourier guidance preserves high-frequency details but fails to reproduce accurate colors, while latent shifting alone maintains global consistency but lacks fine local fidelity. By combining both, FlashI2V achieves videos that are faithful both in global structure and local detail.

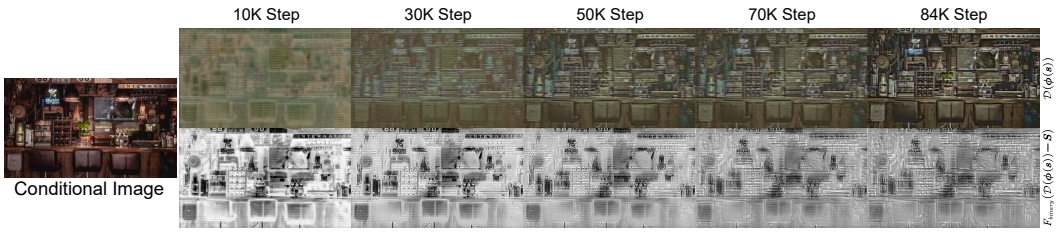
5.4 Analysis

Latent Shifting Encodes Rich Features. The latent shifting operation $\phi(\cdot)$ transforms the original latents into a space where the encoded features carry meaningful structural and high-frequency information. To visualize this effect, we decode $\phi(\boldsymbol{s})$ back to pixel space using the VAE and compute the relative differences between $\mathcal{D}(\phi(\boldsymbol{s}))$ and the original video frames $\boldsymbol{S}$, presenting the results as binary maps (Fig. 8a). Over the course of training, $\phi(\cdot)$ progressively captures richer representations, with high-frequency details increasingly emphasized compared to the original latent $\boldsymbol{s}$. This evolution directly contributes to the model's improved fidelity and more precise reconstructions as training advances.

Controllable Generation Through Fourier Guidance. Fourier guidance enables fine-grained control over the level of detail in generated videos by adjusting the cutoff frequency percentile. As illustrated in Fig. 8b, lowering the cutoff percentile injects more high-frequency components into the generation process, enhancing the clarity of small-scale structures such as text and fine textures. Consequently, this mechanism allows the model to preserve intricate details consistently across the entire video while providing a controllable trade-off between global coherence and local fidelity.

5.5 Visual Understanding

Leveraging the design in which the Multimodal Large Language Model component is kept frozen, our method retains the strong multimodal comprehension abilities of Qwen2.5-VL-7B without requiring any further fine-tuning. This strategy substantially cuts down the demand for training data and



(a) Encoded Features of Latent Shifting

(b) The Influence of Fourier Guidance on Generation Details

Figure 8: **Analysis of latent shifting and fourier guidance**. (a) As training progresses, $\phi(\cdot)$ gradually emphasizes the detailed information in the conditional image. (b) When a lower cutoff frequency percentile is used, more high-frequency information is injected. When the cutoff frequency percentile is set to 0.1, the graphical text at the end of the video remains unchanged, while with the cutoff frequency percentile set to 0.9, the graphical text becomes unrecognizable.

computational resources, while also avoiding the degradation in understanding performance that often appears when models are trained on generative tasks. As shown in Table 2, this architectural design allows UniWorld-Osp2 to deliver outstanding results—surpassing models such as Janus, Show-o, and Emu3 across a range of evaluation metrics—and exhibiting performance that is highly competitive with more recent state-of-the-art systems like BAGEL.

Table 2: **Comparison between different models on Visual Understanding benchmarks.** × indicates the model is incapable of performing the task.

Model	MMB^V	MMB^I	MMMU	MM-vet
Image & Video Understanding				
LLaVA-1.5	×	36.4	67.8	36.3
Video-LLaVA	1.05	60.9	32.8	32.0
Unified Understanding & Generation				
Show-o	×	-	27.4	-
Janus	×	69.4	30.5	34.3
Janus-Pro	×	75.5	36.3	39.8
Emu3	-	58.5	31.6	37.2
BLIP3-o	-	83.5	50.6	66.6
MetaQuery	-	83.5	**58.6**	66.6
BAGEL	-	**85.0**	55.3	**67.2**
GPT-4o	2.15	×	72.9	76.9
UniWorld-OSP2.0	**1.79**	83.5	**58.6**	67.1

6 The Styled Image and Video Dataset

In this section, we describe the construction of our styled image and video dataset, which contains diverse artistic style transformations and corresponding text descriptions. As shown in Fig. 9, the dataset construction process consists of three main steps: source video collection and preprocessing (Section 6.1), keyframe extraction with style generation (Section 6.2), and video synthesis from styled images (Section 6.3). Through this pipeline, we extract keyframes from original short videos,

leverage advanced stylization models to generate images in various artistic styles, and synthesize them into styled videos.

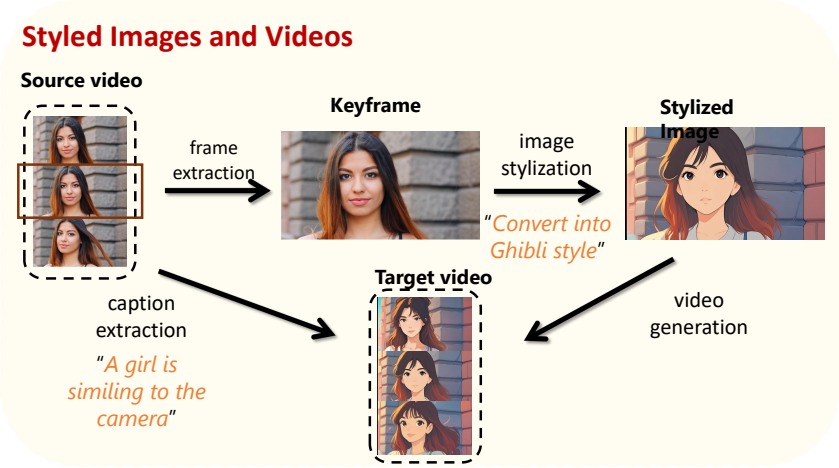


Figure 9: **Illustration of the style data construction process**, including key frame and caption extraction, image stylization, and video generation.

6.1 Source Video Collection and Preprocessing

To construct a styled image dataset with rich visual content and high-quality text descriptions, we collect short videos from multiple video platforms as source materials.

Video selection criteria: During the data collection process, we implement a strict quality control mechanism. First, we prioritize short videos with durations between 5-20 seconds, as this duration ensures scene coherence and content integrity. Shorter videos typically exhibit more focused themes and clearer visual expressions, making them more suitable as inputs for style transformation. Second, we filter videos by resolution, retaining only high-quality videos with 720p or higher resolution to ensure the effectiveness of subsequent stylization processing. Third, we filter videos by aspect ratio, selecting videos with appropriate aspect ratios that are suitable for standard display formats, which helps maintain visual consistency across the dataset.

Video quality assessment: To ensure dataset quality, we employ a human sampling approach for quality assessment. A subset of collected videos is randomly sampled and manually evaluated by human annotators to verify content quality, visual clarity, and appropriateness for stylization. Videos that do not meet our quality standards are excluded from the dataset. Through this rigorous screening process, we ultimately obtain a high-quality, content-rich source video dataset.

6.2 Keyframe Extraction and Style Generation

Keyframe extraction and style generation are the core steps in dataset construction. In this stage, we extract representative frames from source videos and leverage advanced image stylization models to generate diverse artistic style images.

Keyframe extraction strategy: To capture the core visual information of videos, we adopt a middle-frame extraction strategy. Specifically, for each source video, we calculate its total frame count and extract the keyframe located at the midpoint of the timeline as the representative image. This strategy is based on the following observation: the middle frame of short videos typically contains the most important visual information and best represents the thematic content of the entire video. Compared to random sampling or scene-change-based keyframe extraction methods, the middle-frame extraction strategy is computationally simple and stable, ensuring semantic consistency between the extracted keyframes and video captions.

Stylization model application: We employ Step1X-Edit Liu et al. (2025b) as the core stylization generation engine. Step1X-Edit is an advanced open-source image editing model that provides powerful image-to-image transformation capabilities. While originally designed for general image editing tasks, we leverage its strong image transformation abilities to perform artistic style transfer. In our dataset, we implement 12 distinct artistic styles, including: Ghibli (Studio Ghibli animation style),

3D rendering, Cyberpunk, Ink painting, Oil painting, Pixel Art, Vaporwave, Disney (Disney animation style), Fairy Tale, Gotham Noir (film noir aesthetic), Lego (brick style), and Chibi (super-deformed style). For each keyframe, we apply specific style transformations to generate high-quality images.

The stylization process is formulated as follows: given a keyframe I_{key} and target style description s, the Step1X-Edit model generates a styled image I_{style}:

$$I_{style} = \text{Step1X-Edit}(I_{key}, s)$$

where the style description s comes from a predefined style template library containing detailed artistic style instructions. To ensure generation quality and style diversity, we configure specific model parameters for each style, including style strength and detail preservation hyperparameters.

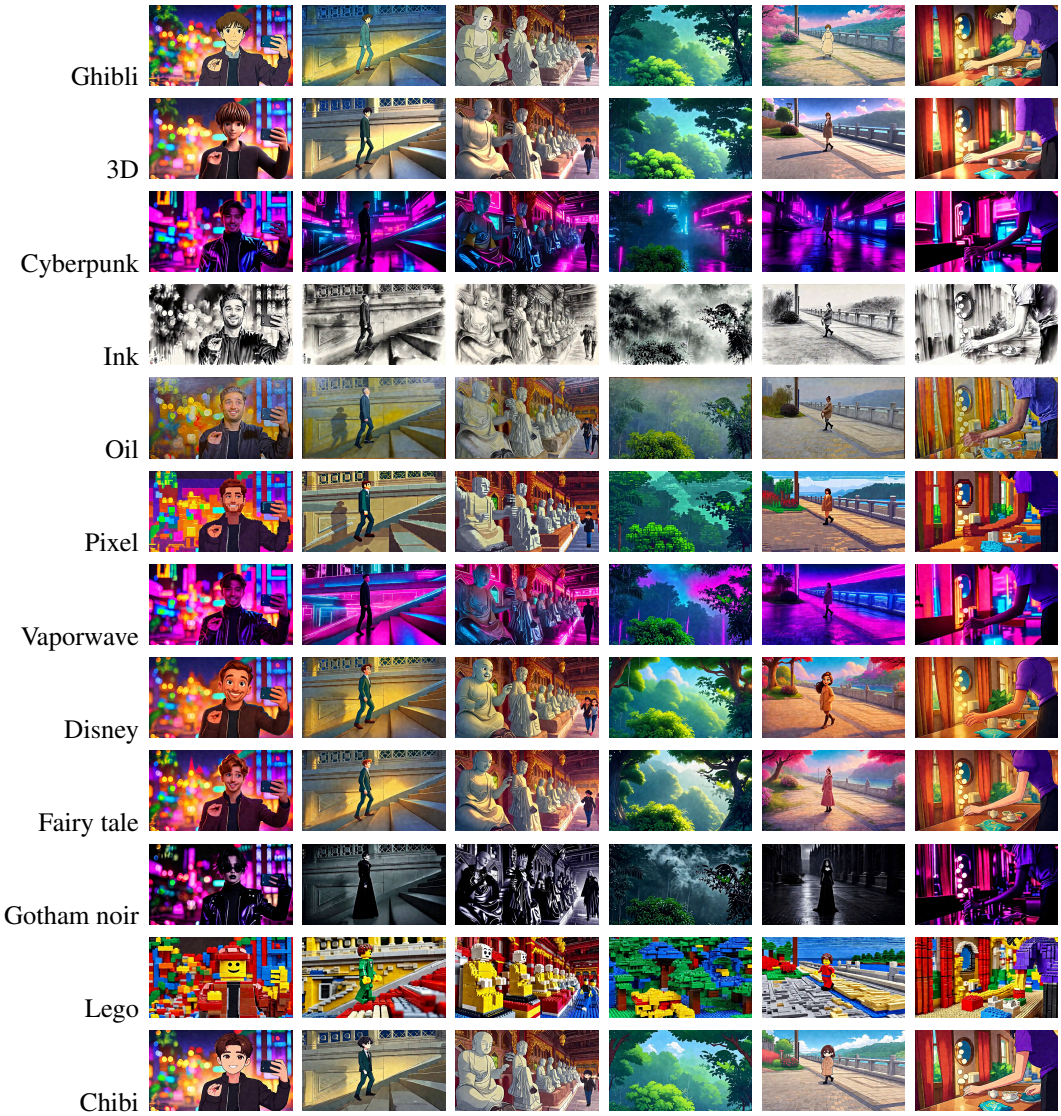


Figure 10: **Illustration of visual examples from 12 different styles in styled image dataset.** Each row corresponds to a specific style, and each column shows sampled images across different instances.

Style diversity guarantee: To construct a dataset with rich style diversity, we carefully design a style library containing 12 distinct artistic styles. These styles cover multiple artistic categories: (1) *Animation styles*, including Ghibli, Disney, Chibi, and Fairy Tale, encompassing Japanese animation, American animation, and fairytale illustration aesthetics; (2) *Traditional art styles*, including Oil and Ink, representing classic Western oil painting and Eastern ink wash painting forms; (3) *Modern digital styles*, including 3D rendering, Pixel Art, Lego, and Vaporwave, reflecting the diversity of contemporary digital art; (4) *Thematic styles*, including Cyberpunk and Gotham Noir, showcasing

science fiction futurism and film noir aesthetics. Each style is carefully calibrated to ensure artistic expressiveness and visual consistency of the generated results. The visual examples of all 12 styles are shown in Fig. 10.

Quality control and post-processing: Generated styled images undergo rigorous quality inspection. We employ a multi-dimensional quality assessment mechanism, including: (1) style consistency assessment to ensure generated images genuinely embody the target artistic style characteristics; (2) content preservation assessment to ensure the main content and structure of original images are retained during stylization; (3) visual quality assessment to filter images with obvious artifacts or distortions. Generated results that do not meet quality standards are discarded or regenerated.

Dataset statistics: Through the above pipeline, we construct a large-scale styled image dataset. Table 3 presents the detailed statistics of each style category in the dataset.

Table 3: **Styled image dataset statistics.** The number of images per visual style is summarized.

Style	# Images	Style	# Images	Style	# Images
Ghibli	∼200,000	3D Rendering	∼200,000	Chibi	∼20,000
Cyberpunk	∼20,000	Ink Painting	∼20,000	Oil Painting	∼20,000
Pixel Art	∼20,000	Vaporwave	∼20,000	Disney	∼20,000
Fairy Tale	∼20,000	Gotham Noir	∼20,000	Lego	∼20,000
Total ∼600,000 images					

The dataset exhibits the following notable characteristics: (1) **Large scale**: containing approximately 600,000 styled images covering 12 distinct artistic styles; (2) **Style diversity**: encompassing multiple artistic categories including animation, traditional art, digital art, and thematic styles, with Ghibli and 3D rendering each containing around 200,000 images and the other 10 styles each containing approximately 20,000 images, ensuring dataset balance; (3) **Content richness**: source videos cover diverse visual categories including people, landscapes, animals, and objects, providing extensive training samples for stylization across different scenes; (4) **High-quality annotations**: each image is accompanied by detailed text descriptions, including original video captions and style labels, providing rich semantic information for multimodal learning.

These characteristics make the dataset an ideal resource for training and evaluating styled image generation models, particularly suitable for tasks such as text-guided artistic style transfer and multi-style image generation, providing strong support for multimodal generation research.

6.3 Video Synthesis from Styled Images

After obtaining high-quality styled images across various artistic domains, the final stage of our pipeline is transforming these static images into temporally coherent styled videos. This step enables the styled image dataset to extend beyond frame-level stylization, providing dynamic video content suitable for multimodal generation and temporal consistency.

Image-to-video generation: We employ Wan2.1-I2V-14B-720P-Diffusers Wan et al. (2025), a large-scale image-to-video diffusion model as the core engine for video synthesis. This model supports high-fidelity video generation at 720p resolution and demonstrates strong temporal stability and motion realism. Unlike traditional frame interpolation or optical-flow-based synthesis methods, Wan2.1-I2V directly models spatiotemporal correlations through diffusion-based video generation, allowing it to produce smooth, style-consistent motion while preserving both the semantic and artistic attributes of the input image.

Given a styled image I_{style} and its corresponding textual description T, the model synthesizes a video sequence V_{style} as:

$$V_{\text{style}} = \text{Model}_{\text{i2v}}(I_{\text{style}}, T)$$

where the text prompt T provides high-level contextual guidance, such as motion dynamics, scene semantics, and camera movement, ensuring that the generated motion aligns with the artistic and narrative intent of the source material.

Generation configuration: In our synthesis pipeline, each generated video clip lasts 5 seconds with a frame rate of 24 fps, providing a good balance between temporal smoothness and generation efficiency. Through this synthesis process, we generate over 600,000 high-quality styled video clips covering all 12 visual styles introduced in Section 6.2. Each clip is paired with its textual description

and corresponding style label, providing rich multimodal supervision for downstream tasks. The video samples corresponding to different styles are shown in Fig. 11.

The resulting styled videos exhibit: (1) **Temporal coherence**: smooth and stable motion across frames; (2) **Artistic consistency**: faithful preservation of the input style across the full temporal sequence; (3) **Semantic alignment**: motion patterns aligned with textual guidance and original scene.

Figure 11: **Illustration of visual examples from 12 different styles in styled video dataset.** Each row shows a specific style, represented by six frames sampled uniformly from the generated video.

6.4 Training Details

Training Setup: We train our model in a multi-node environment using PyTorch FSDP with one GPU per node and BF16 mixed precision. Gradient checkpointing and low-VRAM optimizations are enabled to reduce memory usage. All images and videos are resized to 512×512, and videos are uniformly sampled to 49 frames with a stride of 1. We use a batch size of 1 per GPU, uniform

sampling, and random aspect-ratio adaptation. The transformer3d backbone is fine-tuned with LoRA (rank 8, $\alpha = 4$) under an inpainting-style training mode. Video-token length conditioning and bucketed training are enabled. We adopt AdamW with a learning rate of 1×10^{-4}, weight decay of 3×10^{-2}, and $\epsilon = 10^{-10}$. Training is performed for 1 epochs with a maximum gradient norm of 0.05. A fixed seed (42) is used for reproducibility.

Sampling Setup: During inference, we enable TeaCache to accelerate sampling, using a cache threshold of 0.10 and skipping the first 5 denoising steps to minimize quality degradation. TeaCache offloading is disabled to avoid CPU transfer overhead. Classifier-free guidance skipping is not applied (*cfg_skip_ratio* = 0). We adopt the *Flow_UniPC* sampler with a noise-schedule shift of 3, which is suitable for 480p–720p video generation. Riflex enhancement is disabled in all experiments. We load the transformer3d-14B model with LoRA and adapter weights fine-tuned on our dataset. All videos are generated at a spatial resolution of 384×672, with a temporal length of 49 frames at 16 fps. Inference is performed in `bfloat16`.

Training Results: Building upon the synthesized dataset and training pipeline described above, our trained model is able to perform high-quality multi-style video generation, achieving 13 distinct visual effects. This demonstrates that the model not only fits the provided style distribution but also generalizes to produce additional stylistic variations. Through extensive experiments, we observe that the model achieves stable and reliable inference across all styles. Overall, these results validate the effectiveness of our training strategy and demonstrate that the model possesses high-quality style transfer, controllable generation, and strong multimodal alignment, establishing a solid foundation for downstream creative and visual understanding tasks.

7 Extended Application

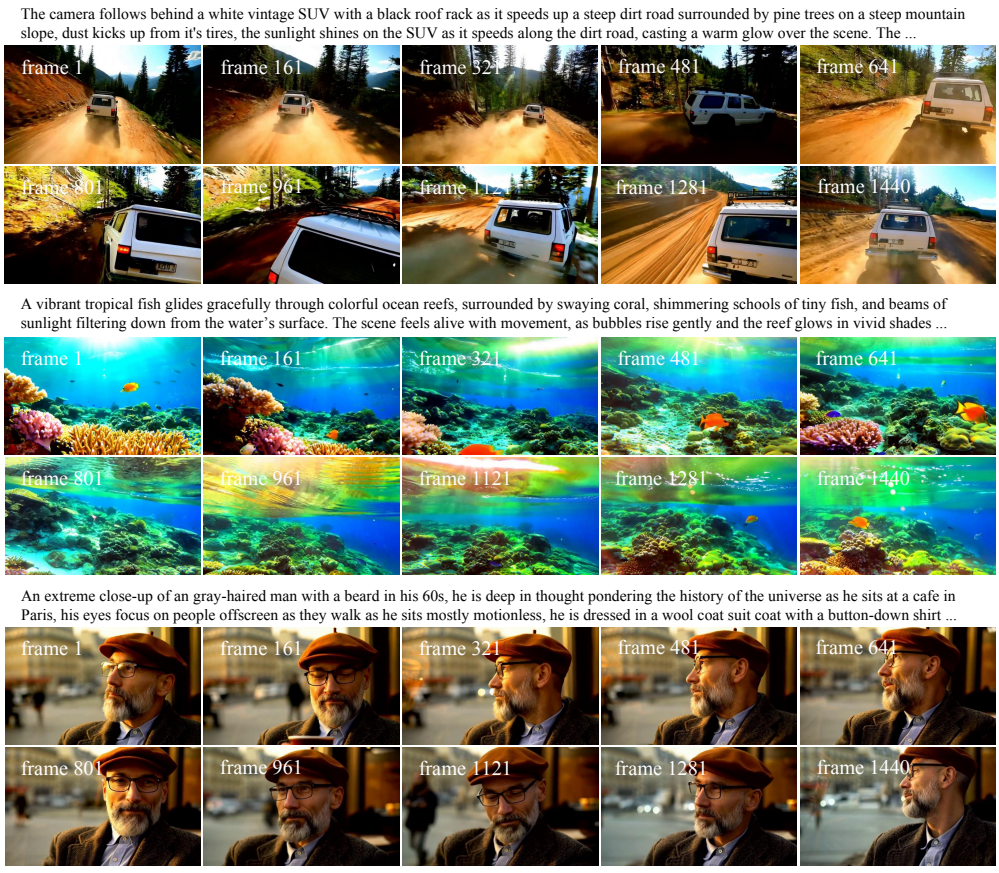


Figure 12: **Showcases of Real-Time Infinity Video Generation.**

7.1 Real-Time Infinity Video Generation

Current video generation methods face significant efficiency bottlenecks. Even with high-end hardware, generating just a few seconds of video often requires tens of minutes of processing time, severely limiting the practical applicability of the technology. While generation quality has improved substantially in recent years, the lengthy processing time remains a critical barrier to adoption in real-world workflows, and existing models are generally limited to generating only 5-10 second videos. To overome these limitations, we trained a real-time video generation model with $10\times$ larger scale than existing models—OSP-RealTime 14B. This model achieves a text-to-video inference speed of 10fps on a single NVIDIA H100 GPU. Figure 12 demonstrates examples of real-time long video generation. Building upon Wan2.1-T2V Wan et al. (2025), we achieved this breakthrough by:

(1) For Infinite Video Generation Strategy. Unlike existing methods that typically employ a sliding window mechanism combined with causal masking to convert bidirectional models into autoregressive ones Huang et al. (2025); Yin et al. (2025); Cui et al. (2025); Yang et al. (2025); Liu et al. (2025a), we reformulate long video generation as an infinite video continuation task. Specifically, we concatenate previously generated video frames as context with the current noise latent along the temporal dimension. This strategy maximally preserves the same inference paradigm as the pretrained model, thereby achieving a higher quality floor compared to conventional approaches.

(2) For Real-time Generation Acceleration. To enable real-time video generation, we implement three key optimizations. First, we reduce the number of noise latent frames from 21 to 9, meaning each continuation generates 9 frames. Since the computational complexity of attention mechanisms is $\mathcal{O}(n^2)$, this adjustment significantly reduces the computational overhead per forward pass. Second, we break away from the traditional approach of maintaining full resolution across all sampling stages. Instead, we partition the inference process into multiple stages with progressively increasing resolution, following a pipeline where $\text{res}_1 < \text{res}_2 < \ldots < \text{res}_n$, further reducing inference overhead. Finally, we apply DMD Yin et al. (2024a,b) to compress the inference steps from 50 to 4. Through the synergistic combination of these optimization strategies, we successfully achieve real-time video generation on 14B model.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Kling AI, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512, 2023a.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7310–7320, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. arXiv preprint arXiv:2406.04325, 2024b.

Ricky T. Q. Chen. torchdiffeq, 2018.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In The Twelfth International Conference on Learning Representations, 2023b.

June Suk Choi, Kyungmin Lee, Sihyun Yu, Yisol Choi, Jinwoo Shin, and Kimin Lee. Enhancing motion dynamics of image-to-video models via adaptive low-pass guidance. arXiv preprint arXiv:2506.08456, 2025.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283, 2025.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural networks, 107:3–11, 2018.

Yunyang Ge, Xinhua Cheng, Chengshu Zhao, Xianyi He, Shenghai Yuan, Bin Lin, Bin Zhu, and Li Yuan. Flashi2v: Fourier-guided latent shifting prevents conditional image leakage in image-to-video generation. arXiv preprint arXiv:2509.25187, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21807–21818, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023.

Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan. Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 17778–17788, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122, 2023.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131, 2024a.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131, 2024b.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. arXiv preprint arXiv:2401.15947, 2024c.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161, 2025a.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. arXiv preprint arXiv:2209.14577, 2022.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025b.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048, 2024.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371, 2024.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020b.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. arXiv preprint arXiv:2302.00482, 2023.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision, 133(5):3059–3078, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In European Conference on Computer Vision, pages 399–417. Springer, 2025.

Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. arXiv preprint arXiv:2405.18991, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 15703–15712, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In NeurIPS, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In CVPR, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 22963–22974, 2025.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Rui-Jie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. Advances in Neural Information Processing Systems, 37:21236–21270, 2024.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. arXiv preprint arXiv:2505.20292, 2025a.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 12978–12988, 2025b.

Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025c.

Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. Advances in Neural Information Processing Systems, 37:30300–30326, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404, 2024.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852, 2023.