This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Video2LoRA: Unified Semantic-Controlled Video Generation via Per-Reference-Video LoRA

Zexi Wu[1], Baolu Li[1], Jing Dai[1], Yiming Zhang[1],
Yue Ma[2]✉, Qinghe Wang[1]✉, Xu Jia[1], Hongming Xu[1]✉
[1]Dalian University of Technology [2]HKUST

https://github.com/BerserkerVV/Video2LoRA

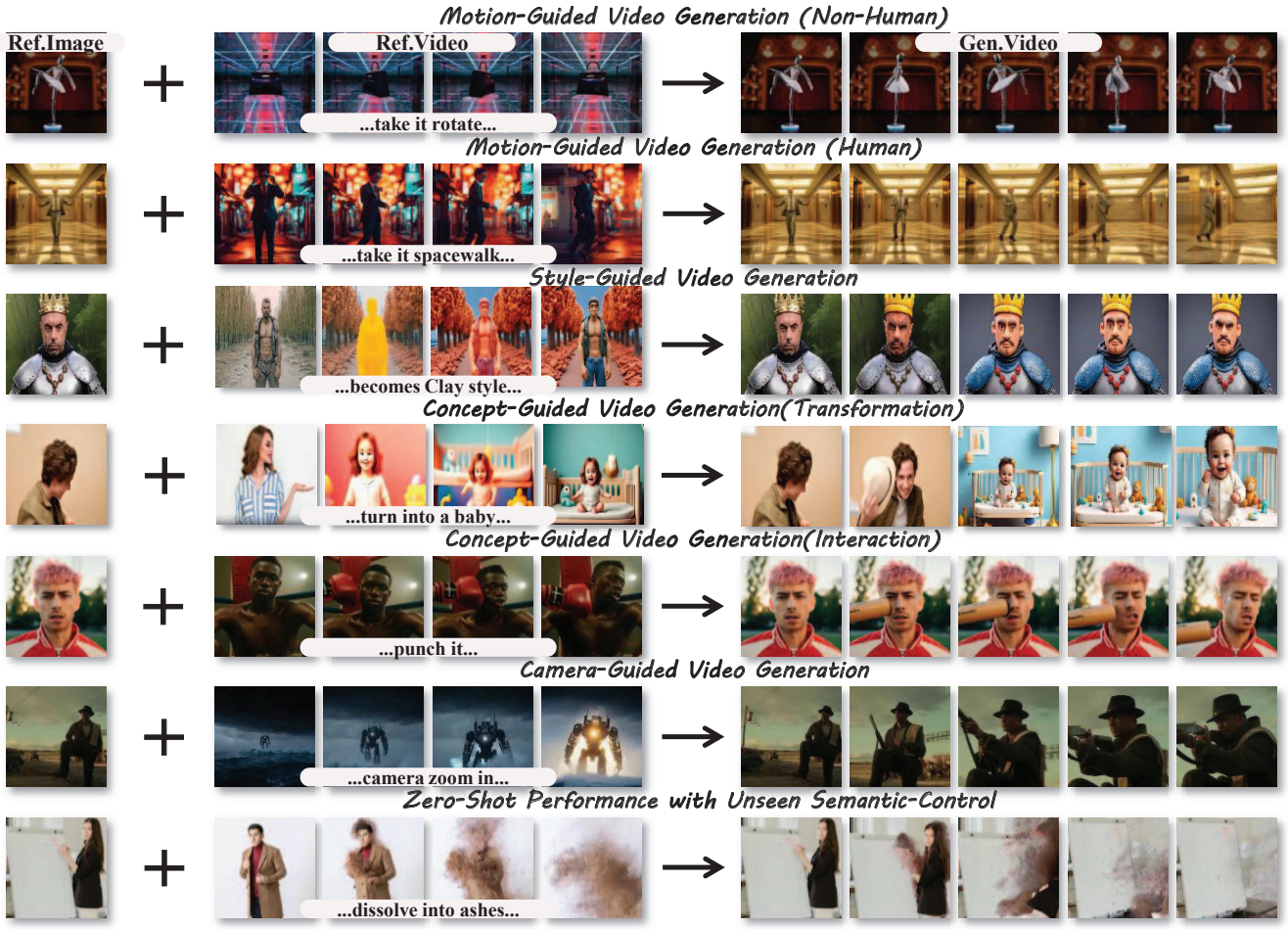


Figure 1. Video2LoRA is a unified framework for semantic-controllable video generation. It takes a reference video containing the desired semantics as input and employs a HyperNetwork to generate lightweight, semantic-specific LoRA modules. By integrating these adaptive components into a frozen video diffusion backbone, Video2LoRA achieves high-quality video generation in both within-domain and out-of-domain scenarios.

Abstract

Achieving semantic alignment across diverse video gener-

✉ Corresponding Author

ation conditions remains a significant challenge. Methods that rely on explicit structural guidance often enforce rigid spatial constraints that limit semantic flexibility, whereas models tailored for individual control types lack interoperability and adaptability. These design bottlenecks hinder progress toward flexible and efficient semantic video generation. To address this, we propose ***Video2LoRA****, a scalable and generalizable framework for semantic-controlled video generation that conditions on a reference video. Video2LoRA employs a lightweight hypernetwork to predict personalized LoRA weights for each semantic input, which are combined with auxiliary matrices to form adaptive LoRA modules integrated into a frozen diffusion backbone. This design enables the model to generate videos consistent with the reference semantics while preserving key style and content variations, eliminating the need for any per-condition training. Notably, the final model weights less than 150MB, making it highly efficient for storage and deployment. Video2LoRA achieves coherent, semantically aligned generation across diverse conditions and exhibits strong zero-shot generalization to unseen semantics.*

1. Introduction

The rapid evolution of generative AI has profoundly transformed visual content creation, enabling unprecedented levels of efficiency, controllability, and expressiveness. In particular, large-scale pretrained video diffusion models [21, 46, 53, 56] have exhibited impressive capabilities in semantic comprehension and temporally coherent synthesis. Recent advances in controllable video generation [27, 32, 51, 52] have predominantly focused on spatially aligned paradigms,leveraging modalities such as depth maps [38, 51], human poses [15], edge sketches [11], keypoints [12, 16], or optical flow [18] to impose spatially consistent guidance during generation. Unified frameworks operating under these pixel-aligned conditions have been investigated [17], demonstrating stable and precise control over structural attributes. In contrast, semantic-controlled video generation that encompasses aspects such as visual effects [4, 22, 23, 35], camera motion [5], and personalized styles remains [54] relatively underexplored despite its close alignment with real-world creative demands. Such high-level controls are inherently more intuitive to human users but challenging to formalize or acquire, as they often lack explicit spatial or parametric representations (e.g., camera trajectories or semantic annotations). Consequently, establishing a unified, generalizable, and user-friendly framework for semantic video control remains an open and pressing challenge.

Existing controllable video generation methods exhibit limited scalability and generalization due to their condition-specific designs. A prevalent line of research fine-tunes either the diffusion backbone or dedicated Low-Rank Adapter (LoRA) [14] for each semantic condition [23, 35], effectively memorizing condition-specific representations. Although such approaches yield satisfactory control within individual domains, they are computationally expensive, storage-inefficient, and fail to generalize across heterogeneous or composite semantics. Another research direction introduces task-specific architectures or inference branches customized for distinct control types, encoding prior knowledge directly into the model design [5, 54, 55]. However, these handcrafted solutions lack interoperability and are inherently constrained to the semantics on which they are trained. As a result, existing paradigms remain fragmented, requiring substantial reconfiguration for new conditions and exhibiting poor zero-shot generalization to unseen semantic domains.

Inspired by the strong cross-domain personalization capability of hypernetworks demonstrated in HyperDreamBooth [41], we hypothesize that similar meta-adaptive mechanisms can endow video generation models with dynamic semantic adaptability. Specifically, hypernetworks [24] can generate semantic-dependent lightweight parameters to modulate a frozen diffusion backbone, enabling flexible semantic conditioning without the need for task-specific retraining. Building upon this insight, we seek to extend the generalization ability of hypernetworks from personalized image synthesis to the broader domain of semantic video generation.

In this work, we introduce Video2LoRA, a unified and generalizable framework for semantic-controlled video generation that conditions on a reference video to synthesize semantically aligned outputs across both in-domain and out-of-domain scenarios. Video2LoRA achieves strong semantic adaptability through a novel hypernetwork-based generation paradigm, in which the hypernetwork predicts a set of lightweight LoRA weights, each less than 50 KB per semantic condition, and merges them with auxiliary matrices to form adaptive LoRA modules injected into a frozen diffusion backbone. This design enables the model to dynamically modulate generation behavior according to diverse semantic cues while requiring no per-condition fine-tuning. Unlike HyperDreamBooth [41], which relies on pre-trained personalized weights as supervision and a three-stage training pipeline with rank relaxation, Video2LoRA is trained end-to-end in a single stage using only the diffusion loss. Our approach eliminates the need for any pre-training or fine-tuning phases, enabling the hypernetwork to directly learn and generalize semantic representations from raw video data without explicit supervision. Extensive experiments on both in-domain and out-of-domain benchmarks demonstrate that Video2LoRA achieves high-fidelity, semantically aligned video generation under diverse control conditions while exhibiting strong generalization to unseen

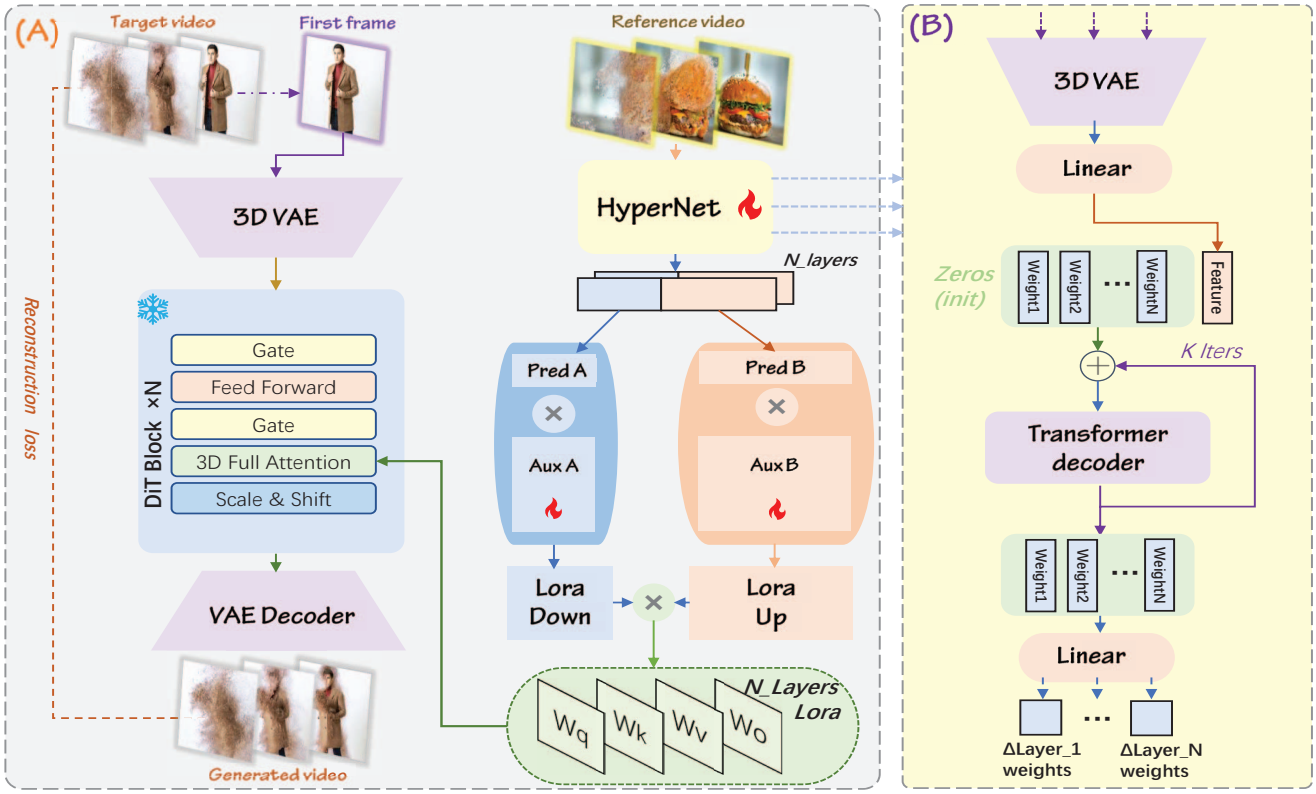


Figure 2. Overview of the proposed **Video2LoRA** framework. A reference semantic video is first fed into the HyperNetwork, where a 3D-VAE encoder extracts spatio-temporal latent features that are linearly projected into the layer-wise LightLoRA subspaces. The projected features are concatenated with zero-initialized weight tokens and processed by a Transformer decoder, which iteratively predicts the LightLoRA components $(A_{\text{pred}}, B_{\text{pred}})$ for each diffusion layer. These predicted components are then fused with the trainable auxiliary matrices $(A_{\text{aux}}, B_{\text{aux}})$ to form the final semantic-specific LoRA weights. The resulting LoRA adapters are injected into the frozen DiT backbone and optimized end-to-end with the vanilla diffusion loss, enabling semantic-controllable video generation from reference videos.

semantics. Our contributions are summarized as follows:

- **Lightweight LoRA representation.** We propose a compact LoRA formulation by training the video generation model within a low-dimensional, trainable weight subspace constructed from a random orthogonal incomplete basis in the low-rank adaptation space. Each semantic condition requires less than 50 KB of parameters.
- **Novel hypernetwork architecture.** We design a novel hypernetwork that leverages the lightweight LoRA configuration to dynamically predict semantic-specific LoRA components for a given video condition, enabling efficient and adaptive control within a unified diffusion backbone.
- **End-to-end semantic generalization.** Unlike prior approaches that rely on pre-trained semantic weights or explicit supervision for each condition, Video2LoRA trains the hypernetwork directly using diffusion objectives, allowing it to implicitly capture semantic relationships and generalize to unseen conditions.

2. Related work

2.1. Video Generation

Recent advances in video generation have been largely driven by diffusion models [8, 9, 13, 25, 28, 30, 31, 33, 34, 36, 47, 48], particularly those adopting Diffusion Transformer (DiT) architectures [37], which integrate the generative strength of diffusion processes with the contextual modeling power of transformers [45]. Such designs greatly enhance temporal coherence and improve motion dynamics. For example, Open-Sora [56] demonstrates efficient long-duration synthesis through scalable transformer blocks and optimized spatiotemporal attention; CogVideoX [53] employs full 3D self-attention to jointly model spatial-temporal dependencies, significantly improving frame-to-frame consistency; and Wan 2.2 [46] incorporates a Mixture-of-Experts (MoE) design to achieve scalable specialization across heterogeneous video content. Despite these advances, most pre-trained DiTs [26, 29, 49, 50] remain limited to text-only or frame-based conditioning, re-

stricting fine-grained semantic control. To address this limitation, recent efforts introduce task-specific modules or customized inference mechanisms to enable more flexible and user-driven video generation.

2.2. Controllable Video Generation

Controllable video generation can be broadly categorized into spatial-alignment and semantic-control paradigms:

The spatial-alignment paradigm leverages explicit structural cues, such as depth [38, 51], pose [15], mask [7], optical flow [18], or motion trajectories [11] to impose pixel-level constraints on synthesis. LongVie [10] integrates multimodal depth and keypoint guidance to achieve temporally coherent ultra-long video generation. Animate Anyone [15] employs pose-based conditioning with spatial attention to achieve appearance-consistent character animation. Motion Prompting [11] introduces trajectory-based motion cues ("motion prompts") to flexibly control object and camera dynamics. While these methods excel at structure-aware control and fine-grained synthesis, they depend on labor-intensive signal extraction or external annotations, making them less suitable for abstract or semantic-level conditioning.

The semantic-control paradigm involves high-level, concept-driven manipulations such as visual effects, camera motion (e.g., trajectories, zooms, or orbits), and personalized stylization (e.g., Ghibli, anime, or Minecraft styles). Existing approaches, such as VFXCreator [23] and GS-DiT [6], achieve control by fine-tuning diffusion backbones or condition-specific Low-Rank Adapters (LoRA) for each semantic condition, including motion type, visual style, or camera behavior, thereby "memorizing" domain-specific representations. Although effective within isolated domains, these methods incur substantial computational cost, suffer from poor parameter efficiency, and fail to generalize across heterogeneous or compositional semantics. Other works, including StyleMaster [54], DiT-Flow [39], and VD3D [3], introduce task-specific architectures for style extraction, motion guidance, or camera-based 3D reasoning, embedding prior knowledge directly into the model structure. However, such specialization inherently limits flexibility and hinders generalization to unseen semantics. OmniEffects [35] attempts to integrate multiple video semantics via a Mixture-of-Experts (MoE) framework but remains confined to in-domain compositions without achieving true cross-domain adaptability. To overcome these limitations, we propose Video2LoRA, a unified semantic video generation framework that enables zero-shot generalization across diverse semantic conditions..

3. Method

We introduce Video2LoRA, a unified and generalizable framework for controllable video generation that learns semantic control end-to-end by dynamically producing lightweight LoRA parameters for any video semantic. Unlike prior approaches [23] that depend on pre-trained semantic experts or condition-specific finetuning pipelines, Video2LoRA directly adapts a diffusion-based video backbone using semantic cues extracted from reference videos, enabling stronger flexibility, scalability, and zero-shot generalization.

Our method is built upon three core components. First, Sec. 3.1 reviews the fundamentals of the CogVideoX diffusion backbone and LoRA-based parameter-efficient adaptation. Then, Sec. 3.2 introduces our *LightLoRA* representation—a compact and trainable low-dimensional parameterization that enables the HyperNetwork to generate semantic-adaptive LoRA weights efficiently. Next, Sec. 3.3 describes the proposed Transformer-based HyperNetwork that predicts these semantic-dependent LoRA components by analyzing spatio-temporal features extracted from a reference video. Finally, Sec. 3.4 presents the full end-to-end training pipeline, where predicted LoRA weights, auxiliary matrices, and the CogVideoX backbone are jointly optimized under the standard image-to-video diffusion objective.

3.1. Preliminaries

Video Generation Backbone. We build our framework upon **CogVideoX-5B-I2V** [53], an image-to-video diffusion model that integrates a **3D Variational Autoencoder (VAE)** [20], a **Diffusion Transformer (DiT)** backbone, and a **T5-based text encoder** [40]. Given an input image $I \in \mathbb{R}^{h\times w\times c}$ and a textual prompt, the model synthesizes a video $V \in \mathbb{R}^{f\times h\times w\times c}$ containing f frames.

During training, the 3D VAE compresses each target video into latent representations that capture spatial-temporal structures, while the first frame is separately encoded for temporal alignment. These latent features are concatenated and passed through the DiT backbone, which iteratively refines the noisy latent sequence through a diffusion-based denoising process guided by text embeddings. This diffusion formulation enables CogVideoX to learn coherent temporal dynamics and high-fidelity visual content, serving as a strong foundation for our controllable video generation framework.

Low-Rank Adaptation (LoRA) [14] provides a parameter-efficient fine-tuning strategy that has been widely adopted in diffusion-based generation frameworks. Instead of updating all network parameters, LoRA optimizes low-rank residual matrices that are added to the frozen model weights. Formally, for a given layer l with a weight matrix $W \in \mathbb{R}^{n\times m}$, LoRA introduces a learnable residual term:

$$\Delta W = AB, \tag{1}$$

where $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times m}$, with rank $r \ll \min(n, m)$. This low-rank decomposition significantly re-

duces the number of trainable parameters while maintaining expressive adaptation capacity.

3.2. Light Weight Lora Representation

To enable direct generation of semantic-specific weight subsets through a HyperNetwork while maintaining semantic fidelity, editability, and generalization, we propose a novel low-dimensional trainable weight space for semantic control in video diffusion models. This compact formulation enables multi-semantic LoRA models that are over **150**× **smaller** than the CogVideoX backbone and more than **20**× **smaller** than single-semantic LoRA-CogVideoX variants.

The core idea of our *LightLoRA* is to further decompose the rank-1 LoRA weight space while maintaining trainability of the decomposed factors. As illustrated in Figure 2(A), this can be understood as decomposing the `Down(A)` and `Up(B)` matrices in Eq. (2) into two components:

$$A = A_{\text{aux}} A_{\text{pred}}, \quad B = B_{\text{pred}} B_{\text{aux}}, \tag{2}$$

where $A_{\text{aux}} \in \mathbb{R}^{n\times a}$ and $B_{\text{aux}} \in \mathbb{R}^{b\times m}$ are auxiliary matrices initialized with row-wise orthogonal vectors of constant magnitude and set to be trainable. The matrices $A_{\text{pred}} \in \mathbb{R}^{a\times r}$ and $B_{\text{pred}} \in \mathbb{R}^{r\times b}$ are dynamically predicted by the HyperNetwork for each semantic condition. Consequently, the residual weight in each linear layer is expressed as:

$$\Delta W x = A_{\text{aux}} A_{\text{pred}} B_{\text{pred}} B_{\text{aux}}, \tag{3}$$

where $r \ll \min(n, m)$, $a < n$, and $b < m$. Two newly-introduced hyperparameters, a and b, control the dimensionality of the auxiliary subspace. In our experiments, setting $a = 100$ and $b = 50$ yields only **23K trainable variables** (approximately **30 KB** in `bf16`), while preserving strong semantic adaptability and zero-shot generalization. Unlike LiDB [41], where the auxiliary matrices are frozen, our trainable A_{aux} and B_{aux} act as *semantic priors* that encode generalizable video semantics. During training, the HyperNetwork learns to combine these priors via the dynamically predicted A_{pred} and B_{pred}, producing condition-specific rank-1 LoRA adapters.

3.3. HyperNetwork Architecture

As illustrated in Figure 2(B), the proposed HyperNetwork $\mathcal{H}_\eta$ is composed of a 3D-VAE encoder, a linear projection layer, and a Transformer-based decoder. The encoder shares the same architecture as the video backbone's 3D-VAE to ensure feature-level alignment between the adaptation module and the generative model. Since the encoder is sequentially dependent on the weights of different layers, effective model personalization requires capturing inter-layer dependencies in the generated LoRA parameters. Previous works [23] overlook this dependency, treating layer weights as conditionally independent. In contrast, our Transformer decoder explicitly models these positional dependencies through learned positional embeddings, analogous to how language models capture contextual relationships among tokens. This design enables the HyperNetwork to reason over structured relationships between layers rather than generating them independently.

The encoder first extracts spatio-temporal latent features f from the input reference video, capturing both motion dynamics and semantic content. These features are projected through a linear layer and passed into the Transformer decoder, which sequentially predicts the semantic-specific LoRA components (A_{pred}, B_{pred}) across layers.

To further enhance inter-layer consistency, we adopt an iterative refinement mechanism similar to recurrent inference. At each iteration k, the decoder refines its prediction based on the previous output:

$$\theta_{pred}^{(k)} = \mathcal{T}(f, \theta_{pred}^{(k-1)}), \tag{4}$$

where $\mathcal{T}$ denotes the Transformer decoder and $\theta_{pred}^{(0)}$ is initialized to zero. The refinement process continues until $k = s$, where s specifies the number of refinement steps. This iterative design effectively enforces semantic stability and temporal coherence while remaining computationally efficient, since the video encoding f is computed only once and reused throughout the refinement process.

3.4. HyperNetwork for Video Semantic Adaptation

To enable fast and unified semantic control in video generation, we adopt a HyperNetwork-based adaptation mechanism for image-to-video diffusion models. A HyperNetwork $\mathcal{H}_\eta$, parameterized by η, takes the 3D-VAE feature representation x_i of a reference semantic video as input and predicts the low-rank *LightLoRA* residuals $\hat{\theta} = \mathcal{H}_\eta(x_i)$. Each predicted $\hat{\theta}$ corresponds to the LoRA component of a specific attention layer and is fused with the auxiliary matrices defined in Eq. (3) to construct the final adaptive LoRA weights. These LoRA adapters are injected into the frozen DiT backbone of CogVideoX-I2V, enabling semantic conditioning within a single diffusion training stage.

In contrast to prior personalization methods [23, 35, 41] that require pre-optimized semantic weights or condition-specific finetuning, our framework jointly trains both the HyperNetwork and the auxiliary matrices solely under the standard I2V diffusion objective. This unified learning scheme allows the HyperNetwork to absorb semantic priors directly from diffusion dynamics, yielding strong generalization across diverse in-domain and out-of-domain semantic conditions.

During training, the 3D-VAE encodes the target semantic video into latent features z, while the first frame is replaced by a placeholder token (-1) for temporal alignment and encoded as z_i. The concatenated latent pair

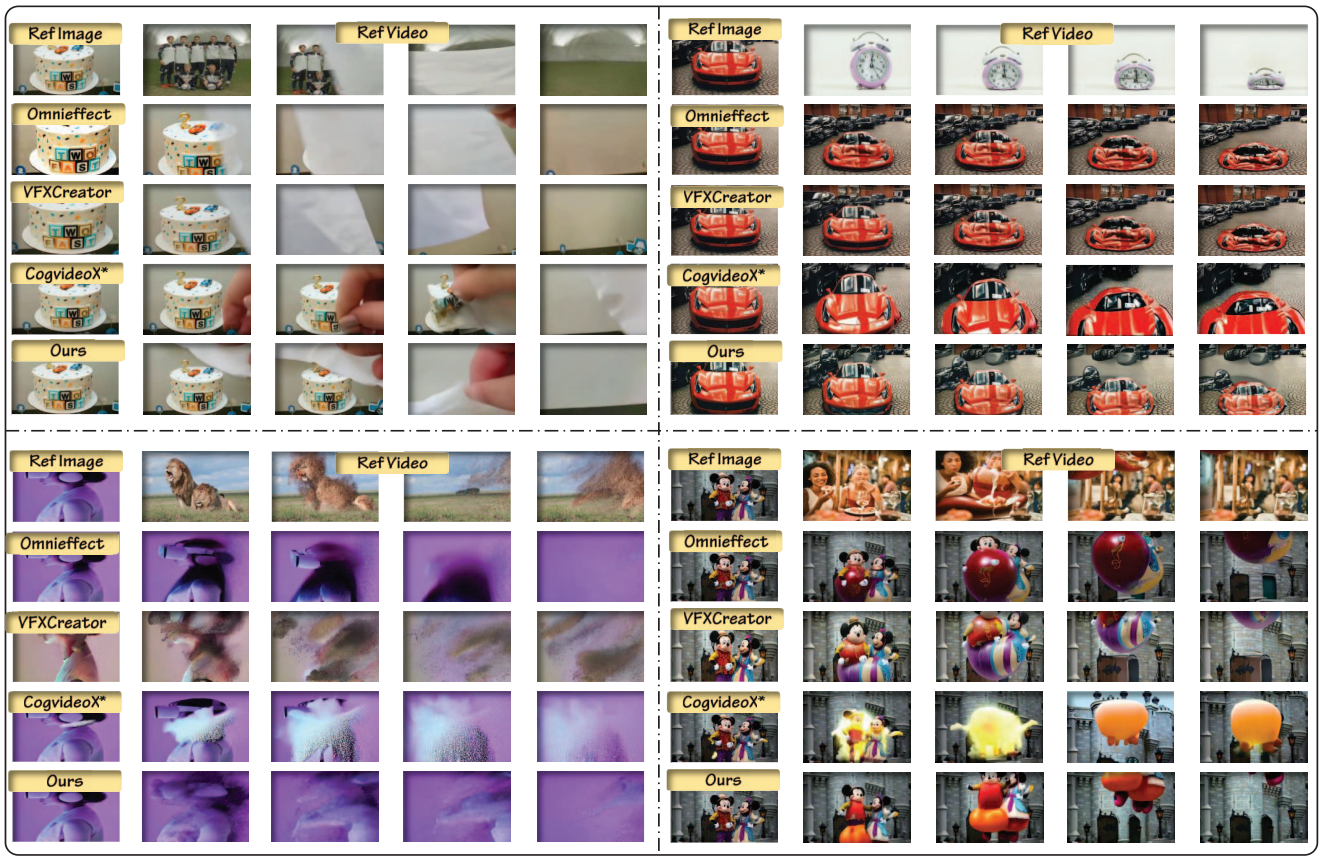


Figure 3. Qualitative comparison with VFXCreator [23] and Ominieffect [35] on the OpenVFX dataset. CogVideoX* refers to the CogVideoX model after supervised fine-tuning on our dataset.

$[z_i, z]$ is then passed into the DiT denoiser, where the HyperNetwork-predicted LoRA weights and auxiliary matrices jointly modulate the denoising process. The denoising network ϵ_Θ is trained using the standard diffusion objective:

$$\mathcal{L}_{\text{diff}}(\Theta) = \mathbb{E}_{t,z,\epsilon}\left[\|\epsilon - \epsilon_\Theta(z_t, t, g)\|_2^2\right], \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, z_t is the noisy latent at timestep t, g is the text embedding, and Θ denotes the parameters of the denoising model. This diffusion-driven supervision propagates gradients through the injected LoRA modules, effectively training the HyperNetwork and auxiliary matrices in an end-to-end manner. Figure 2 provides an overview of the full training pipeline.

4. Experiments

4.1. Implementation Details

We employ **CogVideoX-I2V-5B** [53] as the frozen backbone for all experiments. During training, the *LightLoRA* weights predicted by the HyperNetwork are combined with the auxiliary matrices to form rank-1 low-rank adapters, which are injected into the 3D Transformer blocks of the backbone. The Video2LoRA framework is trained on the **4k** dataset and randomly pairs samples from the same semantic category to construct reference-target video pairs. Each video is uniformly sampled to 49 frames at 8 fps and resized to a resolution of 480×720 pixels. The reference videos are zero-padded to match the spatial and temporal dimensions of the target videos. We use the **AdamW** [19] optimizer with a learning rate of 1×10^{-4} and train only the parameters of the HyperNetwork and auxiliary matrices while keeping the backbone frozen. Training is performed on 8 NVIDIA A800 GPUs for approximately 20K iterations.

4.2. Datasets

Our training dataset is constructed from multiple sources, including the open-source **Open-VFX** [23] dataset, commercial video platforms such as **Higgsfield** [1] and **PixVerse** [2], as well as publicly available online resources. In total, the dataset comprises approximately **4K video samples** spanning over **200 distinct semantic categories**, covering a diverse range of effects, including character trans-

Table 1. **Performance comparison on OpenVFX dataset.** CogvideoX* refers to CogVideoX after supervised fine-tuning on our dataset. Avg. represents the average score over all effects. The highest metric values are highlighted in **bold**.

Metrics	Methods	Cake	Crumble	Crush	Decap	Deflate	Dissolve	Explode	Eye-pop	Harley	Inflate	Levitate	Melt	Squish	Ta-da	Venom	Avg.
FVD↓	CogvideoX*	1732	1849	1195	1937	1664	1916	2427	1649	**2232**	2169	1473	2941	1938	1431	**2792**	1956
	VFX Creator	1776	1580	1156	1754	1997	1607	1886	**1447**	2815	2089	1143	2547	1880	1107	3062	1856
	Omini-Effects	**1548**	1410	1136	**1263**	**1037**	1543	2044	1559	2501	1464	1295	2418	1923	1368	2678	1679
	Ours	1573	**1358**	**1107**	1677	1294	**1412**	**1125**	1528	2466	**1162**	**1005**	**2193**	**1606**	**1027**	2973	**1568**
Dynamic Degree↑	CogvideoX*	1.0	1.0	0.6	0.6	0.4	0.4	1.0	0.0	1.0	0.4	0.0	0.6	1.0	0.8	1.0	0.65
	VFX Creator	1.0	1.0	0.0	0.6	0.0	**0.8**	1.0	0.0	1.0	1.0	0.0	0.6	1.0	1.0	1.0	0.67
	Omini-Effects	1.0	1.0	0.6	0.6	0.2	0.4	1.0	0.2	1.0	1.0	0.0	**0.8**	1.0	0.8	1.0	0.71
	Ours	**1.0**	**1.0**	**0.8**	0.6	**0.6**	**0.8**	**1.0**	**0.2**	**1.0**	**1.0**	**0.2**	0.6	**1.0**	**1.0**	**1.0**	**0.78**
Motion Smoothness↑	CogvideoX*	97.25	96.80	98.10	97.95	98.42	98.15	97.83	98.67	**99.02**	98.45	98.76	98.01	97.56	98.05	97.48	98.17
	VFX Creator	97.84	97.10	98.23	97.68	98.51	**98.60**	97.96	98.72	98.88	98.32	98.69	98.14	97.70	98.26	97.62	98.16
	Omni-Effects	97.66	**97.58**	98.34	97.83	98.47	98.41	**98.22**	98.69	98.95	98.48	98.71	98.25	97.92	98.30	97.73	98.24
	Ours	**98.02**	97.34	**98.56**	**99.24**	**99.28**	98.42	98.06	**99.39**	97.04	**99.01**	**99.55**	**98.46**	**98.30**	**98.48**	**96.51**	**98.50**
Aesthetic Quality↑	CogvideoX*	0.49	0.52	0.50	0.47	0.55	0.52	0.46	0.54	0.50	0.47	0.54	0.51	0.49	0.53	0.48	0.506
	VFX Creator	0.50	0.51	0.53	0.46	0.57	0.54	0.49	**0.57**	0.56	0.48	0.52	0.50	0.47	0.55	0.50	0.519
	Omini-Effects	0.52	0.54	0.55	0.49	0.58	0.56	0.51	0.54	**0.58**	**0.53**	0.55	**0.52**	0.48	0.57	0.52	0.537
	Ours	**0.58**	**0.59**	**0.57**	**0.55**	**0.59**	**0.58**	**0.54**	0.56	0.57	0.52	**0.58**	0.51	**0.53**	**0.59**	**0.55**	**0.565**

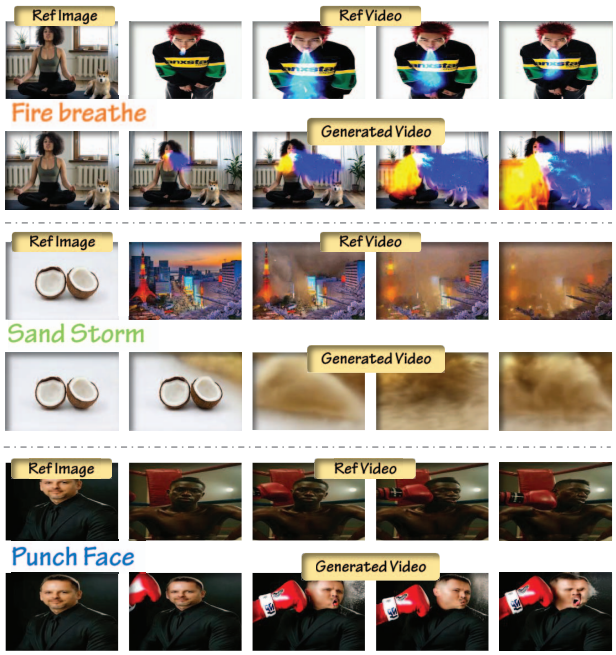


Figure 4. Out-of-Domain Comparison

formations, environmental transitions, camera motion dynamics, object stylization, and artistic style variations.

To further evaluate the robustness and generalization capability of our framework, we curate a dedicated **out-of-domain (OOD)** test set containing unseen semantic conditions. This dataset enables a systematic assessment of the model's ability to adapt to novel visual effects and semantic distributions beyond the training domain.

4.3. Evaluation Metrics

Following previous studies [7, 23, 35], we comprehensively evaluate the proposed method using multiple quantitative metrics, including FVD [44], dynamic degree [43], motion smoothness [40], and aesthetic quality [42]. These metrics collectively reflect different aspects of video generation performance, and detailed descriptions are omitted for brevity.

To quantitatively evaluate the in-domain performance of our approach, we conduct experiments on 15 semantic categories selected from the **Open-VFX** test set. As shown in Table 1, we comprehensively compare **Video2LoRA** against two state-of-the-art VFX generation methods and a baseline model fine-tuned on the same dataset. The results demonstrate that our **Video2LoRA** consistently outperforms all competing approaches in terms of average scores across all evaluation metrics, exhibiting superior visual fidelity, motion coherence, aesthetic appeal, and dynamic range. Notably, for complex effects involving particle dynamics or strong subject interactions, such as *Crumble*, *Crush*, *Decap*, and *Inflate*, our model achieves substantially higher realism and temporal consistency. These findings indicate that **Video2LoRA** not only learns to capture semantic information from reference videos but also reproduces such semantics with higher fidelity and stability over time.

To further evaluate the model's generalization capability beyond the training domain, we conduct a zero-shot out-of-domain (OOD) evaluation, with results reported in Table 2. The evaluation shows that the model's performance on unseen videos is comparable to that observed in the in-domain setting, demonstrating that **Video2LoRA** can generate high-quality, temporally coherent videos even for pre-

Table 2. **Ablation study and zero shot generationon Video2LoRA.** We evaluate the impact of different iterative steps k and auxiliary matrix settings (a, b) on performance. The upper part compares different k values, while the lower part analyzes the influence of (a, b).

Methods	FVD↓	Dynamic Degree↑	Motion Smoothness↑	Aesthetic Quality↑
Ours	**1358**	**0.72**	**98.50**	**0.57**
Ours (Zero-Shot)	1492	0.71	98.37	0.54
Ours (k=1)	1764	0.63	97.45	0.51
Ours (k=2)	1598	0.67	97.92	0.53
Ours (k=8)	1439	0.70	98.23	0.55
Ours (a=60, b=30)	1512	0.69	98.10	0.54
Ours (a=160, b=80)	1384	0.71	98.42	0.56

viously unseen semantic effects. These results further validate the framework's robust zero-shot generalization and semantic adaptation capabilities.

4.4. Qualitative Comparison

We conduct a qualitative comparison between **Video2LoRA** and three representative models across four distinct visual effect categories, as illustrated in Figure 3. Compared with the fine-tuned **CogVideoX-5B** and two state-of-the-art VFX generation frameworks, **Video2LoRA** produces results with notably higher visual fidelity and semantic accuracy. For instance, under the *Dissolve* effect, our model not only captures the gradual disintegration of the subject with fine-grained temporal consistency but also realistically simulates secondary physical behaviors, such as the natural fall of the subject's VR headset after dissolution. Similarly, for the *Levitate* effect, **Video2LoRA** generates smooth and coherent motion trajectories while maintaining semantic alignment with the reference.

Furthermore, Figure 4 presents the results of zero-shot experiments. Even on unseen videos, **Video2LoRA** generates content whose visual style and semantic effects are well-aligned with the reference videos, accurately capturing the intended semantics. For example, in the *Punch Face* effect, the model successfully generates the entire reaction process of a punch to the face, including precise facial deformations and realistic motion of fluids, demonstrating high-fidelity motion and semantic accuracy. These results highlight the model's strong zero-shot generalization capability. Overall, **Video2LoRA** achieves visual quality and semantic coherence that match or surpass current open-source state-of-the-art models, emphasizing its superior ability in precise controllable video effect synthesis.

4.5. Ablation Study

We first analyze the influence of the *LightLoRA* configuration by varying the dimensions of the predicted matrices $(A_{\text{pred}}, B_{\text{pred}})$, which are controlled by hyperparameters a and b. Specifically, we experiment with three settings: $(60, 30)$, $(100, 50)$, and $(160, 80)$. As shown in Table 2, the configuration $(100, 50)$ achieves the best overall video generation quality. The smallest setup $(60, 30)$ fails to capture sufficient semantic diversity due to its limited representational capacity, while increasing the parameter size to $(160, 80)$—a 1.6× increase—does not lead to further improvements and even slightly degrades performance, likely due to overfitting and reduced semantic sparsity. These results indicate that an appropriately compact latent LoRA space is crucial for effective semantic adaptation.

We further investigate the effect of the iterative prediction mechanism introduced in Sec. 3.4. Specifically, we compare our full model (with $k = 4$ refinement steps) against variants with fewer iterations ($k = 1$, $k = 2$) and a deeper refinement setup ($k = 8$). As summarized in Table 2, the performance improves as the number of iterations increases up to $k = 4$, beyond which further refinement yields diminishing returns and slightly higher computational cost. This demonstrates that four refinement rounds provide an optimal balance between semantic consistency, stability, and efficiency.

5. Conclusion

In this work, we introduce **Video2LoRA**, a unified framework for semantic-controlled video generation that leverages a hypernetwork to predict semantic-specific LoRA weights from a reference video. By decoupling semantic adaptation from backbone modification, freezing the diffusion model while only training a compact hypernetwork and auxiliary matrices, our approach eliminates the need for per-condition fine-tuning or pre-trained adapters. This design enables strong zero-shot generalization to unseen semantic domains while maintaining high visual fidelity and temporal coherence. Extensive experiments on the Open-VFX dataset demonstrate that Video2LoRA outperforms existing methods across multiple metrics, including FVD, motion smoothness, and aesthetic quality, despite using significantly fewer parameters. Ablation studies further validate the effectiveness of our lightweight LoRA representation and iterative refinement strategy. We believe this paradigm opens a scalable path toward truly general-purpose semantic control in generative video models.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant No. 82102135, 62472065, U23B2010), the Liaoning Province Science and Technology Joint Program (Grant No. 2024-MSLH-065), the Fundamental Research Funds for Central Universities (Grant No. DUT25Z2514, DUT24YG201).

References

[1] Higgsfield. https://higgsfield.ai/, 2025. Accessed: June 1, 2025. 6

[2] Pixverse: Ai-powered image and video editing platform. https://app.pixverse.ai/, 2025. Accessed: June 1, 2025. 6

[3] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 4

[4] Jianhong Bai, Xiaoshi Wu, Xintao Wang, Xiao Fu, Yuanxing Zhang, Qinghe Wang, Xiaoyu Shi, Menghan Xia, Zuozhu Liu, Haoji Hu, et al. Semanticgen: Video generation in semantic space. *arXiv preprint arXiv:2512.20619*, 2025. 2

[5] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 2

[6] Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *arXiv preprint arXiv:2501.02690*, 2025. 4

[7] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 4, 7

[8] Yiyang Chen, Xuanhua He, Xiujun Ma, and Yue Ma. Contextflow: Training-free video object editing via adaptive context enrichment. *arXiv preprint arXiv:2509.17818*, 2025. 3

[9] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 3

[10] Jianxiong Gao, Zhaoxi Chen, Xian Liu, Jianfeng Feng, Chenyang Si, Yanwei Fu, Yu Qiao, and Ziwei Liu. Longvie: Multimodal-guided controllable ultra-long video generation. *arXiv preprint arXiv:2508.03694*, 2025. 4

[11] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025. 2, 4

[12] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 4

[15] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2, 4

[16] Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7276–7287, 2025. 2

[17] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 2

[18] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2040–2049, 2025. 2, 4

[19] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[21] 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. 2

[22] Baolu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, et al. Vfxmaster: Unlocking dynamic visual effect generation via in-context learning. *arXiv preprint arXiv:2510.25772*, 2025. 2

[23] Xinyu Liu, Ailing Zeng, Wei Xue, Harry Yang, Wenhan Luo, Qifeng Liu, and Yike Guo. Vfx creator: Animated visual effect generation with controllable diffusion transformer. *arXiv preprint arXiv:2502.05979*, 2025. 2, 4, 5, 6, 7

[24] Yewei Liu, Xiyuan Wang, Yansheng Mao, Yoav Gelbery, Haggai Maron, and Muhan Zhang. Shine: A scalable in-context hypernetwork for mapping context to lora in a single pass. *arXiv preprint arXiv:2602.06358*, 2026. 2

[25] Zeqian Long, Mingzhe Zheng, Kunyu Feng, Xinhua Zhang, Hongyu Liu, Harry Yang, Linfeng Zhang, Qifeng Chen,

and Yue Ma. Follow-your-shape: Shape-aware image editing via trajectory-guided region control. *arXiv preprint arXiv:2508.08134*, 2025. 3

[26] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4132–4141, 2022. 3

[27] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2

[28] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[29] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025. 3

[30] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025. 3

[31] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025. 3

[32] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 2

[33] Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, et al. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*, 2025. 3

[34] Yue Ma, Zhikai Wang, Tianhao Ren, Mingzhe Zheng, Hongyu Liu, Jiayi Guo, Mark Fong, Yuxuan Xue, Zixiang Zhao, Konrad Schindler, et al. Fastvmt: Eliminating redundancy in video motion transfer. *arXiv preprint arXiv:2602.05551*, 2026. 3

[35] Fangyuan Mao, Aiming Hao, Jintao Chen, Dongxia Liu, Xiaokun Feng, Jiashu Zhu, Meiqi Wu, Chubin Chen, Jiahong Wu, and Xiangxiang Chu. Omni-effects: Unified and spatially-controllable visual effects generation. *arXiv preprint arXiv:2508.07981*, 2025. 2, 4, 5, 6, 7

[36] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 3

[37] 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. 3

[38] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 2, 4

[39] Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22911–22921, 2025. 4

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 4, 7

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6527–6536, 2024. 2, 5

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 7

[43] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 7

[44] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[46] 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. 2, 3

[47] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *Advances in Neural Information Processing Systems*, 37:96541–96565, 2024. 3

[48] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024. 3

[49] Qinghe Wang, Xu Jia, Xiaomin Li, Taiqing Li, Liqian Ma, Yunzhi Zhuge, and Huchuan Lu. Stableidentity: Inserting anybody into anywhere at first sight. *IEEE Transactions on Multimedia*, 2025. 3

[50] Qinghe Wang, Baolu Li, Xiaomin Li, Bing Cao, Liqian Ma, Huchuan Lu, and Xu Jia. Characterfactory: Sampling consistent characters with gans for diffusion models. *IEEE Transactions on Image Processing*, 2025. 3

[51] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 2, 4

[52] Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025. 2

[53] 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. 2, 3, 4, 6

[54] Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2630–2640, 2025. 2, 4

[55] Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. Flexiact: Towards flexible action control in heterogeneous scenarios. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 2

[56] 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. 2, 3