LiON-LoRA: Rethinking LoRA Fusion to Unify Controllable Spatial and Temporal Generation for Video Diffusion

Yisu Zhang[1,2*‡], Chenjie Cao[2,3*], Chaohui Yu[2,3], Jianke Zhu[1†]
[1]Zhejiang University [2]DAMO Academy, Alibaba Group [3]Hupan Lab
{zhyisu,jkzhu}@zju.edu.cn, {caochenjie.ccj,huakun.ych}@alibaba-inc.com,

Abstract

*Video Diffusion Models (VDMs) have demonstrated remarkable capabilities in synthesizing realistic videos by learning from large-scale data. Although vanilla Low-Rank Adaptation (LoRA) can learn either specific spatial or temporal movement to driven VDMs with constrained data, it is still challenging to achieve precise control over both camera trajectories and object motion due to the unstable fusion and non-linear scalability. To address these issues, we propose **LiON-LoRA**, a novel framework that rethinks LoRA fusion through three core principles: **Linear scalability**, **Orthogonality**, and **Norm consistency**. Firstly, we analyze the orthogonality of LoRA features in shallow VDM layers, enabling decoupled low-level controllability. Secondly, norm consistency is enforced across layers to stabilize fusion during complex camera motion combinations. Thirdly, a controllable token is integrated into the diffusion transformer (DiT) to linearly adjust motion amplitudes for both cameras and objects with a modified self-attention mechanism to ensure decoupled control. Finally, we extend LiON-LoRA to temporal generation by leveraging static-camera videos, unifying spatial and temporal controllability. Experiments demonstrate that LiON-LoRA outperforms state-of-the-art methods in trajectory control accuracy and motion strength adjustment, achieving superior generalization with minimal training data. Project Page: https://fuchengsu.github.io/lionlora.github.io/*

1. Introduction

Rapid advancement of foundational video diffusion models (VDMs) [3, 4, 7, 9, 18, 49, 51, 59, 65, 72, 73] has enabled them to learn realistic interactions from the physical world, resulting in impressive video outputs. Using vast amounts of video data available on the Internet, VDMs have significant potential to understand 3D spatial and even 4D temporal information in the real world [1, 20, 24, 25, 45, 68, 70]. Consequently, reactivating the potentials of VDMs with good controllability would largely promote many downstream applications, such as augmented reality experiences, interactive video editing, and simulation of virtual environments. Despite the abundance of in-the-wild video data, the lack of proper annotations presents a challenge in taming VDMs for 3D and 4D generation, particularly in terms of unified control of both camera and motion strength, as illustrated in Figure 1.

Recent studies made use of either explicit camera presentations [16, 53, 60, 63, 74] or motion guidance [23, 50, 53, 67] to control VDM. However, these manners usually undermine the generalization of VDMs, leading to inferior synthesis quality as a compromise. Thus, it is still challenging to equip VDMs with effective controllability using limited labeled video data. Fortunately, Parameter-Efficient Fine-Tuning (PEFT) [10] alleviates this challenge by reducing the dependency on extensive annotated training data while having the merit of lightweight model weights, thus mitigating the degradation of generalization in VDMs. Low-Rank Adaptation (LoRA) [19] has garnered considerable attention among these PEFT approaches due to its flexibility and versatility. LoRA utilizes low-rank factorized matrices with a minimal number of trainable parameters, making it widely applicable in text-to-image and image-to-image personalization [37, 43, 76]. Additionally, AnimateDiff [20] employs 8 distinct LoRAs to drive the respective camera controls within the VDM. DimensionX [45] further decouples the temporal and spatial factors through unified LoRAs for advanced 4D generation.

Despite the pioneering advances mentioned above, implicitly driving video generation via LoRA fails to achieve as precise control as explicit injection, which could be attributed to two factors: 1) *Unstable LoRA fusion.* Unlike high-level LoRA fusion techniques used for tasks such as style and subject editing [22, 37, 43, 64], we find that combining different camera control LoRAs suffers from obvious instability, such as abrupt changes in direction, as illustrated in Figure 2. 2) *Non-linear LoRA scalability.* The conven-

*Equal Contribution. †Corresponding author. ‡Work done during internship at DAMO Academy, Alibaba Group

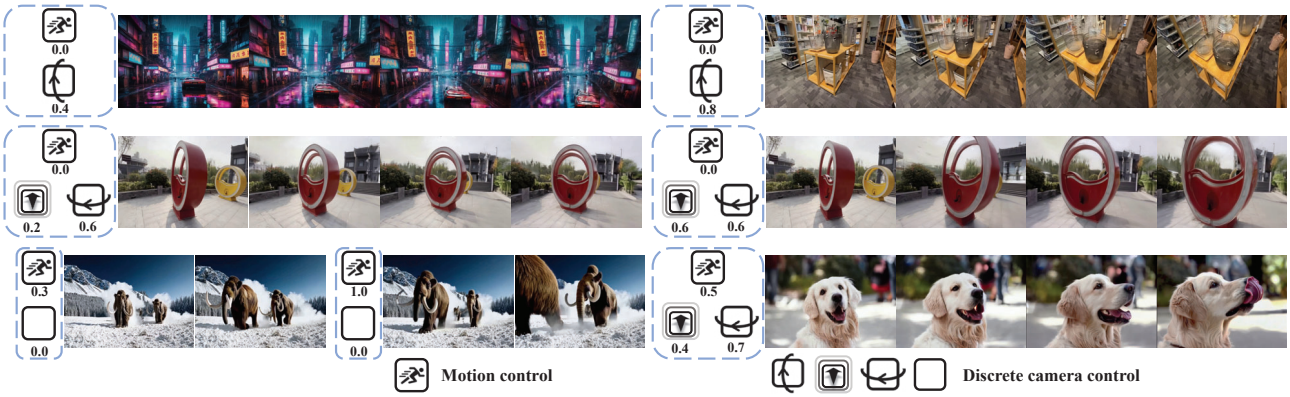

Figure 1. **Camera and motion control results of our LiON-LoRA.** LiON-LoRA can linearly control both camera trajectory and object motion in videos generated by the video diffusion model. Furthermore, based on LoRA fine-tuning, LiON-LoRA achieves excellent generalization with minimal training data.

tional approach of adjusting the multiplying coefficients of LoRA features is adequate for modifying high-level properties, such as style transfer. However, it fails to precisely control camera movements, which demands a deeper low-level spatial understanding.

The underlying issue of these challenges lies in the tendency of most previous works to focus on high-level LoRA fusion, neglecting an in-depth discussion on inherent LoRA mechanisms to quantify low-level controllability. In this paper, we rethink LoRA mechanisms from three aspects, including **Linear scalability**, **Orthogonality**, and **Norm consistency**, and further present **LiON-LoRA**, a framework with unified spatial and temporal controls for VDM, fully compatible with all aforementioned features.

Formally, we first analyze the orthogonality of distinct camera LoRAs and find that LoRA features from shallow layers of VDM exhibit low feature correlation, which is theoretically sufficient to decouple low-level controllability. Moreover, we recognized that distinct camera LoRAs possess disparate magnitudes (norms), leading to unstable LoRA fusions. Therefore, we normalize the output norms from each layer of the LiON-LoRA, ensuring norm consistency and prominently enhancing the stability of LoRA fusion for complex camera motion combinations. Subsequently, we achieve the linear scalability of LiON-LoRA by inserting a dedicated *scaling token* within the LoRA training. Specifically, we concatenated an additional global token to the diffusion transformer (DiT) feature sequence, linearly encoding the motion amplitudes of both cameras and objects. When fusing multiple LoRAs, we adjust the receptive field of self-attention to attend to separate scaling tokens from different LoRAs, enjoying decoupled linear control. Additionally, we collect extensive video clips with static camera positions, extending the linear scalability of LiON-LoRA into temporal control. This unifies both 3D and 4D generation within VDM. Experiments of camera control with both simple and complicated trajectories demonstrate that LiON-LoRA outperforms other methods, offering superior controllability and generalization with minimal training samples. We also evaluate the Pearson correlation, denoting that LiON-LoRA can effectively control motion strength for 4D generation.

We summarize the key contributions of LiON-LoRA as

- Unlike previous LoRA works that focus primarily on high-level fusion, we revisit the LoRA fusion in low-level camera control and present LiON-LoRA, acquiring linear scalability, orthogonality, and norm consistency.
- We propose a simple yet effective way to normalize the LiON-LoRA feature's norm to enable stable camera movement fusion.
- We introduce a scaling token for LiON-LoRA, which enables linear scalability to adjust the influence of camera and object movements.
- Although LiON-LoRA is highly training-efficient, it successfully unifies both spatial and temporal controllability, facilitating impressive 4D generation.

2. Related Work

Image-to-Video Generation. Recent years have witnessed remarkable advances in video generation technology, significantly propelling the development of content synthesis fields. Early research efforts [3, 7, 8, 13, 18, 20, 49, 49, 51] Pioneered text-to-video (T2V) generation by adapting pre-trained text-to-image (T2I) models through architectural adjustments or fine-tuning strategies. Subsequently, commercial video generation engines *e.g.*, Sora [4], Gen-3 [42], and Kling [29] have demonstrated expanded capabilities in T2V synthesis, along with extended functionalities that include image-to-video (I2V) conversion and specialized video effect generation. These closed-source systems typically employ comprehensive video generation pipelines incorporating extensive pre-processing and post-processing modules. Recent open source initiatives have improved methodological transparency and accessibility for research

communities [28, 32, 59, 65, 75], providing modular frameworks for video synthesis. While both CogVideoX [65] and HunyuanVideo [28] adopt the MMDiT structure, a modified Transformer variant employing full attention mechanisms [11]. In contrast, Open-Sora leverages the DiT backbone [34, 39, 69], demonstrating competitive performance through spatial-temporal modeling in latent space.

Camera Controllable Video Generation. Controlled generation has gained increasing attention in both image [35, 40, 44, 55, 58, 66, 71] and video generation [6, 14, 20, 40], with control signals evolving from simple text descriptions to include audio [17, 46, 47], identity [5, 52, 56], user interaction patterns [67], and camera trajectory specifications [6, 31, 36, 54, 63]. Among these, text-based camera control is one of the simplest yet most effective approaches, leveraging natural language descriptions to guide camera motion. Early research [3, 15] used fine-tuning techniques, such as LoRA, to parse control signals from text inputs. Furthermore, training-free methods [21, 23] have been proposed to generate coarse-grained camera movements, though with limited precision. Although these text-based approaches enable camera trajectory control, they often struggle with precise control and exhibit limited controllability. To address the challenge of camera control, some methods [16, 30, 53, 63] explicitly inject camera pose as an additional input into the model, while others [60, 74] leverage epipolar attention to improve the multiview consistency. Furthermore, some works [12, 70] take advantage of recent advances in monocular depth estimation models [61, 62] to reconstruct 3D point clouds from images as explicit control, allowing accurate camera movement for free-view rendering. However, these methods require intensive training preparation that involves the extraction of motion trajectories from large-scale video data. Additionally, such explicit controls struggle to handle dynamic scenes due to their overly rigid conditions. Despite these explorations, the challenges of achieving precise and controllable camera and object motion using easily constructed, small-scale datasets persist. Our method aims to address these limitations.

Motion Control in Video Generation Object motion control has emerged as a critical interaction channel for video manipulation, with recent methods predominantly employing bounding-box annotations or drag operations as control signals. Bounding box-based approaches [23, 50, 74] enable object trajectory specification through sequential bounding boxes that outline target positions across frames. Alternative drag-based methods [53, 57, 67] leverage sparse point trajectories to manipulate object movements. These methods suffer from limited trajectory capacity and complex user interactions, while still generating implausible motions from inadequate real-world dynamics modeling. In contrast to these trajectory-based control mechanisms, our method introduces a parameterized motion scale that jointly encodes motion properties in 3D space. Besides, this control paradigm avoids complex interactions while preserving the plausibility of natural motions.

3. Method

Overview. We first briefly review LoRA and its usage for camera control in Section 3.1. Based on this formulation, we then analyze the challenges of vanilla LoRA fusion for camera control (Section 3.2). Subsequently, we propose LiON-LoRA with orthogonality and norm consistency (Section 3.3), as well as linear scalability (Section 3.4) respectively. Finally, we introduce the training-free fusion of LiON-LoRA in Section 3.5.

3.1. Preliminary: Discrete LoRA Camera Control

Complex camera movements in 3D space can be decomposed into combinations of fundamental motion primitives. Through LoRA [19] fine-tuning, we demonstrate that VDMs can effectively adapt to basic motion patterns using only a minimal set of reference videos with limited training iterations, as evidenced in Table 3. Therefore, we modify open-source CogVideoX [65] by injecting LoRA modules into the linear layers of the transformer blocks. This lightweight adaptation enables precise control over camera motion patterns while preserving the model's original generative capabilities. The parameter update process for each motion primitive is formalized as:

$$\Delta W_i = A_i B_i, \quad A_i \in \mathbb{R}^{d\times r}, B_i \in \mathbb{R}^{r\times k}, \tag{1}$$

$$W_i = W_i^{\text{base}} + \lambda \Delta W_i, \tag{2}$$

where i indexes distinct motion primitives, r, d, and k represent the LoRA rank, input, and output dimensions, respectively. In addition, W_i^{base} and ΔW_i represent the original CogVideoX weights and low-rank updates introduced by LoRA, while λ controls the intensity of LoRA injection to achieve the final model weights W_i.

In this paper, three types of fundamental camera primitives are included: (1) *horizontal/vertical offset*, (2) *forward/backward movement*, (3) *orbital rotation around the scene center*. Each primitive requires dedicated LoRA weight training, which requires motion-specific video data preparation. To achieve videos to learn distinct camera control LoRAs, we render fixed and stable trajectories through 3D Gaussian Splatting (3DGS) [27] trained on the multiview dataset, DL3DV [33]. We adaptively decide the moving distance according to the medium depth of each scene. More details are discussed in the supplementary.

3.2. Rethinking LoRA Fusion for Camera Control

Pre-trained foundational generative models typically enable linear combinations of multiple LoRA weights to achieve

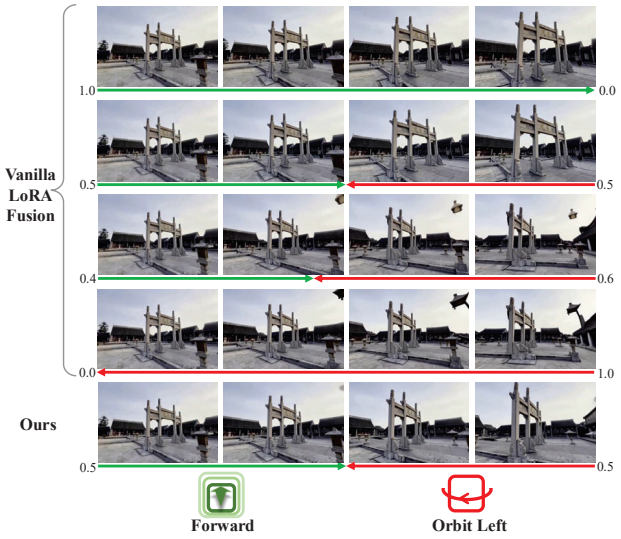

Figure 2. **Comparison between the vanilla LoRA fusion and ours.** The vanilla fusion of "forward" and "orbit left" LoRA controls with adjusted adapter scales suffers from abrupt changes. While our LiON-LoRA enhanced with norm consistency enjoys a much smoother combination without parameter adjusting.

multitask capabilities, where the fusion strategy used between different LoRA modules is particularly significant. The most straightforward method for LoRA fusion involves linearly combining multiple LoRAs ΔW_i through coefficients $\lambda_i, i \in \{1, 2, ..., k\}$ as:

$$\Delta W = \sum_{i=1}^{k} \lambda_i \Delta W_i, \tag{3}$$

where ΔW denotes the weight of the fused LoRA module. Although directly linearly combining LoRA weights performs well for high-level controllable tasks, such as style and subject fusion [37, 43, 76], this vanilla strategy does not manage satisfactory low-level control, such as camera movements fusion, as illustrated in Figure 2. Specifically, the LoRA fusion of the vanilla camera suffers from abrupt changes (*e.g.*, suddenly switching from forward movement to orbit left) rather than a smooth progression in blending these two camera motions, even when the combining coefficients are carefully adjusted.

To address these issues, we rethink the LoRA fusion from the perspective of vector space principles, including orthogonality, norm consistency, and precise linear scalability. **Linear scalability**: Although learning LoRA primitives is an implicit way to control camera trajectories, explicit linear scalability is essential to achieve accurate camera control in linear space, which cannot be achieved simply by adjusting the LoRA coefficient λ. This property not only facilitates flexible camera control but also enhances the LoRA fusion process by enabling smoother movements. **Orthogonality**: When multiple tasks share a base model, high-level

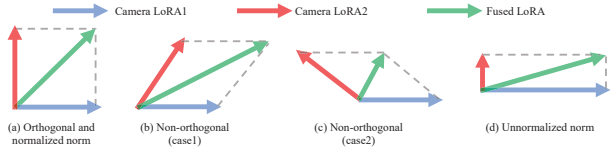

Figure 3. **Illustration of the dilemma for LoRA fusion in camera control.** We show the 2D toy case to demonstrate the importance of orthogonality and norm consistency. Non-orthogonality results in undesired enhancements or conflicts of fusions (b, c), while significantly different norms lead to unstable fusion (d).

correlations among LoRA features may lead to undesired enhancements or conflicts, as shown in Figure 3(b) and Figure 3(c), compromising the characteristics of each LoRA during fusion. Thus, ensuring the orthogonality of the features generated by each camera LoRA is a fundamental requirement for effective fusion, allowing the model to preserve the distinct characteristics of each motion. **Norm consistency**: Since different camera control LoRAs are trained independently, their low-rank matrices exhibit divergent Frobenius norms. This disparity results in disproportionate amplification or suppression of specific LoRA weights during the combination, allowing one module to dominate over others, as shown in Figure 3(d). To this end, we introduce **LiON-LoRA** method to encompass all three properties, which effectively decouples camera control capabilities within the parameter space while maintaining their combined expressive power.

3.3. Orthogonality and Norm Consistency

Orthogonality. As mentioned above, the orthogonality should be the base to get the complicated trajectories combined from discrete ones. As shown in Figure 4, our layer-wise similarity analysis reveals some factors of camera LoRAs' orthogonality. LoRA outputs in earlier transformer blocks exhibit strong orthogonality, with an average cosine similarity of 0.06±0.06, while deeper blocks show a notable increase in correlation. This phenomenon suggests that camera LoRAs have achieved almost orthogonal characteristics in the shallower layers. As demonstrated in [1, 45], the low-frequency camera controls are primarily encoded within the initial transformer blocks of VDMs, with subsequent layers emphasizing high-frequency generation. This observation aligns with our theoretical analysis, *i.e.*, the LoRA features corresponding to different camera trajectories exhibit good orthogonality in early blocks. It suggests that the orthogonality should be sufficient to handle the low-frequency camera control.

Norm Consistency. As illustrated in Figure 5, we observe significant norm disparities among different LoRAs related to various camera trajectories and transformer blocks across different samples. These disparities stem from variations in the underlying distribution of the training data. Notably, the

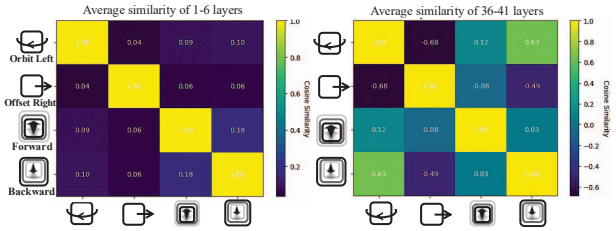

Figure 4. **Cosine similarity of different camera LoRAs.** In the shallow layers, LoRAs maintain orthogonality, while in the later layers, coupling effects begin to emerge.

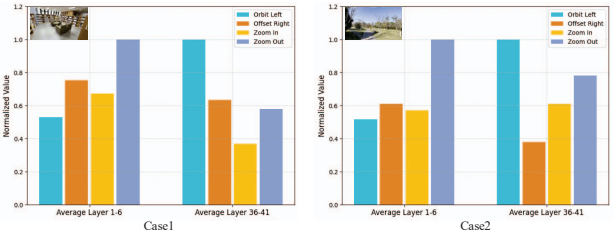

Figure 5. **Norm Consistency of different LoRAs.** The norms of outputs from different camera control LoRAs vary significantly for different samples, which occur in both early and later layers, reflecting that LoRA weights fail to ensure norm consistency.

norm discrepancies are not uniform across the transformer blocks, making it challenging to establish a pre-set adapter scale that effectively balances the LoRA outputs. To address this issue, we propose a post-normalization approach tailored to individual LoRA outputs to ensure a balanced contribution across different LoRAs as follows:

$$\Delta\hat{W}_i = \frac{\alpha}{\|\Delta W_i\|}\Delta W_i, \tag{4}$$

α serves as a unified scaling factor for k different LoRAs, which is typically set to $\alpha = \sum(\|\Delta W_i\|)/k$ for balanced norm preservation. This simple yet effective layer-wise norm normalization scheme significantly improves the smoothness of LoRA fusion, as shown in Figure 2.

3.4. Scaling Controllable Video Diffusion

Linear Scalability. Our empirical pilot studies reveal that adjusting the adapter value λ of each camera LoRA during training fails to enable explicit linear control. We summarize the limitations of conventional LoRA scaling as follows. (1) λ is not integrated as the inputs of VDMs, without explicitly affecting other video and textual features to control generations. (2) Increasing the value of λ undermines the overall generalization of the base VDM, contradicting the primary purpose of LoRA. (3) λ determines the intensity of LoRA injection that facilitates high-level feature fusion. However, this adjustment struggles to achieve fine-grained precision in controlling camera trajectories.

To address these limitations, we adopt a flexible control mechanism by incorporating an additional *scaling token* into the visual and textual token sequence to explicitly represent the scale of camera motion. Specifically, we define a scaling value $\mathcal{S} \in [s, 1]$ to present the amplitude of each camera movement, where s denotes the minimal scaling value. Then we render corresponding videos from DL3DV which contains 600 frames for each video clip. Consequently, we could prepare training pairs of $\mathcal{S}$ and $V = 49$ video frames uniformly sampled from the first $600{\cdot}\mathcal{S}$ frames of each clip, while the minimal $s = 49/600$. To improve the fine-grained representation of $\mathcal{S}$, we employ the Fourier-based positional embedding $\gamma(\cdot)$ before sending it to the linear projection as follows:

$$\gamma(S) = \sum_{j=0}^{J-1}[\sin(2^j\pi S), \cos(2^j\pi S)], \tag{5}$$

$$\mathcal{E} = \text{linear}(\gamma(S)), \tag{6}$$

where $j \in \{0, ..., J-1\}$ denote the frequency indice, and $\mathcal{E}$ is the scaling token. Subsequently, $\mathcal{E}$ is further concatenated with the visual and textual token sequence H as below:

$$H' = [H; \mathcal{E}] \in \mathbb{R}^{(n+1)\times d}, \tag{7}$$

where n, d denote the original sequential length and channels for VDM's transformer blocks, respectively. Note that we use specific project linear to encode $\mathcal{S}$ for different camera LoRAs. Thereby, it avoids coupling that could hinder orthogonality. LiON-LoRA remains parameter-efficient since the additional trainable parameters associated with the linear projection are negligible. As shown in Figure 10, the scaling token scheme outperforms the adapter value. Our systematic comparison demonstrates that the scaling token achieves superior and more stable linear scalability.

Motion Strength Scaling. The proposed LoRA scaling paradigm can seamlessly be extended to control the scale of motion strength. Instead, the scaling value $\mathcal{S}$ is defined to present the speed of object motion. Similar to the linear scalability in controllable camera motion, scaling value $S \in [s, 1]$ represents the magnitude of object motion. Different from the setting in camera motion, the maximum frame of the object motion video is 240, resulting in the minimal $s = 49/240$. To ensure that the scaling token focuses exclusively on motion strength unimpeded by undesirable camera movements, we train the motion strength scaling only on videos with static camera positions. To validate the effectiveness of our proposed motion scaling, we compute the Pearson correlation [2] between $\mathcal{S}$ and the optical flow magnitude over consecutive frames, as illustrated in Figure 6. The results demonstrate that our method achieves rapid convergence during the early stages of fine-tuning LiON-LoRA. In contrast, the vanilla adapter scaling fails to establish a meaningful correlation for effective control. Moreover, we show the qualitative results in Figure 7, where the scaling token successfully controls the motion strength with linear scalability.

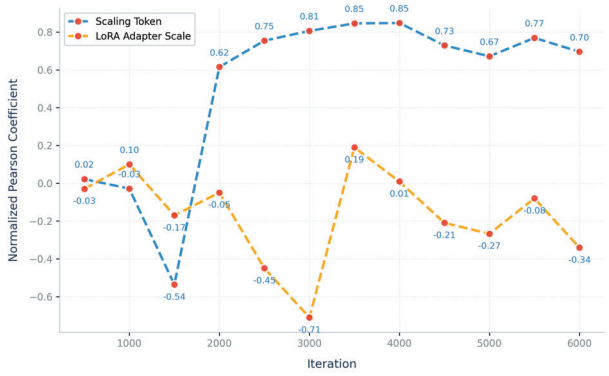

Figure 6. **Pearson correlation of motion strength controllability.** We quantify motion controllability through Pearson correlation analysis between optical flow magnitudes and two control types (scaling scale vs LoRA adapter scale). While the adapter scale exhibits poor motion linearity, our scaling token scheme enables precise control over object motion strength.

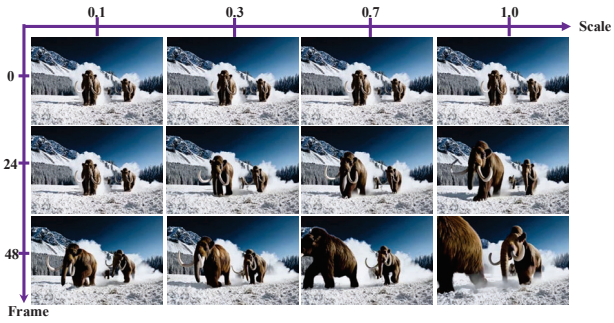

Figure 7. **Illustration of motion strength scaling.** Our LiON-LoRA method enables the control of object motion amplitude through the scaling token without complex bounding box annotations or drag-based interactions. Furthermore, this parameter-driven approach ensures that the generated videos exhibit both temporal continuity and visual consistency.

3.5. Training Free Fusion

The proposed LiON-LoRA achieves both orthogonality and norm consistency without altering the existing LoRA fusion process. However, incorporating new scaling tokens during the LoRA fusion could potentially affect the video sequence and compromise the orthogonality of LoRA outputs, thereby inhibiting independent control. Therefore, we propose a novel fusion framework that enables independent scaling control for multiple LoRAs as illustrated in Figure 8. Given k scaling values $\{S_1, ...S_k\}$, each value is processed through dedicated encoding layers to corresponding scaling tokens $\{\mathcal{E}_1, ..., \mathcal{E}_k\}$. These scaling tokens are then concatenated with original hidden states of video sequence H to form H':

$$H' = [H; \mathcal{E}_1; \ldots; \mathcal{E}_k] \in \mathbb{R}^{(n+k)\times d}. \tag{8}$$

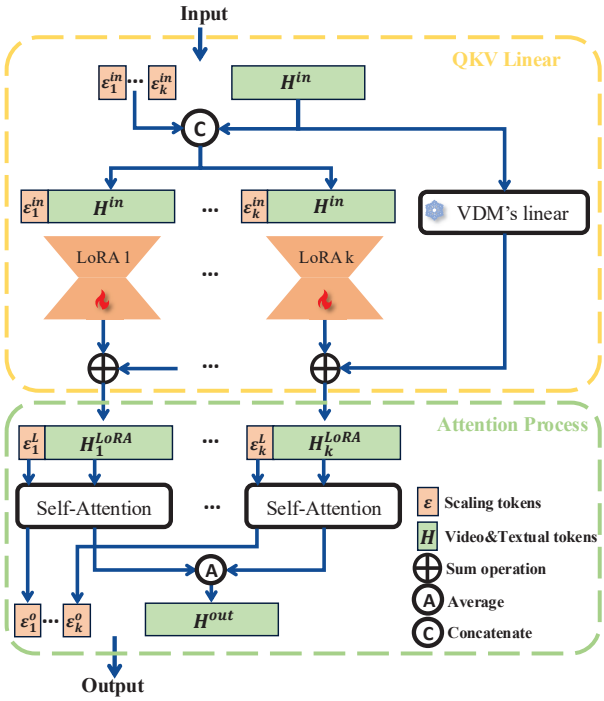

Figure 8. **Pipeline of multiple LiON-LoRA fusion.** LiON-LoRA first splits scaling tokens and then assigns them to the individual LoRA encoding and attention process. Finally, we average k groups of output visual and textual tokens as H^{out}, while k output scaling tokens are spatially concatenated.

To maintain individual control without feature coupling, each LoRA output is designed to attend exclusively to a specific scaling token $\mathcal{E}_i$ within its own independent attention subspace, denoted as $H_i{}' = [H; \mathcal{E}_i]$. After the self-attention, different scaling tokens are spatially concatenated while the hidden states are averaged. The resulting output tokens $H^{out} \in \mathbb{R}^{(n+k)\times d}$ then propagate through the subsequent blocks, ensuring that control is properly disentangled. Notably, LiON-LoRA fusion is training-free and highly versatile, allowing for multi-trajectory camera fusions and hybrid camera-motion controls without any joint training.

4. Experiments

4.1. Experiment Setup

Dataset. We select the large-scale, multi-category scene dataset DL3DV [33] as the training dataset of camera control. For each basic camera trajectory, we reconstruct the scene using 3DGS [27] and render videos by moving the virtual camera along pre-defined trajectory patterns, including horizontal/vertical offset, forward/backward movement, and orbital rotation around the scene center. Ultimately, we reconstruct 100 scenes and render the corresponding videos for each fundamental camera primitive. To train controllable object motion, we collected 172 videos from the Inter-

Table 1. **Quantitative results of basic camera poses.** The DimensionX-S* indicates the re-implemented version of S-Director of DimensionX [45].

Method	RotErr ↓	TransErr ↓	ATE ↓	FVD ↓
CogVideoX [65]	4.974	0.765	0.980	387.6
MotionCtrl [53]	2.254	0.269	0.408	290.9
CameraCtrl [16]	1.737	0.192	0.458	218.9
CamI2V [74]	1.033	0.215	0.370	294.6
DimensionX-S* [45]	1.223	0.201	0.359	193.3
Ours	**0.776**	**0.167**	**0.295**	**136.0**

Table 2. **Quantitative results of complex fused camera poses.**

Method	RotErr ↓	TransErr ↓	ATE ↓	FVD ↓
CogVideoX [65]	4.104	0.717	0.842	383.2
MotionCtrl [53]	2.063	0.317	0.426	357.2
CameraCtrl [16]	1.947	0.201	0.440	261.8
CamI2V [74]	**0.924**	0.217	0.398	285.6
DimensionX-S* [45]	1.510	0.264	0.405	240.7
Ours	1.044	**0.197**	**0.345**	**172.8**

net, which contain videos with flexible object movements while the camera remains static. These videos with static cameras are filtered by heuristic processing within boundary tracking points detected from Cotracker [26]. We further employed the optical flow to filter videos with too small motions. This design facilitates the decoupling of camera and object motion, enabling precise control over both modalities. For quantitative performance evaluation, we selected 100 samples from DL3DV for each motion primitive and conducted evaluations on both basic and complex camera poses. For qualitative results, we conduct comprehensive experiments across diverse data sources including DL3DV, web images, and generated images.

Implementation Details. Our LiON-LoRA is based on CogVideoX [65], a VDM trained on 49-frame sequences. This LoRA-based training approach can be seamlessly adapted to other video diffusion models, such as HunYuan [28] and DynamicCrafter [59] with its flexibility. During training, we supervise the diffusion model using videos at a resolution of 480×720, following the CogVideoX pretraining scheme. We set the LoRA rank at 256 and fine-tuned the LoRA layers for 4,000 steps with a learning rate of 5×10^{-4}. We train LiON-LoRA and variants on 4 NVIDIA H20 GPUs with a batch size of 16. For inference, we employ DDIM sampling with 50 steps, and the classifier-free guidance scale is set to 6.

Metrics and Baselines. Following CameraCtrl [16], we evaluate the accuracy of camera pose predictions using three metrics. RotErr (rotation error), TransErr (translation error), and ATE (absolute trajectory error). To compute these metrics with the structure-from-motion method, we use GLOMAP [38] to convert multiview images into camera poses for the predicted videos as in [74] . Additionally, we assess video generation quality using FVD [48]. We compare our method with several state-of-the-art baselines, including DimensionX [45], MotionCtrl [53], CameraCtrl [16], and Cami2V [74].

4.2. Qualitative Results

The qualitative results are presented in Figure 9. It can be seen that our model outperforms MotionCtrl, CameraCtrl, and CamI2V in terms of higher resolution, better consistency, less hallucinated content, and longer sequence generation. Furthermore, our approach requires only $4,000$ iterations of fine-tuning, significantly fewer than the at least $50,000$ steps typically needed by other competitors. This efficient training strategy helps preserve the generalization of the VDM. A particularly notable comparison is with the DimensionX baseline, where our model achieves linear controllability and multi-motion fusion under the same foundation model, CogvideoX. Furthermore, CameraCtrl, MotionCtrl, and CamI2V may suffer from hallucinated content and temporal flickering due to limitations in their base models and potential overfitting to specific training datasets.

4.3. Quantitative Results

The evaluation encompasses both fundamental camera poses and their complex combinations to assess motion control fidelity. As demonstrated in Table 1 and Table 2, our method achieves significant improvements in both camera controllability and visual quality. For fundamental camera poses, quantitative results reveal superior performance across all metrics: Rotation Error (RotErr) $\downarrow 24.8\%$, Translation Error (TransErr) $\downarrow 13.0\%$, and Average Trajectory Error (ATE) $\downarrow 21.5\%$. The enhanced control capability stems from introducing the scaling token that enables precise control of camera trajectories. Currently, due to the efficient training architecture, the reduced number of fine-tuning iterations enables our framework to maintain enhanced temporal consistency during long-form video generation (49 frames vs. 16-frame), demonstrating a 37.8% reduction in FVD compared to CamI2V. As for complex camera motion combinations, our method also achieves superior performance except RotErr metric, as demonstrated in Table 2.

4.4. Ablation Study

Dataset Size and Training Iterations. Compared to existing Plücker [41] embeddings as global positional embedding to inject camera parameters, which typically require large-scale datasets and extensive fine-tuning (10k+ iterations), our method achieves precise linear controllability with significantly fewer training samples. We conducted comparative experiments on the "orbit left" trajectory using 7k versus 100 training instances, as illustrated in Table 3. The results demonstrate that our model attains satisfactory controllability after only 4k iterations with minimal data.

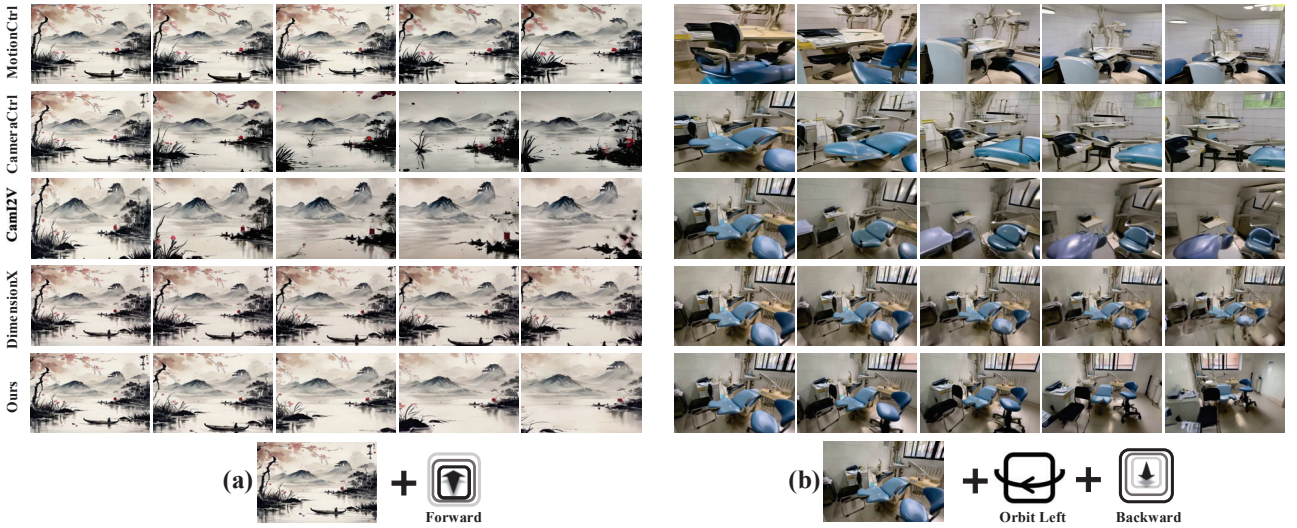

Figure 9. **Qualitative comparisons on camera pose control.** (a) Our model achieves diverse styles with strong generalization and adaptability. Existing methods (MotionCtrl [53], CameraCtrl [16], CamI2V [74]) struggle with inconsistent scene geometry and hallucinations across styles. The re-implemented DimensionX [45] has limited camera movement amplitude and lacks linear control. (b) Under complex trajectories, our method effectively integrates diverse camera movements while maintaining high quality. Others face limitations in large-scale motion and generalization, while DimensionX suffers from LoRA conflicts, losing some camera movements (orbit left).

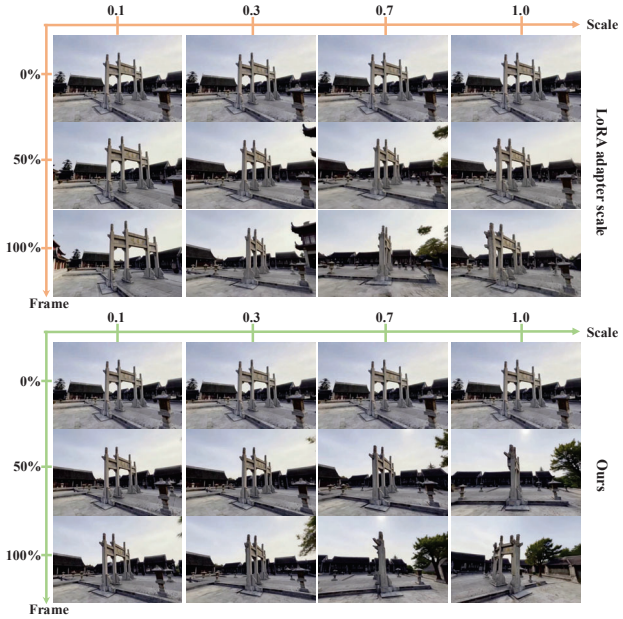

Figure 10. **Linear scalability of LoRA.** Vanilla LoRA adjusts injection strength via the global "adapter scale" λ, which lacks precision in camera motion control. Specifically, an insufficient adapter scale fails to control the camera of VDM, while a large adapter scale may lead to poor quality. In contrast, our LiON-LoRA enables linear controllability with better VDM generalization.

Each Components. Our method achieves precise linear controllability by injecting control signals through additional scaling tokens in the transformer architecture. As compared to the vanilla adapter scaling approach of LoRA in Figure 10, our framework not only enables accurate control but also effectively decouples the intensity of LoRA injection from control signals. To further quantify the influence of individual component on overall model performance, we provide a detailed ablation analysis in Table 4.

Table 3. **Ablation study on dataset size and training iterations.**

Samples	Iteration	RotErr ↓	TransErr ↓	ATE ↓	FVD ↓
100	4k	0.802	0.143	0.331	**175.3**
100	10k	**0.710**	0.146	**0.327**	184.6
7k	4k	1.227	0.252	0.390	232.5
7k	10k	0.896	**0.140**	0.341	202.8

Table 4. **Ablation study on each components.**

Method	RotErr ↓	TransErr↓	ATE ↓	FVD ↓
W/O scale token	1.317	0.230	0.377	179.3
W/O LiON-fusion	1.863	0.269	0.427	254.5
W/O LoRA norm	1.436	0.265	0.392	227.8
Full model	**1.044**	**0.197**	**0.345**	**172.8**

5. Conclusion

We present LiON-LoRA, a parameter-efficient framework for unified spatial-temporal control in VDMs, based on linear scalability, orthogonality, and norm consistency. We find camera LoRAs exhibit strong orthogonality in shallow VDM layers, while we further normalize LoRA outputs for norm consistency, facilitating smooth camera LoRA fusion. Moreover, we propose a scaling token to control camera movements and motion strength with linear scalability. A training-free LoRA fusion framework is proposed for independent control across these modules. Extensive experiments validate the efficacy of our proposed LiON-LoRA approach with minimal training data.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2023YFF0905104) and also by National Natural Science Foundation of China under Grants (62376244).

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 1, 4

[2] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009. 5

[3] 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. 1, 2, 3

[4] 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. 1, 2

[5] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *arXiv preprint arXiv:2407.08674*, 2024. 3

[6] Changgu Chen, Junwei Shu, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. Motion-zero: Zero-shot moving object control framework for diffusion-based video generation. *arXiv preprint arXiv:2401.10150*, 2024. 3

[7] 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*, 2023. 1, 2

[8] 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, 2024. 2

[9] 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*, 2023. 1

[10] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023. 1

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[12] Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol-camera: Precise video camera control with adjustable motion strength. *arXiv preprint arXiv:2411.06525*, 2024. 3

[13] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023. 3

[15] 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. 3

[16] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 1, 3, 7, 8

[17] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2273, 2024. 3

[18] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 1, 2

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 3

[20] 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. 1, 2, 3

[21] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024. 3

[22] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023. 1

[23] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 8079–8088, 2024. 1, 3

[24] Yanqin Jiang, Li Zhang, Jin Gao, Weiming Hu, and Yao Yao. Consistent4d: Consistent 360° dynamic object generation from monocular video. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[25] Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. Animate3d: Animating any 3d model with multi-view video diffusion. *Advances in Neural Information Processing Systems*, 37:125879–125906, 2025. 1

[26] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024. 7

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3, 6

[28] 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. 3, 7

[29] Kuaishou. Kling. *https://klingai.kuaishou.com*, 2024. 2

[30] Teng Li, Guangcong Zheng, Rui Jiang, Tao Wu, Yehao Lu, Yining Lin, Xi Li, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025. 3

[31] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24153, 2024. 3

[32] 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*, 2024. 3

[33] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 3, 6

[34] 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. 3

[35] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3

[36] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 3

[37] Ziheng Ouyang, Zhen Li, and Qibin Hou. K-lora: Unlocking training-free fusion of any subject and style loras. *arXiv preprint arXiv:2502.18461*, 2025. 1, 4

[38] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 7

[39] 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

[40] 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. 3

[41] Julius Plücker. *Analytisch-Geometrische Entwicklungen*. GD Baedeker, 1828. 7

[42] RunwayML. Gen-3 alpha. *https://runwayml.com/research/introducing-gen-3-alpha*, 2024. 2

[43] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. 2023. 1, 4

[44] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. *arXiv preprint arXiv:2404.05674*, 2024. 3

[45] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 1, 4, 7, 8

[46] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[47] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 3

[48] 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

[49] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2

[50] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[51] 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. *arXiv preprint arXiv:2309.15103*, 2023. 1, 2

[52] Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024. 3

[53] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 3, 7, 8

[54] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*, 2024. 3

[55] Tao Wu, Xuewei Li, Zhongang Qi, Di Hu, Xintao Wang, Ying Shan, and Xi Li. Spherediffusion: Spherical geometry-aware distortion resilient diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6126–6134, 2024. 3

[56] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*, 2024. 3

[57] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 3

[58] Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*, 2024. 3

[59] 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, 2024. 1, 3, 7

[60] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 1, 3

[61] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 3

[62] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 3

[63] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 1, 3

[64] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, and Wei Liu. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv: 2403.11627*, 2024. 1

[65] 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. 1, 3, 7

[66] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[67] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 1, 3

[68] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 1

[69] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. *arXiv preprint arXiv:2403.14148*, 2024. 3

[70] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1, 3

[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[72] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 1

[73] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7747–7756, 2024. 1

[74] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 1, 3, 7, 8

[75] 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. 3

[76] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*, 2024. 1, 4