DreamRelation: Relation-Centric Video Customization

Yujie Wei[1], Shiwei Zhang[2,*], Hangjie Yuan[2], Biao Gong[3], Longxiang Tang[2], Xiang Wang[2], Haonan Qiu[4], Hengjia Li[5], Shuai Tan[3], Yingya Zhang[2], Hongming Shan[1,†]

[1]Fudan University [2]Alibaba Group
[3]Ant Group [4]Nanyang Technological University [5]Zhejiang University

yjwei22@m.fudan.edu.cn, zhangjin.zsw@alibaba-inc.com, hmshan@fudan.edu.cn

Project page: https://dreamrelation.github.io

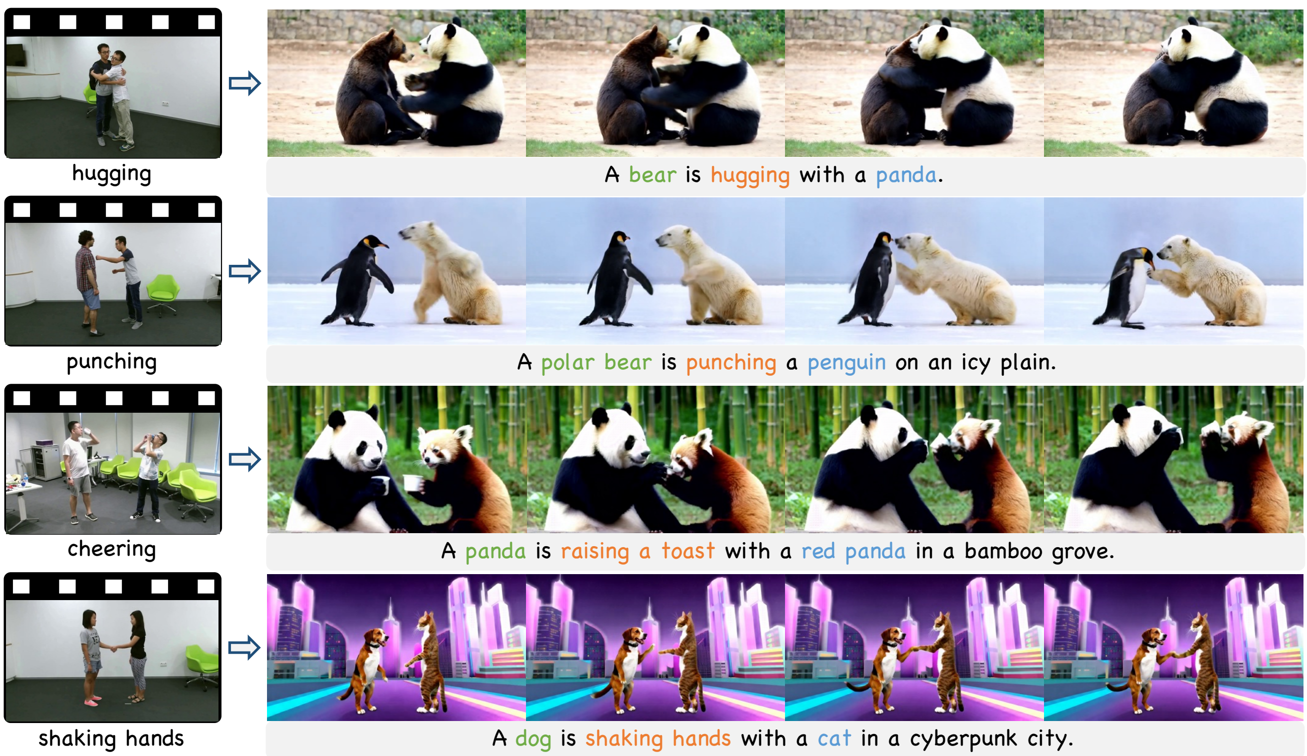


Figure 1. **Relational video customization results of DreamRelation.** Given a few exemplar videos, our method can customize specific relations and generalize them to novel domains, where animals mimic human interactions.

Abstract

Relational video customization refers to the creation of personalized videos that depict user-specified relations between two subjects, a crucial task for comprehending real-world visual content. While existing methods can personalize subject appearances and motions, they still struggle with complex relational video customization, where precise relational modeling and high generalization across subject categories are essential. The primary challenge arises from the intricate spatial arrangements, layout variations, and nuanced temporal dynamics inherent in relations; consequently, current models tend to overemphasize irrelevant visual details rather than capturing meaningful interactions. To address these challenges, we propose ***DreamRelation****, a novel approach that personalizes relations through a small*

* Project Leader
† Corresponding Author: Hongming Shan is with Institute of Science and Technology for Brain-inspired Intelligence, MOE Frontiers Center for Brain Science, Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, and State Key Laboratory of Brain Function and Disorders, Fudan University, Shanghai, China

set of exemplar videos, leveraging two key components: Relational Decoupling Learning and Relational Dynamics Enhancement. First, in Relational Decoupling Learning, we disentangle relations from subject appearances using relation LoRA triplet and hybrid mask training strategy, ensuring better generalization across diverse relationships. Furthermore, we determine the optimal design of relation LoRA triplet by analyzing the distinct roles of the query, key, and value features within MM-DiT's attention mechanism, making DreamRelation the first relational video generation framework with explainable components. Second, in Relational Dynamics Enhancement, we introduce space-time relational contrastive loss, which prioritizes relational dynamics while minimizing the reliance on detailed subject appearances. Extensive experiments demonstrate that DreamRelation outperforms state-of-the-art methods in relational video customization.

1. Introduction

Recent advancements in text-to-video (T2V) generation, particularly through powerful video diffusion transformers (DiT) [5, 65, 104], have significantly propelled customized video generation [38, 93, 108]. While existing methods succeed in customizing subject appearances and single-object motions [85, 97, 112], the challenging task of customizing higher-order interactions between subjects (*i.e.*, Relational Video Customization) remains under-explored due to its intrinsic complexity. Enhancing video generation through customized relations is crucial for real-world applications such as filmmaking, enabling a more profound comprehension and production of complex relational visual content.

We formulate the task of Relational Video Customization as follows: given exemplar videos representing a relational pattern <subject, relation, subject>, the model aims to generate videos that exhibit the specified relation within the pattern, as shown in Fig. 1. While general text-to-video DiTs like Mochi [79] can generate videos depicting certain relational concepts, they often fail to: (1) produce unconventional or counter-intuitive interactions, such as animals engaging in human-like relationships as illustrated in Figs. 2, even when provided with detailed prompts; (2) generate videos that adhere to precise relational dynamics, such as *"two people approaching each other from predefined positions."* These issues highlight the need for a novel video generation method to precisely customize desired relations.

A straightforward approach involves adapting existing video subject or motion customization methods to customize relations between subjects. However, while subject customization techniques like Dreamix [62] capture detailed appearances using low-level reconstruction loss, they may hinder high-level relation learning due to severe appearance leakage. Similarly, motion customization methods

(a) Results of the base model Mochi.

(b) Results of our DreamRelation.

Figure 2. (a) General Video DiT models like Mochi [79] often struggle to generate unconventional or counter-intuitive interactions, even with detailed descriptions. (b) Our method can customize a specific relation to generate videos on new subjects.

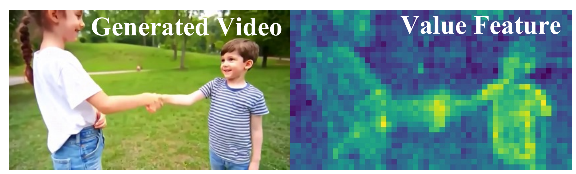


Figure 3. Averaged value feature across all layers and frames in Mochi. We identify that the relations encompass intricate spatial arrangements, layout variations, and nuanced temporal dynamics, presenting challenges in relational video customization.

such as MotionInversion [85] excel in transferring single-object motions but struggle to precisely capture relational dynamics between two subjects. We identify that the key challenge stems from the complexity inherent in the relations, which involve intricate *spatial arrangements*, *layout variations*, and *nuanced temporal dynamics*. To illustrate this, we visualize the Value features in Fig. 3 and provide detailed analysis in Sec. 3.3. This tangled nature may prevent accurate modeling of relations and cause models to focus on irrelevant subject appearances. This raises a critical research question: *How can we decouple relations and subject appearances while accurately modeling relational dynamics to enhance generalizability?*

To that end, we propose ***DreamRelation***, a relational video customization method that personalizes user-specified relations from exemplar videos through two concurrent processes: relational decoupling learning and relational dynamics enhancement. In relational decoupling learning, we decompose the relational pattern from input videos into relational and appearance information using devised relation LoRA triplet, a composite LoRA [32] set comprising relation LoRA sets and subject LoRA sets. To facilitate this decoupling, we introduce hybrid mask training strategy that guides the two types of LoRAs to focus on designated regions with corresponding masks, achieved by a LoRA selection strategy and an enhanced diffusion loss based on masks to amplify the learning in target areas.

Furthermore, building on the MM-DiT [18] architecture, we analyze the query, key, and value features within the full attention, and empirically identify that the query, key, and value matrices serve distinct roles in the relation customization task. This insight motivates our design of relation LoRA triplet, particularly in determining the optimal placement of LoRA components within the model architecture to maximize relational customization effectiveness.

To explicitly enhance relational dynamics learning, we propose a novel space-time relational contrastive loss, which emphasizes relational dynamics while reducing the focus on detailed appearances during training. Concretely, we pull relational dynamics representations closer through frame differences in model outputs of videos depicting the same relation, while distancing them from appearance representations derived from single-frame outputs.

We curate a dataset comprising 26 human interactions from publicly available action recognition datasets [49, 71] to comprehensively evaluate relational video customization. Each video is annotated with a textual prompt, and approximately 20 videos per relation type are randomly selected for training. The evaluation is conducted on diverse subjects using 40 designed textual prompts. Extensive experimental results demonstrate that our DreamRelation outperforms state-of-the-art methods in this task.

Our contributions are summarized as follows:

- We make the first attempt at the Relational Video Customization task by presenting **DreamRelation**, a method that generates videos depicting customized relations based on the MM-DiT architecture.
- We devise relation LoRA triplet with hybrid mask training strategy to explicitly decouple relation and subject appearances. To determine the optimal model design of our method, we further analyze the roles of query, key, and value features in MM-DiT full attention.
- We propose a novel space-time relational contrastive loss to enhance relation learning by emphasizing relational dynamics while reducing focus on appearances.
- Extensive experimental results demonstrate that DreamRelation achieves state-of-the-art performance on relational video customization.

2. Related Work

Text-to-video diffusion models. Text-to-video generative models have achieved breakthroughs in generating high-quality and controllable videos using textual prompts and diverse conditions [1, 3, 4, 7, 17, 24, 29, 41, 48, 56–59, 67, 74–77, 86–90, 98, 103, 107, 109]. VDM [30] introduces diffusion models into video generation by modeling video distribution in pixel space. ModelScopeT2V [84] and VideoCrafter [11, 13] integrate spatiotemporal blocks for text-to-video generation. With the success of DiT [65] that introduces Transformers [82] as the backbone of diffusion models, the generated video quality has improved with increased parameters [5, 19, 43, 55, 113]. CogVideoX [104] incorporates 3D VAE and expert transformers, enhancing video coherence. Mochi [79] proposes an Asymmetric Diffusion Transformer architecture to scale parameters. HunyuanVideo [42] enhances architecture design and model training, achieving leading performance. These advancements pave the way for relational video customization.

Customized video generation. Building upon achievements in image generation and personalization [8, 15, 16, 21, 28, 35, 44, 66, 69, 70, 92, 99, 110, 115], customized video generation has garnered growing attention [9, 26, 46, 60, 62]. Many studies focus on generating personalized videos using a few subject or facial images [45, 72, 93–97, 108, 111, 114], while others tackle the challenging multi-subject video customization [10, 12, 14, 34, 91]. Besides subject customization, motion customization or motion transfer have also gained significant interest [37, 38, 68, 80, 81, 100, 105, 112]. For example, MotionInversion [85] integrates motion embeddings into the temporal attention of video diffusion models to learn motion dynamics. While these methods effectively capture the subject appearances or single-object motions, the challenging task of customizing interactions between two subjects remains underexplored due to its inherent complexity. In this work, we pioneer this relational video customization task by presenting DreamRelation, which can personalize specific relations and generate diverse videos aligned with text prompts.

Relation generation. Early works on relational image generation focus on human-object interactions using additional conditions like bounding boxes [22, 31, 33]. Recently, inspired by image customization methods, several works have explored relational image customization to personalize user-specific interactions from a few relational images [23, 36, 73]. For instance, ReVersion [36] utilizes inversion techniques to capture relational information in the text embedding space. Despite these advancements, existing methods are confined to the relatively simple relations depicted in images. Direct adaptation of these image-based methods for relational video customization often leads to inaccurate relation modeling since dynamic and sequential interactions cannot be fully represented in a single image. In contrast, we design our method based on Video DiT architecture and precisely model relations through relational decoupling learning and relational dynamics enhancement.

3. DreamRelation

Our DreamRelation aims to generate videos depicting a specified relation expressed in a few exemplar videos while

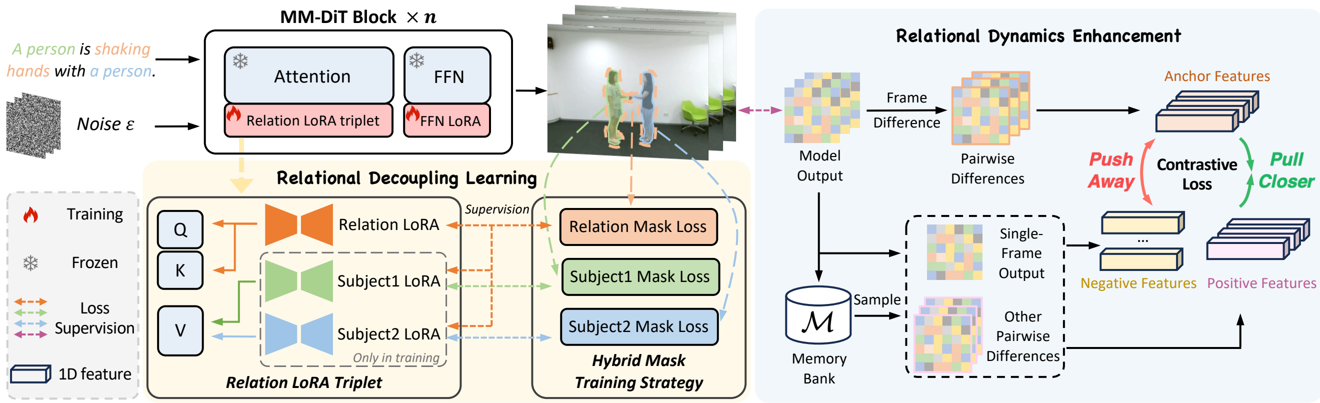


Figure 4. **Overall framework of DreamRelation**. Our method decomposes relational video customization into two concurrent processes. (1) In Relational Decoupling Learning, Relation LoRAs in relation LoRA triplet capture relational information, while Subject LoRAs focus on subject appearances. This decoupling process is guided by hybrid mask training strategy based on their corresponding masks. (2) In Relational Dynamics Enhancement, the proposed space-time relational contrastive loss pulls relational dynamics features (anchor and positive features) from pairwise differences closer, while pushing them away from appearance features (negative features) of single-frame outputs. During inference, subject LoRAs are excluded to prevent introducing undesired appearances and enhance generalization.

aligning with textual prompts, as illustrated in Fig. 4. We begin by introducing preliminaries in Sec. 3.1. We then detail relational decoupling learning and relational dynamics enhancement in Secs. 3.2 and 3.4, respectively, along with an analysis of the query, key, and value features in Sec. 3.3.

3.1. Preliminaries of Video DiT

Text-to-video diffusion transformers (DiTs) show growing attention due to their capacity to generate high-fidelity, diverse, and long-duration video. Current Video DiTs [79, 104] predominantly adopt MM-DiT [18] architecture with full attention and employ diffusion processes [28] in latent space with a 3D VAE [39]. Given latent code $\boldsymbol{z}_0 \in \mathbb{R}^{f\times h\times w\times c}$ from video data $\boldsymbol{x}_0 \in \mathbb{R}^{F\times H\times W\times 3}$ with its textual prompt $\boldsymbol{c}$, the optimization process is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{z},\epsilon,\boldsymbol{c},t}\left[\, \|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t)\|_2^2 \,\right], \tag{1}$$

where $\epsilon \in \mathcal{N}(0,1)$ is random noise from a Gaussian distribution, and $\boldsymbol{z}_t$ is a noisy latent code at timestep t based on $\boldsymbol{z}_0$ with the predefined noise schedule. In this work, we choose Mochi [79] as our base Video DiT model.

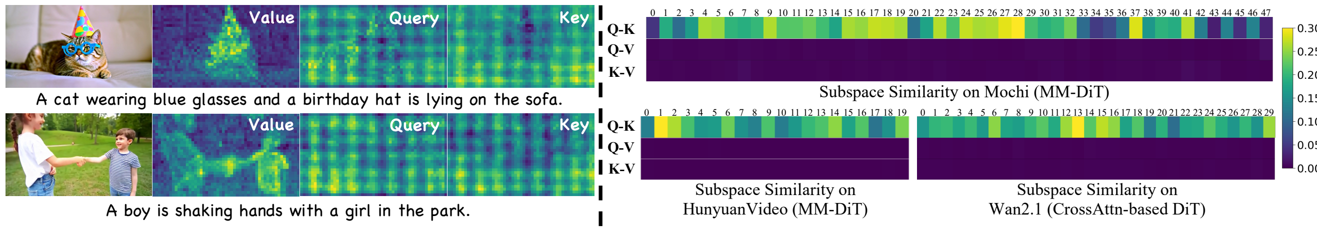
3.2. Relational Decoupling Learning

Relation LoRA triplet. To customize complex relations between subjects, we decompose the relational pattern from exemplar videos into distinct components emphasizing subject appearances and relations. Formally, given a few videos depicting interactions between two subjects, we represent their relational patterns as a triplet <subject, relation, subject>, denoted as <S_1, R, S_2> for brevity, where S_1 and S_2 are two subjects and R is the relation [106].

To differentiate relations and subject appearances in the relational pattern, we introduce relation LoRA triplet, a composite LoRA set comprising Relation LoRAs to model relational information and two Subject LoRAs to capture appearance information, as depicted in Fig. 4. Specifically, we inject Relation LoRAs into the query and key matrices of the MM-DiT full attention. Concurrently, we design two Subject LoRAs corresponding to the two subjects involved in the relation and inject them into the value matrix. This design is motivated by our empirical findings that the query, key, and value matrices serve distinct roles within the MM-DiT full attention. More details on the analysis are provided in Sec. 3.3. Additionally, we devise an FFN LoRA to refine the outputs of the Relation and Subject LoRAs and inject it into the linear layers of full attention. Note that the two branches of text and vision tokens in MM-DiT are processed by different LoRA sets.

Hybrid mask training strategy. To achieve the decoupling of relational and appearance information in the introduced relation LoRA triplet, we propose hybrid mask training strategy (HMT) to guide Relation and Subject LoRAs to focus on designated regions using corresponding masks. We first employ Grounding DINO [50] and SAM [40] to derive masks for the two individuals in a video, indicated as Subject Masks M_{S_1} and M_{S_2}. Inspired by representative relation detection approaches [78, 101, 102] that utilize minimum enclosing rectangles to delineate subject-object interaction zones, we define the Relation Mask M_R as the union of the two Subject Masks to indicate the relation area. Since the 3D VAE in Video DiT compresses the video's temporal dimensions by a factor of T_c, we average the masks over every T_c frame to represent the latent masks.

We then devise a LoRA selection strategy and an enhanced diffusion loss for better disentanglement during training. Specifically, we randomly select either the Relation LoRAs or one type of Subject LoRAs in relation



(a) Visualization of Query, Key, and Value Features in MM-DiT Full Attention. (b) Subspace similarity among the Query, Key, and Value Matrices across different models.

Figure 5. **Features and subspace similarity analysis of MM-DiT**. (a) Value features across different videos encapsulate rich appearance information, and relational information often intertwines with these appearance cues. Meanwhile, query and key features exhibit similar patterns that differ from those of value features. (b) We perform singular value decomposition on the query, key, and value matrices of each DiT block and compute the similarity of the subspaces spanned by their top-k left singular vectors, indicating query and key matrices share more common information while remaining independent of the value matrix. This observation holds for both MM-DiT (Mochi [79] and HunyuanVideo [42]) and CrossAttn-based DiT (Wan2.1 [83]) architectures.

LoRA triplet to update for each training iteration. When the Relation LoRAs are chosen, the two Subject LoRAs are trained simultaneously to provide appearance cues, assisting the Relation LoRAs in concentrating on relational information. This process facilitates the decoupling of relational and appearance information. The FFN LoRAs are consistently engaged throughout training to refine outputs from the selected Relation or Subject LoRAs.

Following LoRA selection, we apply the corresponding masks to amplify the loss weight within the focused area, which can be defined as:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\boldsymbol{z},\epsilon,\boldsymbol{c},t}\big(\lambda_m \mathbf{M}_l + 1\big) \cdot \big\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t)\big\|_2^2, \quad (2)$$

where $l \in \{S_1, S_2, R\}$ indicates the selected mask type, and λ_m is the mask weight. By employing the LoRA selection strategy and the enhanced diffusion loss, Relation and Subject LoRAs are encouraged to concentrate on their designated area, facilitating effective relation customization and improving the generalization capacity.

Inference. During inference, we exclude Subject LoRAs to prevent undesired appearances and inject only Relation LoRAs and FFN LoRAs into the base Video DiT to maintain learned relations and enhance generalization.

3.3. Analysis on Query, Key, and Value Features

To determine the optimal model design, we analyze query, key, and value features and matrices in MM-DiT's full attention via visualization and singular value decomposition, revealing their impacts on relational video customization.

Visualization analysis. We start with two types of videos: a single-subject video with multiple attributes, and a two-subject interaction video, as illustrated in Fig. 5(a). We compute the averaged query, key, and value features across all layers and attention heads at timestep 60, focusing solely on those associated with vision tokens. These features are then reshaped into an $f \times h \times w$ format, and we visualize the averaged features across all frames with shape $h \times w$. From the observations in Fig. 5(a), we draw two conclusions: **1)** *Value features across different videos encapsulate rich appearance information, and relational information often intertwines with these appearance cues*. For instance, in the single-subject video, high-value feature responses occur at locations like "blue glasses" and "birthday hat." In the two-subject video, high values are observed both in regions of relations (*e.g.*, handshakes) and appearances (*e.g.*, human face and clothing), indicating the entanglement of relational and appearance information within the features. **2)** *Query and key features exhibit highly abstract yet similar patterns, distinctly diverging from the value features*. Unlike the obvious appearance information in value features, query, and key features exhibit homogeneity across different videos, clearly differing from value features. To further validate this point, we analyze query, key, and value matrices from a quantitative perspective.

Subspace similarity analysis. We further analyze the similarity of the subspace spanned by the singular vectors of the query, key, and value matrix weights from the base Video DiT model Mochi. This similarity reflects the degree of overlap in contained information between two matrices. For the query and key matrices, we apply singular value decomposition to obtain left-singular unitary matrices U_Q and U_K. Following [32, 52], we select the top r singular vectors from U_Q and U_K, and measure their normalized subspace similarity based on the Grassmann distance [25] using $\frac{1}{r}\left\|U_Q^{r\top} U_K^r\right\|_F^2$. The other similarities are calculated in a similar way. The results in Fig. 5(b) demonstrate that the subspaces of the query and key matrices are highly similar, whereas their similarity to the value matrix is minimal. This suggests that *the query and key matrices in MM-DiT share more common information while remaining largely independent of the value matrix*. In other words, the query and key matrices exhibit a strongly non-overlapping relationship with the value matrix, which facilitates the design of our decoupling learning. This observation aligns with the visualization results in Fig. 5(a). To further verify the generalizability of this finding, we conduct similar analyses on various models, *e.g*. HunyuanVideo [42] and

Wan2.1 [83]. The results in Fig. 5(b) indicate that *the higher similarity between the query and key matrices remains consistent across different MM-DiT models and other DiT architectures (CrossAttn-based DiT)*.

Building on these observations, we empirically argue that the query, key, and value matrices serve distinct roles in relational video customization, motivating our design of relation LoRA triplet. Specifically, given that value features are rich in appearance information, we inject Subject LoRAs into the value matrix to focus on learning appearances. In contrast, due to the homogeneity observed in the query and key features and their non-overlapping nature with the value matrix, which facilitates decoupling learning, we inject Relation LoRAs into both query and key matrices to disentangle relations from appearances. The results in Tab. 3 confirm our analysis, showing this design achieves optimal performance. We believe our findings can advance video customization research based on DiT architecture.

3.4. Relational Dynamics Enhancement

To explicitly enhance relational dynamics learning, we propose a novel space-time relational contrastive loss (RCL), which emphasizes relational dynamics while reducing the focus on detailed appearance during training. Specifically, at each timestep t, we compute the pairwise differences of the model output along the frame dimension, denoted as $\bar{\epsilon} \in \mathbb{R}^{(f-1)\times h\times w\times c}$. We then reduce dependency on pixel-level information by averaging these differences across the spatial dimensions, resulting in 1D relational dynamics features $A \in \mathbb{R}^{(f-1)\times c}$, which serve as anchor features. Subsequently, we sample n_{pos} 1D relational dynamics features from other relation videos as positive samples $P \in \mathbb{R}^{(f-1)\times n_{\text{pos}}\times c}$. For each frame in A, we sample n_{neg} 1D features from single-frame model outputs $\epsilon_i \in \mathbb{R}^{1\times h\times w\times c}$ as negative samples $N \in \mathbb{R}^{(f-1)\times n_{\text{neg}}\times c}$, which capture appearance information while excluding relational dynamics.

Our objective is to learn representations with relational dynamics by pulling together the pairwise differences from different videos depicting the same relation, while distancing them from spatial features of single-frame outputs to mitigate appearance and background leakage. Following InfoNCE [61, 64] loss, we formulate the proposed loss as:

$$\mathcal{L}_{\text{RCL}} = -\log \sum_{i=1}^{f-1} \frac{\sum_{j=1}^{n_{\text{pos}}} \exp(\frac{A_i^\top P_{ij}}{\tau})}{\sum_{j=1}^{n_{\text{pos}}} \exp(\frac{A_i^\top P_{ij}}{\tau}) + \sum_{k=1}^{n_{\text{neg}}} \exp(\frac{A_i^\top N_{ik}}{\tau})}, \quad (3)$$

where τ is the temperature hyper-parameter.

Additionally, we maintain a memory bank $\mathcal{M}$ to store and update the positive and negative samples, both randomly selected from the 1D features of the current batch videos and previously seen videos. This online dynamic update strategy can enlarge the number of positive and negative samples, enhancing the contrastive learning effect and training stability. At each iteration, we store all current anchor features A and the 1D features of ϵ_i into $\mathcal{M}$. The memory bank is implemented as a First In, First Out queue.

Overall, the training loss $\mathcal{L}_{\text{total}}$ consists of both reconstruction and contrastive learning loss, defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{RCL}}, \quad (4)$$

where λ_1 is the loss balancing weight.

4. Experiment

4.1. Experimental Setup

Datasets. We conduct experiments on the NTU RGB+D Action Recognition Dataset [49, 71]. We select 26 types of human relations, such as handshakes and hugs, each labeled with a text prompt like "A person is shaking hands with a person." For evaluation, we design 10×26 prompts with uncommon subject interactions, such as "A dog is shaking hands with a cat", to assess generalization to novel domains. More details are provided in Appendix A.

Baselines. Given the absence of existing methods for relational video customization, we define four baseline categories: *1)* The base model Mochi. *2)* Direct LoRA finetuning. *3)* Adapted relational image customization methods. We reproduce ReVersion [36] on Mochi for relational video customization. *4)* Motion customization methods, which mostly rely on Temporal Attention Layers that are absent in MM-DiT, face challenges in direct adaptation. Thus, we choose the recent and adaptable MotionInversion [85] as a baseline, reproducing it on Mochi for comparison.

Evaluation metrics. We evaluate our method by focusing on four aspects: *1)* Relation Accuracy. Instead of using biased classifiers trained on test sets with limited diversity like previous methods [23, 36], which hinders test accuracy and generalizability, we propose the Relation Accuracy metric to assess relations using *advanced Vision-Language Models (VLMs)*. Specifically, we input generated videos to Qwen-VL-Max [2], a leading VQA model [20, 53], asking if the video matches the specified relation, and converting yes/no responses into relation accuracy percentages. We repeat this process 10 times to calculate the average accuracy. *2)* Text Alignment. We employ CLIP image-text similarity (CLIP-T) to measure alignment with text prompts. *3)* Temporal Consistency, which computes the average cosine similarity across consecutive frames [17]. *4)* Video Quality. We use FVD to evaluate the video quality. The reference videos are 800 videos from the AnimalKingdom test dataset [63].

Implementation details. We adopt Mochi [79] as our base model. We use AdamW [54] optimizer with a learning rate of 2e-4 and weight decay of 0.01. The training iteration is 2400. We set LoRA rank to 16, λ_m to 50, and λ_1 to 0.01. Generated video resolution is 61×480×848, and the

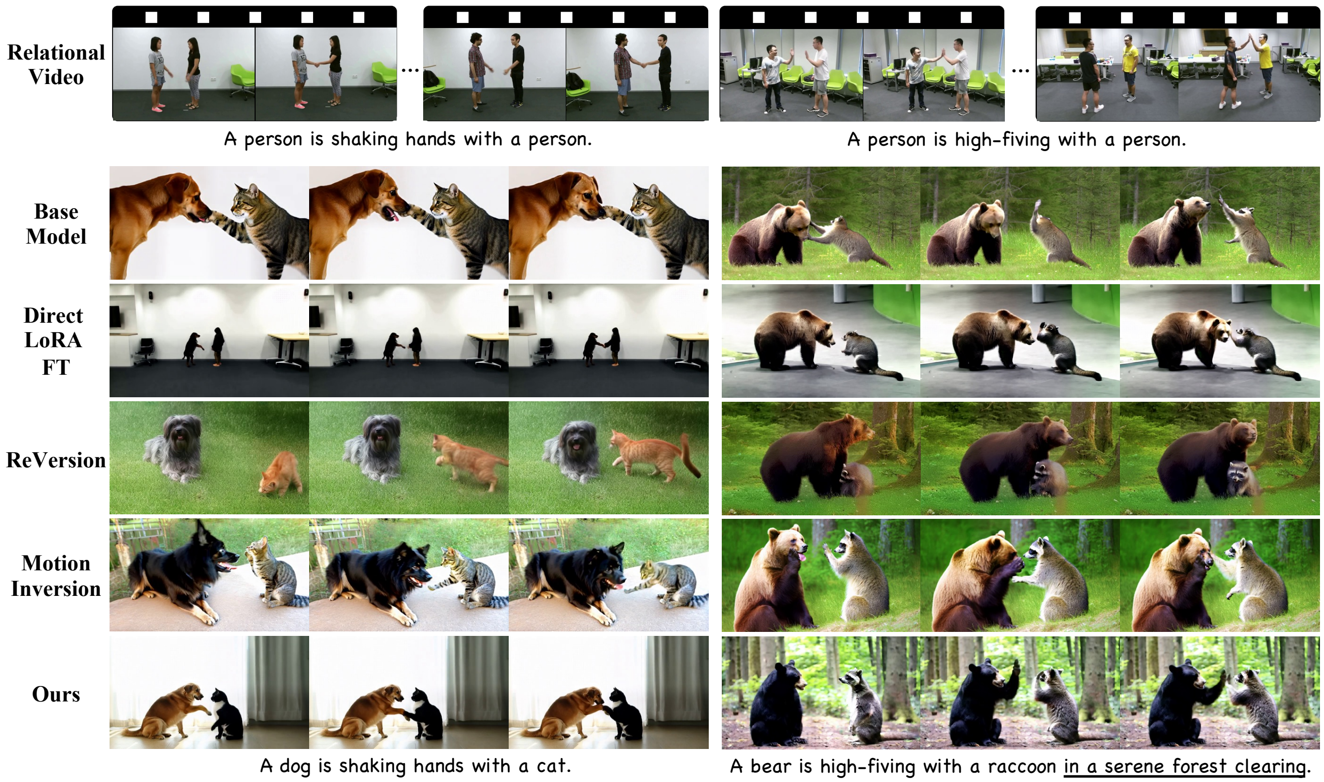


Figure 6. **Qualitative comparison results**. Our method outperforms all baselines in precisely capturing the intended relation and mitigating appearance and background leakage.

Table 1. **Quantitative comparison results.**

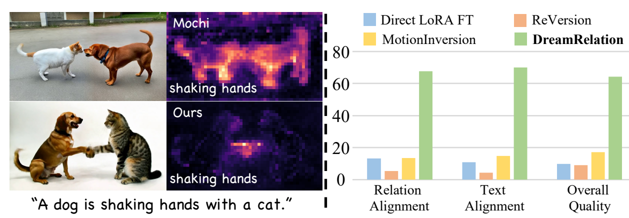
Method	Relation Accuracy	CLIP-T	Temporal Consistency	FVD↓
Mochi (base model) [79]	0.2623±0.04	0.3237	0.9888	**2047.37**
Direct LoRA finetuning	0.3258±0.05	0.2966	0.9945	2229.08
ReVersion [36]	0.2690±0.01	0.3013	0.9921	2682.69
MotionInversion [85]	0.3151±0.03	0.3217	0.9855	2084.51
DreamRelation	**0.4452**±0.01	**0.3248**	**0.9954**	2079.87

batch size is 1. We set n_{pos} to 4 and n_{neg} to 10. The memory bank size is 64, and τ is 0.07. During inference, we generate 30-fps videos using Euler Discrete method [47, 51] with 64 steps. The classifier-free guidance [27] scale is 6.0.

4.2. Main Results

Qualitative results. Qualitative comparisons in Fig. 6 reveal that all baseline methods, including the base model Mochi, fail to generate videos that match the relations defined in exemplar videos. For example, Direct LoRA finetuning struggles with appearance and background leakage, while other methods like MotionInversion cannot capture desired relational dynamics due to the complexity inherent in relations. In contrast, our DreamRelation precisely generates videos with intended relations and diverse subjects, effectively preventing appearance and background leakage.

Quantitative results. Tab. 1 presents the quantitative comparison results. Direct LoRA finetuning improves the base model's Relation Accuracy but suffers from reduced CLIP-T and FVD due to appearance leakage. Inversion-based methods like ReVersion and MotionInversion achieve better CLIP-T than finetuning but fail to model desired relations accurately. In contrast, while comparable to the base model in FVD, our DreamRelation consistently surpasses baselines across other metrics, verifying its effectiveness.



(a) Visualization of Attention Maps. (b) Human evaluation results.

Figure 7. (a) Our method focuses on the desired relational region. (b) Our method is most preferred by users across all aspects.

Attention map analysis. To verify the effectiveness of our method, we compute averaged attention maps from all layers and heads, extracting values for text tokens of relations like "shaking hands" and all vision tokens [6]. These attention maps are reshaped and visualized in Fig. 7(a). We observe that the base model's attention map for "shaking hands" is messy, leading to poor generation. In contrast, our method's attention map effectively focuses on the relational area, producing more natural results and demonstrating its capability to capture relational information.

User study. We conduct user studies to evaluate our DreamRelation, involving 15 annotators who rate 180 video

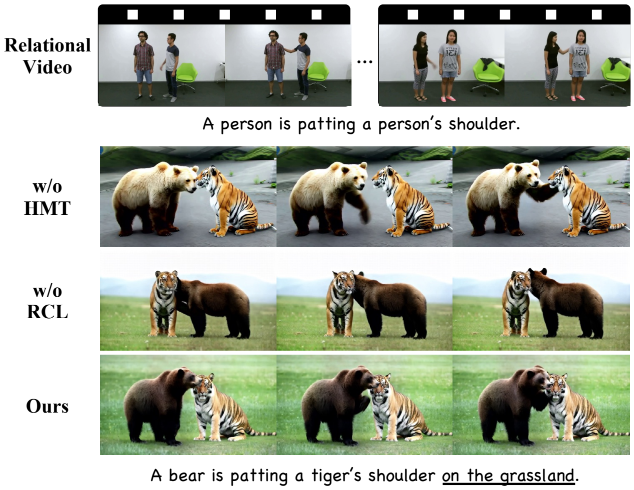


Figure 8. **Qualitative ablation study on each component.**

Table 2. **Ablation studies on effects of hybrid mask training strategy (HMT), space-time relational contrastive loss (RCL), and each type of LoRA.** Removing any of the above components significantly reduces the overall performance.

Method	Relation Accuracy	CLIP-T	Temporal Consistency	FVD↓
w/o HMT	0.3574±0.02	0.3244	0.9938	2248.52
w/o RCL	0.3416±0.03	0.3185	0.9953	2136.95
w/o Relation LoRAs	0.3626±0.02	0.3035	0.9950	2318.49
w/o Subject LoRAs	0.3769±0.04	0.3147	0.9949	2408.59
w/o FFN LoRAs	0.4021±0.03	0.3241	0.9914	2369.98
ours	**0.4452**±0.01	**0.3248**	**0.9954**	**2079.87**

groups generated by four methods. Each group contains four generated videos, a reference video, and a textual prompt. Evaluations are based on majority votes in three aspects: Relation Alignment, Text Alignment, and Overall Quality. Results in Fig. 7(b) indicate that our method is most preferred by users across all aspects. More details about the user study are provided in Appendix B.

4.3. Ablation Studies

Ablation on each component. We perform an ablation study on the effects of each component, as shown in Fig. 8. Without hybrid mask training strategy, the model generates the desired relations but experiences some background leakage due to incomplete decoupling of relational and appearance information. Omitting space-time relational contrastive loss reduces background leakage but results in videos exhibiting inaccurate relations.

Quantitative results in Tab. 2 show that removing hybrid mask training strategy or space-time relational contrastive loss degrades performance across all metrics, confirming that each component is crucial to overall performance; see Appendix C for more ablation studies.

Ablation on each LoRA in relation LoRA triplet. We conduct ablation studies to verify each LoRA's effects. The results in Tab. 2 indicate that removing Relation LoRAs or Subject LoRAs significantly reduces Relation Accuracy and

Table 3. **Ablation study of Relation LoRA position.**

Relation LoRA	Subject LoRA	Relation Accuracy	CLIP-T	Temporal Consistency	FVD↓
V	Q, K	0.3444±0.02	0.3225	0.9953	2233.48
Q	K, V	0.3921±0.03	**0.3301**	0.9951	2284.65
K, V	Q	0.3937±0.04	0.3196	**0.9954**	2180.27
Q, K	V	**0.4452**±0.01	0.3248	**0.9954**	**2079.87**

Table 4. **Effects of space-time relational contrastive loss on motion customization** method (MotionInversion).

Method	Relation Accuracy	CLIP-T	Temporal Consistency	FVD↓
MotionInversion [85]	0.3151±0.03	0.3217	0.9855	2084.51
MotionInversion + RCL	0.3633±0.05	0.3181	0.9862	2063.30

CLIP-T due to insufficient decoupling of appearance and relational information. Excluding FFN LoRAs also lowers accuracy, highlighting the need for refinement.

Ablation on Relation LoRAs position. To determine the optimal position of Relation LoRAs, we experiment with different settings in the query (Q), key (K), and value (V) matrices, as shown in Tab. 3. Inserting Relation LoRAs to the V matrix results in the lowest Relation Accuracy, likely because V features predominantly exhibit appearance information, making it challenging to accurately capture the desired relations. Placing Relation LoRAs in the Q matrix or KV matrices is suboptimal since the overlapping nature of the QK matrices hinders their ability to process different information separately, which is not conducive to decoupling relations from appearances. In contrast, inserting Relation LoRAs to the QK matrices achieves the best Relation Accuracy, consistent with our analysis of full attention in Fig. 5.

Ablation on space-time relational contrastive loss (RCL). To verify the effectiveness of RCL among different methods, we integrate it with MotionInversion [85]. Results in Tab. 4 show that incorporating RCL enhances Relation Accuracy and Temporal Consistency while maintaining comparable CLIP-T, demonstrating its potential for generalization across different methods.

5. Conclusion

In this paper, we present DreamRelation, a novel relational video customization method that accurately models complex relations defined in exemplar videos through relational decoupling learning and relational dynamics enhancement. We introduce relation LoRA triplet to decompose relations into appearance and relational information and further enhance this decoupling with hybrid mask training strategy. Our analysis of query, key, and value features in MM-DiT's full attention motivates and offers interpretability for our model design. To further enhance relation dynamics learning, we propose space-time relational contrastive loss, which prioritizes relational dynamics over detailed appearances. Extensive experimental results demonstrate the superior customization capabilities of DreamRelation.

Acknowledgements This work is partially supported by STI2030-Major Projects (No. 2021ZD0200204), National Natural Science Foundation of China (No. 62471148), and Alibaba Group. Special thanks to Zhihang Liu (University of Science and Technology of China) and Boyuan Cao (Fudan University) for sharing their invaluable experience and advice on model analysis.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6, 14

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 3

[4] 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. 3

[5] 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. 2, 3

[6] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. *arXiv preprint arXiv:2412.18597*, 2024. 7

[7] Yuanhao Cai, He Zhang, Xi Chen, Jinbo Xing, Yiwei Hu, Yuqian Zhou, Kai Zhang, Zhifei Zhang, Soo Ye Kim, Tianyu Wang, et al. Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. *arXiv preprint arXiv:2506.23361*, 2025. 3

[8] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 3

[9] 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

[10] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023. 3

[11] 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. 3

[12] Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. *arXiv preprint arXiv:2405.12796*, 2024. 3

[13] 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. 3

[14] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. *arXiv preprint arXiv:2501.06187*, 2025. 3

[15] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[16] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24209–24218, 2024. 3

[17] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3, 6

[18] 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, 4

[19] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025. 3

[20] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 6

[21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[22] Chen Gao, Si Liu, Defa Zhu, Quan Liu, Jie Cao, Haoqian He, Ran He, and Shuicheng Yan. Interactgan: Learning to generate human-object interaction. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 165–173, 2020. 3

[23] Mengmeng Ge, Xu Jia, Takashi Isobe, Xiaomin Li, Qinghe Wang, Jing Mu, Dong Zhou, Li Wang, Huchuan Lu, Lu Tian, et al. Customizing text-to-image generation with inverted interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10901–10909, 2024. 3, 6

[24] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[25] Jihun Hamm and Daniel D Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, pages 376–383, 2008. 5

[26] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 3

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[29] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[31] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6180–6189, 2024. 3

[32] 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. 2, 5

[33] Tianyu Hua, Hongdong Zheng, Yalong Bai, Wei Zhang, Xiao-Ping Zhang, and Tao Mei. Exploiting relationship for complex-scene image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1584–1592, 2021. 3

[34] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025. 3

[35] Zhizhong Huang and Xiaoming Liu. Generalizable object re-identifcation via visual in-context prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[36] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3, 6, 7, 14

[37] Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. In *European Conference on Computer Vision*, pages 358–376. Springer, 2024. 3

[38] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9212–9221, 2024. 2, 3

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

[40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4

[41] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3

[42] 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, 5, 14, 19

[43] PKU-Yuan Lab and Tuzhan AI etc. Open-sora: Democratizing efficient video production for all, 2024. https://doi.org/10. 5281/zenodo.10948109. 3

[44] Hengjia Li, Yang Liu, Yibo Zhao, Haoran Cheng, Yang Yang, Linxuan Xia, Zekai Luo, Qibo Qiu, Boxi Wu, Tu Zheng, et al. Gca-3d: Towards generalized and consistent domain adaptation of 3d generators. *arXiv preprint arXiv:2412.15491*, 2024. 3

[45] Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. *arXiv preprint arXiv:2411.17048*, 2024. 3

[46] Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. *arXiv preprint arXiv:2503.12689*, 2025. 3

[47] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 7

[48] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024. 3

[49] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 3, 6, 14

[50] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[51] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 7

[52] Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3855–3863, 2024. 5

[53] Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Boqiang Zhang, Nianzu Yang, Pandeng Li, Yinglu Li, Zuan Gao, et al. What is a good caption? a comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *arXiv preprint arXiv:2502.14914*, 2025. 6

[54] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[55] 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

[56] 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. 3

[57] 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.

[58] 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.

[59] 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

[60] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024. 3

[61] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020. 6

[62] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 2, 3

[63] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19023–19034, 2022. 6

[64] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[65] 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. 2, 3

[66] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[67] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2024. 3

[68] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2402.14780*, 2024. 3

[69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[70] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 22500–22510, 2023. 3

[71] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 3, 6, 14

[72] D She, Mushui Liu, Jingxuan Pang, Jin Wang, Zhen Yang, Wanggui He, Guanghao Zhang, Yi Wang, Qihan Huang, Haobin Tang, et al. Customvideox: 3d reference attention driven dynamic adaptation for zero-shot customized video diffusion transformers. *arXiv preprint arXiv:2502.06527*, 2025. 3

[73] Qingyu Shi, Lu Qi, Jianzong Wu, Jinbin Bai, Jingbo Wang, Yunhai Tong, Xiangtai Li, and Ming-Husan Yang. Relationbooth: Towards relation-aware customized object generation. *arXiv preprint arXiv:2410.23280*, 2024. 3, 15, 19

[74] Shuai Tan, Biao Gong, Yutong Feng, Kecheng Zheng, Dandan Zheng, Shuwei Shi, Yujun Shen, Jingdong Chen, and Ming Yang. Mimir: Improving video diffusion models for precise text understanding. *arXiv preprint arXiv:2412.03085*, 2024. 3

[75] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024.

[76] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2024.

[77] Shuai Tan, Bin Ji, and Ye Pan. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26317–26327, 2024. 3

[78] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 4

[79] Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024. 2, 3, 4, 5, 6, 7, 14

[80] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2024. 3

[81] Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024. 3

[82] 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

[83] 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. 5, 6, 15, 19

[84] 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. 3

[85] Luozhou Wang, Ziyang Mai, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*, 2024. 2, 3, 6, 7, 8, 14, 15

[86] 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*, 36:7594–7611, 2023. 3

[87] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023.

[88] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024.

[89] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6572–6582, 2024.

[90] 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. 3

[91] 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

[92] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 3

[93] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 2, 3

[94] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*, 2024.

[95] 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.

[96] Tao Wu, Yong Zhang, Xiaodong Cun, Zhongang Qi, Junfu Pu, Huanzhang Dou, Guangcong Zheng, Ying Shan, and Xi Li. Videomaker: Zero-shot customized video generation with the inherent force of video diffusion models. *arXiv preprint arXiv:2412.19645*, 2024.

[97] 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. 2, 3

[98] Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation. *arXiv preprint arXiv:2502.04299*, 2025. 3

[99] Chao Xu, Yang Liu, Jiazheng Xing, Weida Wang, Mingze Sun, Jun Dan, Tianxin Huang, Siyuan Li, Zhi-Qi Cheng, Ying Tai, et al. Facechain-imagineid: Freely crafting high-fidelity diverse talking faces from disentangled audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1292–1302, 2024. 3

[100] Chao Xu, Mingze Sun, Zhi-Qi Cheng, Fei Wang, Yang Liu, Baigui Sun, Ruqi Huang, and Alexander Hauptmann. Combo: Co-speech holistic 3d human motion generation and efficient customizable adaptation in harmony. *arXiv preprint arXiv:2408.09397*, 2024. 3

[101] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 4

[102] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 4

[103] Nianzu Yang, Pandeng Li, Liming Zhao, Yang Li, Chen-Wei Xie, Yehui Tang, Xudong Lu, Zhihang Liu, Yun Zheng, Yu Liu, and Junchi Yan. Rethinking video tokenization: A conditioned diffusion-based approach. *arXiv preprint arXiv:2503.03708*, 2025. 3

[104] 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

[105] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 3

[106] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, 35:37416–37431, 2022. 4

[107] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6463–6474, 2024. 3

[108] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024. 2, 3

[109] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 3

[110] 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

[111] Yunpeng Zhang, Qiang Wang, Fan Jiang, Yaqi Fan, Mu Xu, and Yonggang Qi. Fantasyid: Face knowledge enhanced id-preserving video generation. *arXiv preprint arXiv:2502.13995*, 2025. 3

[112] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 2, 3

[113] 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, 2024. https://github.com/hpcaitech/Open-Sora. 3

[114] Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, Nanxuan Zhao, Jing Shi, and Tong Sun. Sugar: Subject-driven video customization in a zero-shot manner. *arXiv preprint arXiv:2412.10533*, 2024. 3

[115] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 3