

SA-V2V: Training-Free Subject-Aware Video-to-Video Personalization

Soobin Park¹, Seohyeon Yoo¹, Jiwon Kim², SeonHwa Kim², Kyong Hwan Jin²^{*}, and Eunju Cha¹^{*}

¹ Department of Electrical Engineering, Sookmyung Women’s University, Seoul, South Korea

² School of Electrical Engineering, Korea University, Seoul, South Korea

Abstract. Video personalization aims to generate personalized videos by preserving the motion and background of a reference video while transferring the subject appearance from a given subject image. Such video personalization has achieved tremendous progress alongside recent advancements in video diffusion transformers (DiT). Existing DiT-based approaches predominantly rely on training-based methods that require auxiliary control signals, while independent control over motion, subject appearance, and background remains challenging due to the entangled nature of the unified spatiotemporal attention mechanism. In this paper, we uncover a key structural insight that the spatio-temporal attention matrix of DiT exhibits an inherent functional decomposition—intra-frame blocks primarily encode spatial information, while inter-frame blocks encode motion dynamics. Building on this observation, we introduce **SA-V2V**, a novel training-free video personalization framework that disentangles and enables independent control over these visual attributes. We propose two guidance approaches: (1) For motion guidance, we present stochastic temporal injection, which leverages inter-frame attention maps to precisely control dynamic trajectories. (2) For appearance guidance, we propose a **T**arget-aware **F**eature **G**uidance to enable independent control over subject and background appearance by modulating the synthesized features with masked attention features from the subject image and the reference video, respectively. Extensive experiments demonstrate that the proposed SA-V2V significantly outperforms existing approaches in motion fidelity, subject adherence, and background preservation. [Our code is available at https://parksoobin1219.github.io/SA-V2V/](https://parksoobin1219.github.io/SA-V2V/)

Keywords: Video personalization · Motion guidance · Training-free diffusion

1 Introduction

Diffusion-based text-to-video (T2V) and image-to-video (I2V) models [1, 8, 15, 30, 31, 38] have achieved remarkable success in generating high-quality videos. This

^{*} Corresponding authors.



Fig. 1: Video-to-video personalization. The proposed SA-V2V is designed to enable transferring specific subjects from the subject image (left) while maintaining the motion and background of the reference video (top).

achievement is built upon successful architectural paradigms as follows: temporal extension of 2D UNet [1, 2, 8, 31] and the subsequent transition to DiT-based architectures [15, 30, 38]. Despite their success in high-quality video synthesis, attaining precise control over pre-trained video diffusion models (VDMs) to produce videos with specific motion and subject identities remains challenging.

To enhance controllability of VDMs, various controllable video generation approaches [3, 13, 16, 19, 20, 32–35, 39] have been proposed. Current methods [3, 13, 32, 34, 35, 40] often fine-tune VDMs using LoRAs [11] or auxiliary modules to learn features from reference videos or appearance images. However, aforementioned approaches significantly limit generalizability and require additional resources for fine-tuning. To mitigate these issues, training-free approaches [16, 19, 20] have been introduced. Customized videos are synthesized by swapping internal feature maps or by injecting attention scores from reference videos during generation [6, 16]. Note that training-free methods have been primarily investigated on UNet-based architectures, which facilitate such manipulation through their explicitly decoupled spatial and temporal layers. Adapting such schemes to DiTs is challenging, as 3D full-attention mechanisms of DiT entangle spatial and temporal tokens, making it difficult to control motion and appearance independently without additional adapters or optimizations.

To address these limitations, we propose **Subject-Aware Video-to-Video personalization (SA-V2V)**, a novel training-free and inversion-free framework that enables flexible controls over individual video components within DiT-based VDMs. To evaluate this capability, we introduce a video personalization task illustrated in Fig. 1: given a reference video and a subject image, the objective is to generate a new video that faithfully preserves the motion dynamics and background of the reference while substituting only the subject’s appearance with that of the subject image. This formulation demands fine-grained, disentangled control over three distinct attributes—motion, subject identity, and background—a setting that prior works have not addressed yet.

The key insight enabling this disentanglement lies in the structure of 3D unified attention. Although DiTs jointly encode spatial and temporal informa-

tion without explicit separation, we observe that the spatio-temporal attention matrix exhibits an inherent decomposition: diagonal blocks, where tokens attend within the same frame, predominantly capture spatial appearance, while off-diagonal blocks, where tokens attend across frames, encode temporal dynamics. Building on this observation, we propose two complementary mechanisms for video personalization. For motion control, **stochastic temporal injection** is introduced, which randomly transfers motion dynamics from the reference video while preserving adaptive temporal coherence in the generated output. For appearance control, we present **Target-aware Feature Guidance (TFG)**, a mask-constrained attention modulation scheme that enables selective subject or background guidance without disrupting other characteristics. By synergistically combining the proposed motion injection with TFG, SA-V2V provides faithful video personalization while maintaining the motion and background of the reference video, without requiring per-video optimization or model fine-tuning. To our knowledge, this is the first framework to achieve training-free, disentangle control over motion and appearance within DiT-based VDMs.

Our main contributions are summarized as follows:

- We formulate a novel video personalization task that requires independent control over motion dynamics, subject appearance, and background composition, which have been underexplored in prior works.
- We propose SA-V2V, a training-free framework that enables video personalization through selective motion feature injection and Target-aware Feature Guidance (TFG).
- Extensive experiments validate the superior performance of SA-V2V in terms of motion fidelity, subject adherence, and background preservation, demonstrating the effectiveness of our disentangled control approach.

2 Related works

2.1 Video customization

While early video customization focused on temporal extensions of UNets by leveraging their decoupled spatio-temporal layers [35, 37, 40], recent shifts toward DiT architectures [15, 30, 38] require new paradigms given their unified mechanisms, where spatial and temporal tokens are processed together through 3D full attention. To manage this inherent spatio-temporal entanglement, several training-based approaches [3, 32, 36] have been introduced. JointTuner [3] introduces Gated LoRA (GLORA) to adaptively steer modality-specific representations within DiT blocks. DualReal [32] utilizes an alternating adaptation scheme and a specialized controller to resolve identity-motion conflicts between different denoising stages and DiT depths. DreamRelation [36] introduces a relation LoRA triplet to explicitly decouple complex relational dynamics from subject appearances. However, these training-based methods require intensive per-video optimization, which poses a significant computational hurdle. Furthermore, the lack of spatio-temporal decoupling in DiT architectures presents a fundamental

challenge for training-free approaches that attempt to manipulate motion and appearance independently of each other.

2.2 Consistent generation

Consistent generation aims to synthesize multiple images or videos that depict visually consistent subjects across different contexts. Early personalization-based approaches optimize a text embedding to represent the target identity [5], or directly fine-tune model parameters [5, 17, 25] or identity-specific LoRA models [26]. While effective, these methods require multiple reference images and subject-specific optimization, limiting their scalability and generalization. Training-free approaches manipulate attention mechanisms in pretrained models via concatenation [28, 41] or linear interpolation [4]. Even though these algorithms provide promising results, they may struggle with preserving identity in detail or complex scene variations, especially in video generation.

3 Method

Given a reference video $\mathbf{z}^{\text{ref}}$ and a subject image $\mathbf{z}^{\text{sub}}$, our work aims to generate a personalized video $\mathbf{z}^{\text{gen}}$ that preserves the motion and background characteristics of $\mathbf{z}^{\text{ref}}$ while incorporating the subject identity from $\mathbf{z}^{\text{sub}}$. To this end, we first analyze the unified spatio-temporal attention features in DiT-based VDMs (Sec. 3.2). Building upon this analysis, we propose a training-free, inversion-free approach, called **SA-V2V** (Sec. 3.3). In particular, we propose stochastic temporal injection for motion control and **Target-aware Feature Guidance (TFG)** for appearance guidance. While our model operates in the latent space of the diffusion model, we use the same notation for both pixel-level and latent representations throughout this paper for notational simplicity.

3.1 Preliminary

Given a video latent of shape $F \times H \times W \times C$ encoded by a 3D VAE, DiT-based VDMs patchify it with spatial patch size p , yielding sequence $\mathbf{z}_{\text{vis}} \in \mathbb{R}^{\left(F \cdot \frac{H \cdot W}{p^2}\right) \times C}$. The resulting vision tokens $\mathbf{z}_{\text{vis}}$ are concatenated with text tokens to form an input sequence of VDMs. Unlike UNet-based architectures that decouple spatial and temporal modeling through separate attention layers, DiTs process this sequence through a 3D unified attention:

$$\mathbf{h} := \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \quad (1)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are query, key, and value representations with head dimension d , respectively. In this design, each token attends to every other token across both space and time, thereby entangling appearance and temporal dynamics within a single attention matrix $\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)$.

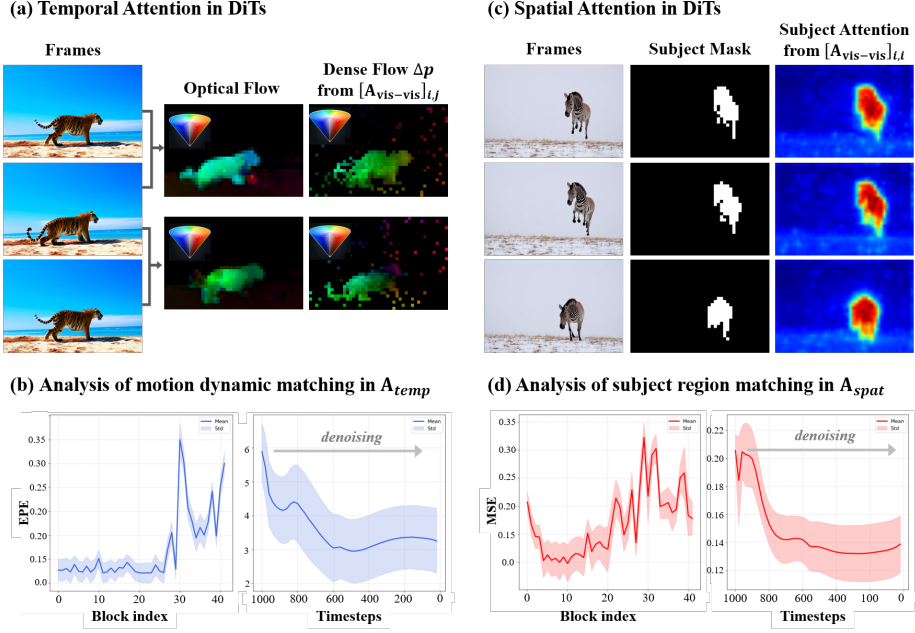


Fig. 2: Analysis of spatio-temporal attention in CogVideoX-5B [38]. (a) Inter-frame attention components (off-diagonal blocks) capture motion dynamics, showing a high correlation with dense optical flow [27]. (b) Quantitative analysis of motion matching using endpoint error (EPE) across DiT blocks and denoising timesteps. (c) Intra-frame attention components (diagonal blocks) effectively localize the subject, aligning with the SAM2 [24]-derived masks. (d) Quantitative analysis of subject region matching using mean squared error (MSE) across DiT blocks and denoising timesteps.

3.2 Motivation: Decomposability of 3D unified attention

While the unified attention matrix $\mathbf{A}$ offers greater modeling flexibility and has demonstrated superior generation quality, it raises a fundamental question: *Does the unified attention implicitly recover the spatial-temporal factorization that UNet-based models impose explicitly?*

We hypothesize that the answer is yes—that the unified attention implicitly recovers spatial-temporal factorization through the geometry of the attention matrix itself. Intuitively, tokens within the same frame share a common temporal context and are more likely to attend to one another for local appearance coherence, whereas tokens across different frames should bridge temporal distance to capture motion and correspondence. If the model has learned to respect this distinction, the attention matrix would exhibit a semantically meaningful block structure, even without any explicit architecture constraint.

Motivated by this hypothesis, we decompose the attention matrix $\mathbf{A}$ by token modality and focus on the vision-to-vision submatrix $\mathbf{A}_{vis-vis}$, which governs all interactions among visual tokens. Specifically, $\mathbf{A}_{vis-vis}$ is partitioned into $F \times F$ frame-wise blocks, where each block $[\mathbf{A}_{vis-vis}]_{i,j}$ captures attention from tokens

in frame i to those in frame j . By construction, diagonal blocks ($i = j$) aggregate intra-frame token interactions, while off-diagonal blocks ($i \neq j$) aggregate inter-frame interactions—structurally analogous to spatial and temporal attention, respectively. We verify whether this semantic correspondence holds through two experiments. For the experiment, we generate 100 videos using CogVideoX-5B [38] with diverse prompts.

Observation 1: Off-diagonal blocks encode temporal dynamics. To verify that off-diagonal blocks capture motion, we derive an attention-based dense flow $\Delta \mathbf{p}$ from the off-diagonal blocks $[\mathbf{A}_{\text{vis-vis}}]_{i,j}$ and compare it against optical flow computed by RAFT [27] on 100 videos. Specifically, for the n -th patch in frame i , we identify its most attended patch in the subsequent frame $i + 1$ as:

$$\hat{m}_n = \underset{m \in [0, \dots, N-1]}{\operatorname{argmax}} [\mathbf{A}_{\text{vis-vis}}]_{i, i+1}, \quad (2)$$

and calculate the spatial displacement vector $\Delta \mathbf{p}_n = \mathbf{p}_{\hat{m}_n} - \mathbf{p}_n$, where $\mathbf{p}$ denotes the 2D patch coordinate. Repeating this for all patches yields a dense flow $\Delta \mathbf{p}$.

As shown in Fig. 2(a), $\Delta \mathbf{p}$ exhibits strong spatial alignment with optical flow, demonstrating that off-diagonal blocks encode inter-frame motion correspondence. To further characterize when this encoding is most reliable, we measure the endpoint error (EPE) between $\Delta \mathbf{p}$ and optical flow across DiT blocks and denoising timesteps. EPE remains low in the early-to-middle blocks but rises sharply beyond the 30th block, suggesting that motion correspondence is predominantly encoded in the shallower layers (Fig. 2(b)). Across denoising timesteps, EPE is initially high but drops sharply after approximately 30% of the denoising process and remains stable thereafter. These findings confirm that off-diagonal blocks faithfully encode inter-frame motion dynamics, establishing the off-diagonal blocks $[\mathbf{A}_{\text{vis-vis}}]_{i,j}$ as an effective guidance signal for transferring temporal dynamics during the generation process.

Observation 2: Diagonal blocks encode spatial appearance. For 100 generated videos, we extract subject masks using SAM2 [24]. In addition, we accumulate $[\mathbf{A}_{\text{vis-vis}}]_{i,i}$ for query tokens corresponding to subject regions and visualize the resulting attention maps in Fig. 2(c). $[\mathbf{A}_{\text{vis-vis}}]_{i,i}$ consistently localizes to semantically meaningful spatial regions within each frame. To further quantify this correspondence, the mean squared error (MSE) between $[\mathbf{A}_{\text{vis-vis}}]_{i,i}$ and the subject mask is calculated across transformer blocks and denoising timesteps. As illustrated in Fig. 2(d), the MSE is notably lower in the early-to-middle blocks, indicating that spatial appearance is most faithfully encoded in these layers. Furthermore, the MSE decreases progressively as denoising proceeds, reflecting that the coarse spatial appearance is determined in the early denoising steps and finer details are progressively refined as the generation converges. Together, these results confirm that diagonal blocks of the unified attention encode spatial appearance within the individual frames.

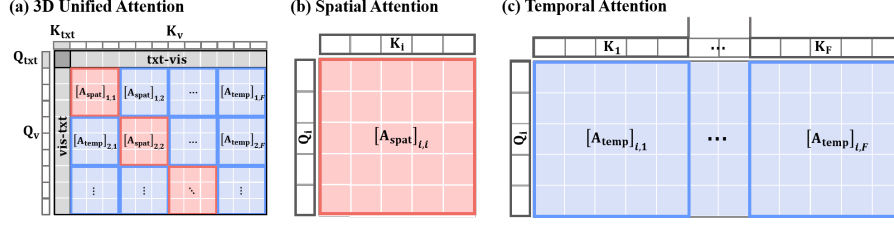


Fig. 3: 3D unified attention in DiTs. (a) The standard unified attention jointly attends over all spatio-temporal tokens in a single matrix. (b) Intra-frame diagonal block ($[\mathbf{A}_{\text{vis-vis}}]_{i,i}$) is defined as spatial attention $\mathbf{A}_{\text{spat}}$ for capturing local appearance. (c) Inter-frame off-diagonal blocks ($[\mathbf{A}_{\text{vis-vis}}]_{i,j}$) effectively act as temporal attention $\mathbf{A}_{\text{temp}}$ for encoding motion dynamics.

These findings consistently hold across other DiT-based architectures. Additional analysis on LTX-Video [9] and WAN 2.2 [30] is provided in the supplementary A.2.

3.3 SA-V2V: Subject-Aware Video-to-Video personalization

Spatio-temporal decomposition of unified attention. Building on the observation in Sec. 3.2 that intra-frame and inter-frame attention encode disentangled roles—appearance and motion dynamics, respectively—we propose to decompose $\mathbf{A}_{\text{vis-vis}}$ into **spatial attention** $\mathbf{A}_{\text{spat}}$ and **temporal attention** $\mathbf{A}_{\text{temp}}$. As illustrated in Fig. 3, $\mathbf{A}_{\text{spat}}$ is defined as a block-diagonal matrix capturing local appearance relationships, while $\mathbf{A}_{\text{temp}}$ aggregates all off-diagonal blocks to represent inter-frame motion dynamics. This decomposition allows us to treat the unified attention mechanism as a functionally factorized operator, providing a precise interface for disentangled video manipulation without retraining the backbone or additional adapters.

Stochastic temporal injection. For motion guidance, we define the temporal guidance feature $\mathbf{h}_{\text{temp}}^{\text{ref}}$ for the i -th frame as:

$$[\mathbf{h}_{\text{temp}}]_i^{\text{ref}} = \left[[\mathbf{A}_{\text{spat}}]_{i,i}^{\text{gen}} \{ [\mathbf{A}_{\text{temp}}]_{i,j}^{\text{ref}} \}_{j \in \{0, F-1\} \setminus i} \right] \left[\begin{array}{c} \mathbf{V}_i^{\text{gen}} \\ \{ \mathbf{V}_j^{\text{ref}} \}_{j \in \{0, F-1\} \setminus i} \end{array} \right], \quad (3)$$

which combines the identity context of $\mathbf{z}^{\text{gen}}$ with the motion trajectories of $\mathbf{z}^{\text{ref}}$. This design facilitates motion transfer while inherently preventing appearance leakage from $\mathbf{z}^{\text{ref}}$. However, $\mathbf{A}_{\text{temp}}$ also concentrates on spatial regions undergoing significant movements in the given video, due to the inherent spatial-temporal coherence of video. This entanglement poses a fundamental challenge for video personalization—when the subject in $\mathbf{z}^{\text{sub}}$ appears at different positions or scales than the moving object in $\mathbf{z}^{\text{ref}}$, fully injecting $\mathbf{h}_{\text{temp}}^{\text{ref}}$ forces motion patterns that are spatially misaligned with the subject’s configuration. For example, as shown

in Fig. 5, parts of the inserted subject (wolf), may extend beyond the spatial region occupied by the original object (white tiger), requiring adaptive motion adjustment rather than direct movement replication.

To resolve this, we propose **stochastic temporal injection**, which applies a random binary mask M_{temp} to sparsely integrate motion features:

$$\mathbf{h}_{temp}^{gen} \leftarrow M_{temp} \odot \mathbf{h}_{temp}^{ref} + (1 - M_{temp}) \odot \mathbf{h}_{temp}^{gen}, \quad (4)$$

where each element of M_{temp} is sampled independently from a Bernoulli distribution with probability p , and $\odot$ denotes element-wise multiplication. This stochastic injection facilitates sparse motion guidance while maintaining sufficient flexibility for the model to adjust dynamics according to the subject’s position and size. In addition, $\mathbf{h}_{temp}^{ref}$ is injected during the first T_{temp} denoising steps, which is grounded in our empirical analysis in Fig. 2(b). The global temporal dynamics are primarily established in the early steps, while the later phase is dedicated to fine-grained spatial refinement rather than motion redirection. Consequently, our scheduling ensures effective motion transfer while allowing the model to focus on high-fidelity visual synthesis as it converges. Further empirical analysis of the proposed injection scheme is presented in the supplementary Sec. C.

Target-aware feature guidance. In our video personalization formulation, it is crucial to preserve the background characteristics of $\mathbf{z}^{ref}$ while substituting its moving object with the subject from $\mathbf{z}^{sub}$. To achieve this disentanglement, we decompose the appearance of $\mathbf{z}^{gen}$ into subject appearance and background appearance. While stochastic temporal injection effectively governs the dynamic trajectories, precisely controlling these decoupled appearances requires careful intervention to maintain high-fidelity identity across the entire temporal frames. Intuitively, one might attempt to restrict the appearance guidance strictly to the intra-frame spatial attention $\mathbf{A}_{spat}$. Although this local adjustment effectively controls the appearance within individual frames, it inherently disrupts the temporal coherence of the subject’s identity across the video.

To address this issue, we compute the attention by combining $\mathbf{Q}^{gen}$ of $\mathbf{z}^{gen}$ with $\mathbf{K}^{tgt}$ and $\mathbf{V}^{tgt}$ of the target appearance across the entire unified attention mechanism, formulated as follows:

$$\mathbf{h}_{app}^{tgt} = \text{Attn}(\mathbf{Q}^{gen}, \mathbf{K}^{tgt}, \mathbf{V}^{tgt}), \quad (5)$$

where the target indicates either the subject from $\mathbf{z}^{sub}$ or background from $\mathbf{z}^{ref}$. Even though $\mathbf{h}_{app}^{tgt}$ provides globally consistent target appearance, directly substituting the generation feature with $\mathbf{h}_{app}^{tgt}$ faces fundamental challenges. The moving object in $\mathbf{z}^{ref}$ and the subject in $\mathbf{z}^{sub}$ often exhibit spatial misalignment in terms of position, size, or extent, as well as semantic variations, such as pose or viewpoint.

To achieve precise control of appearance, we introduce **Target-aware Feature Guidance (TFG)**, which enables spatially disentangled appearance control by

selectively modulating $\mathbf{h}^{\text{gen}}$ using $\mathbf{h}_{app}^{\text{sub}}$ or $\mathbf{h}_{app}^{\text{ref}}$. We first design TFG for conditioning subject appearance of $\mathbf{z}^{\text{sub}}$ to $\mathbf{z}^{\text{gen}}$, named Subject-TFG (**S-TFG**). Drawing on the consistent generation [4, 28], we modulate appearance via linear interpolation of attention:

$$\tilde{\mathbf{h}}_{app}^{\text{gen}} = c^{\text{sub}} \cdot \mathbf{h}_{app}^{\text{sub}} + (1 - c^{\text{sub}}) \cdot \mathbf{h}^{\text{gen}}, \quad (6)$$

where $\tilde{\mathbf{h}}_{app}^{\text{gen}}$ denotes the guided features and $c^{\text{sub}} \in (0, 1)$ is the guidance scale. For elaborate region-specific control, the spatial masking is further incorporated. The subject mask M^{sub} is estimated from $\mathbf{z}^{\text{sub}}$ using SAM2 [24] and resized to the resolution of attention features. With this mask, Eq. (6) is reformulated for selective appearance modulation:

$$\hat{\mathbf{h}}_{app}^{\text{gen}} = c^{\text{sub}} M^{\text{sub}} \cdot \mathbf{h}_{app}^{\text{sub}} + (1 - c^{\text{sub}} M^{\text{sub}}) \cdot \mathbf{h}^{\text{gen}}. \quad (7)$$

By constraining guidance to masked regions, S-TFG ensures that only subject appearance from $\mathbf{z}^{\text{sub}}$ is consistently transferred to $\mathbf{z}^{\text{gen}}$.

Analogously, background appearance of $\mathbf{z}^{\text{ref}}$ can be preserved in $\mathbf{z}^{\text{gen}}$ through Background-TFG (**B-TFG**), which applies the same masked interpolation framework to background regions:

$$\hat{\mathbf{h}}_{app}^{\text{gen}} = c^{\text{bg}} M^{\text{bg}} \cdot \mathbf{h}_{app}^{\text{ref}} + (1 - c^{\text{bg}} M^{\text{bg}}) \cdot \mathbf{h}^{\text{gen}}, \quad (8)$$

where c^{bg} denotes the guidance scale, and M^{bg} is the background mask extracted from $\mathbf{z}^{\text{ref}}$. $\mathbf{K}^{\text{ref}}$ and $\mathbf{V}^{\text{ref}}$ denotes keys and values extracted from $\mathbf{h}_{app}^{\text{ref}}$. Thanks to B-TFG, the background characteristics of $\mathbf{z}^{\text{ref}}$ can be faithfully preserved in $\mathbf{z}^{\text{gen}}$.

To ensure generation stability, S-TFG and B-TFG are applied alternately across denoising timesteps. This alternating scheme prevents interference from simultaneous guidance, allowing for a harmonious control of both identity and background. TFG is performed during the first T_{TFG} steps, during which appearance attributes are predominantly established.

Overall framework Figure 4 illustrates the overall pipeline of our SA-V2V. For the initial generation latent $\mathbf{z}^{\text{gen}}$, we replicate a single random noise F times along the frame axis, where F denotes the number of frames in $\mathbf{z}^{\text{ref}}$, ensuring temporal consistency. During reverse diffusion, we synergistically modulate the attention features to control motion and appearance independently of each other, yet harmoniously. Motion dynamics of $\mathbf{z}^{\text{ref}}$ are transferred by injecting temporal attention feature $\mathbf{h}_{temp}^{\text{ref}}$ according to the stochastic scheme. For appearance, rather than applying multiple constraints simultaneously, the proposed S-TFG and B-TFG are utilized in an alternating manner at each timestep. This alternating strategy prevents potential feature interference, allowing both the subject identity of $\mathbf{z}^{\text{sub}}$ and background context of $\mathbf{z}^{\text{ref}}$ to be properly reflected in $\mathbf{z}^{\text{gen}}$. Finally, the denoised latent is decoded via the pretrained 3D VAE to generate the personalized video.

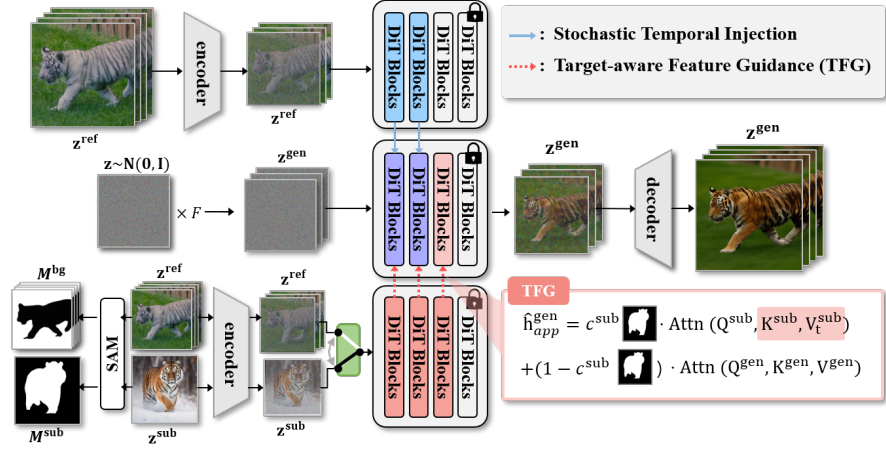


Fig. 4: Flowchart of SA-V2V. Temporal attention features extracted from the reference video are stochastically injected into intermediate transformer blocks to transfer motion dynamics. For appearance control, S-TFG and B-TFG are applied in an alternating manner at each timestep. The figure illustrates a specific timestep where S-TFG is active.

4 Experiments

Dataset. We construct a comprehensive evaluation dataset from DAVIS dataset [22], the Pexels library, and the VBench dataset [12]. These sources cover a wide range of realistic scenarios, ensuring that our benchmark covers various motion dynamics and content categories. Our evaluation set comprises 49 reference videos—25 animal, 14 human, and 5 human-object interaction videos—paired with diverse subjects, which results in 200 subject-reference pairs.

Implementation Details. We employ CogVideoX-5B [38] as our base model. We set the guidance duration as follows: $T_{temp} = 0.3T$, and $T_{TFG} = 0.25T$, where T is the total number of denoising steps. For the TFG guidance scale, we use $c^{bg} = c^{sub} = 0.5$ for all samples. Detailed settings, including control blocks, can be found in the supplementary Sec. B.

Baselines. Due to the lack of methods directly addressing the proposed video personalization task, we compare SA-V2V against closely related approaches: CtrlX [18], MotionClone [19], DiTFlow [21], Disentangled S2V [14], JointTuner [3], and DualReal [32]. For CtrlX [18], which was originally proposed for controllable image generation with a 2D UNet backbone, we extend its framework to operate on a 3D UNet for video generation. For DiTFlow [21] and Disentangled S2V [14], which are designed to control only motion and subject appearance, respectively, we use the reference video and subject image solely for their intended purposes, while the remaining visual attributes are specified through text prompts. Since MotionClone [19], DualReal [32], and JointTuner [3] are designed to support joint conditioning on both reference video and subject image

inputs, we follow the generation procedures described in their respective papers. Note that two training-free algorithms, CtrlX [18] and MotionClone [19], perform guidance via UNet feature-based manipulation and thus adopt AnimateDiff [8] as their base model. The remaining baselines and the proposed SA-V2V are implemented on CogvideoX-5B [38], one of the DiT-based VDMs. All experiments were conducted on NVIDIA A100-80GB GPUs.

Evaluation metrics. We employ eight metrics to evaluate personalization fidelity and overall video quality. To assess personalization fidelity, we utilize CLIP-I scores [23] to assess two decoupled aspects: (1) Subject adherence, evaluating the identity consistency between the generated subject and the given subject image, and (2) background preservation, evaluating the contextual fidelity of the background against the reference video. In addition, to evaluate the overall spatiotemporal realism and dynamic similarity, we calculate the Fréchet Video Distance [29] between the generated and reference videos. While FVD captures the overall distributional gap between generated and real videos, we additionally adopt five metrics from VideoScore [10] to provide fine-grained assessment of perceptual quality and temporal coherence.

5 Results

5.1 Qualitative evaluation

We present the qualitative comparisons between the proposed SA-V2V and the baselines in Fig. 5. CtrlX [18] follows the reference motion, but struggles to reflect proper subject appearance and background details. In particular, it often produces overly darkened backgrounds, suggesting that its feature manipulation mechanism can compromise the overall visual quality of the generated video. MotionClone [19] incorporated the subject image via SparseCtrl [7], but this coupling makes it difficult to achieve independent control over appearance and motion, resulting in videos where either the subject identity or the motion dynamics is poorly preserved. DiTFlow [21] preserves the motion of the reference video, while showing limited capability in personalizing the subject. Although Disentangled S2V [14] provides the intended subject appearance, it often fails to transfer the motion dynamics from the reference video. JointTuner [3] and DualReal [32] exhibit high variance across samples and occasionally fail to precisely composite the subject onto the intended spatial region of the scene. These failure cases underscore the inherent difficulty of our proposed video personalization task, which requires disentangled and accurate control over subject appearance, motion dynamics, and background composition simultaneously. In contrast, the proposed SA-V2V consistently provides high-quality results across diverse personalization scenarios. By decomposing spatiotemporal attention into spatial and temporal components and applying stochastic temporal injection, our method effectively transfers motion dynamics from the reference video while avoiding spatial misalignment with the subject. Furthermore, TFG enables independent and region-specific control over subject appearance and background

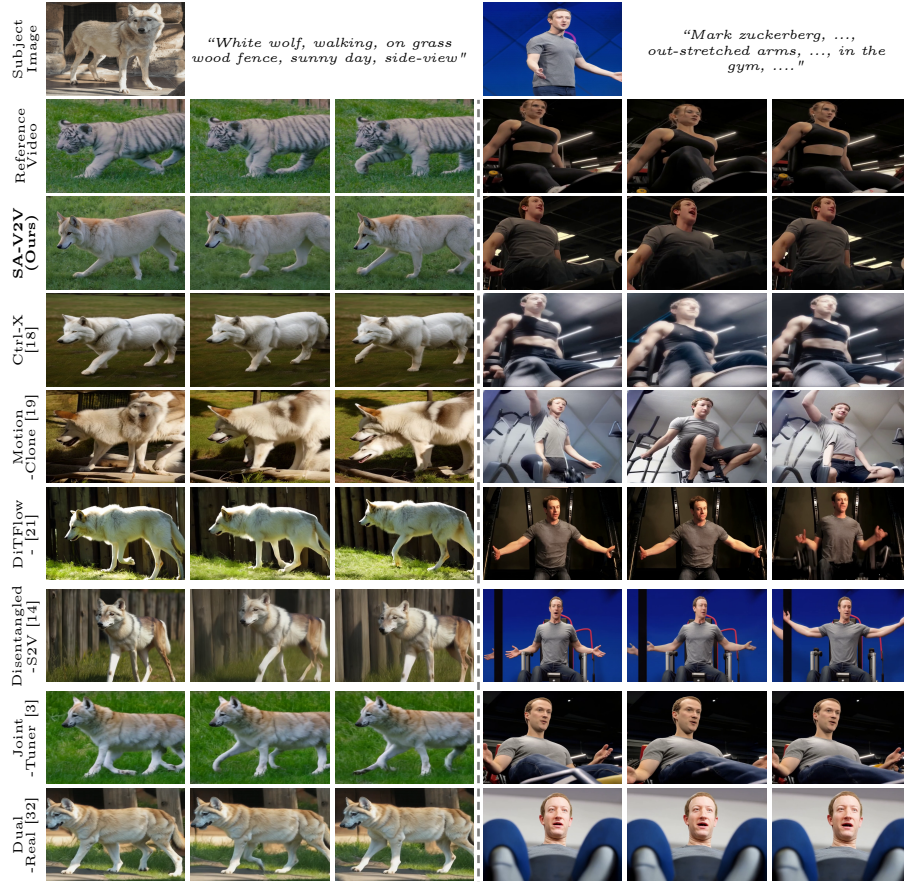


Fig. 5: Qualitative results of video personalization. SA-V2V successfully preserves reference motion and background while transferring subject appearance, demonstrating superior performance compared to existing methods.

composition, achieving a significantly better balance among motion fidelity, subject consistency, and background preservation compared to all baselines.

5.2 Quantitative evaluation

Table 1 shows the superior performance of the proposed SA-V2V for video personalization. Our method achieves the best performance for background preservation (0.8768), motion similarity (66.5532), and four video quality scores (average Videoscore 2.7646). While MotionClone achieves higher subject fidelity, it suffers from significantly degraded background preservation and video quality. This indicates that its image adapter overfits to the subject condition, failing to achieve spatially disentangled personalization. SA-V2V, in contrast, demonstrates robust and superior performance across all metrics, quantitatively validating our disentangled control framework. To further evaluate the effectiveness

Table 1: Quantitative results. SA-V2V demonstrates better quantitative performance compared to the baselines.

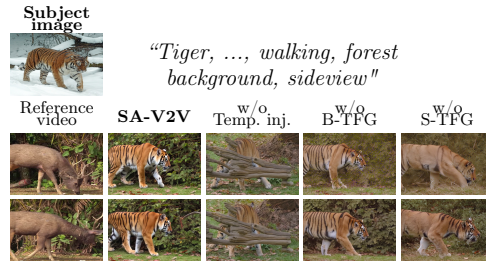
Method	Masked CLIP		FVD ($\downarrow$)	Videoscore [10] (%)					
	Subject ($\uparrow$)	Background ($\uparrow$)		VQ ($\uparrow$)	TC ($\uparrow$)	DD ($\uparrow$)	TVA ($\uparrow$)	FC($\uparrow$)	Average ($\uparrow$)
CtrlX [18]	0.7491	<u>0.8503</u>	<u>67.8538</u>	2.2641	2.0639	3.0562	2.8353	2.0133	2.5854
MotionClone [19]	0.8243	0.7604	85.8680	2.4075	1.9452	3.0892	2.7615	1.9115	2.4230
DualReal [32]	0.8116	0.7984	83.4813	<u>2.6268</u>	<u>2.5410</u>	3.0823	<u>2.9646</u>	<u>2.4525</u>	<u>2.7415</u>
JointTuner [3]	0.8079	0.8240	74.7159	2.5522	2.4721	<u>3.0950</u>	2.9968	2.3929	2.7036
SA-V2V (Ours)	<u>0.8140</u>	0.8768	66.5532	2.6524	2.5491	3.1491	2.8731	2.4995	2.7646

Table 2: Results of user study. The results show that our SA-V2V achieved the highest score in all evaluation criteria for video personalization.

Method	Background ($\uparrow$)	Motion ($\uparrow$)	Subject ($\uparrow$)	Text ($\uparrow$)	Video Quality ($\uparrow$)
CtrlX	0.046	<u>0.106</u>	0.080	0.084	0.058
MotionClone	0.011	0.029	0.034	0.009	0.017
DualReal	0.080	0.095	<u>0.259</u>	<u>0.113</u>	<u>0.121</u>
JointTuner	<u>0.101</u>	0.080	0.132	0.087	0.084
SA-V2V (Ours)	0.761	0.690	0.494	0.707	0.721

of SA-V2V, we designed a user study with 29 participants for four baselines and the proposed method. Our algorithm aims to transfer a specific subject while maintaining the motion and background of the reference video. Thus, participants were asked to select the best result across five evaluation criteria: (1) *preserving the background of the reference video*, (2) *preserving the motion dynamics of the reference video*, (3) *retaining the appearance of the subject*, (4) *alignment with the text prompt*, and (5) *overall quality of the video*.

Note that there is no well-established metric to quantitatively evaluate motion and subject preservation, which involves semantically coherent motion and consistent subject identity rather than mere pixel movements or superficial appearance maintenance. Therefore, the user study results for these criteria are particularly valuable. As shown in Tab. 2, our method outperforms existing approaches across all evaluation criteria. Notably, our method provides significantly improved performance in motion preservation, substantially outperforming competing methods.

**Fig. 6: Ablation study on the proposed components.** We evaluate variants of SA-V2V by individually omitting stochastic temporal injection, B-TFG, and S-TFG.

6 Discussion

6.1 Components of SA-V2V

We validate the contribution of each component of SA-V2V through ablation studies both qualitatively (Fig. 6) and quantitatively (Tab. 3). As shown in Fig. 6, removing stochastic temporal injection causes the model to fail in cloning the reference motion at all. Without TFG for appearance guidance, the synthesized subject no longer reflects the appearance of the subject image.

The quantitative results in Tab. 3 further support these findings. In particular, removing temporal injection leads to a substantial degradation in FVD. These observations indicate that both components are essential for disentangled video personalization.

6.2 TFG application scheme

Figure 7 demonstrates the ablation study on our TFG application scheme. Note that SA-V2V employs both spatial masking and the alternating timestep-wise guidance strategy. Without masking, the appearance and reference video features are not properly separated. When TFG is alternatively applied in a block-wise manner instead of timestep-wise, features within the same block receive conflicting signals, leading to inaccurate appearance guidance. The last column of Fig. 7 shows the result of directly applying spatial feature injection instead of linear interpolation. Due to the strong constraints of the subject image, the reference background is overwritten, preserving only the subject features. This ablation study clearly demonstrates the effectiveness of the proposed TFG design.

6.3 Computational cost

To evaluate the computational efficiency, we assess the GPU peak memory usage, training time, and inference time. For a fair comparison, we compare SA-V2V with DualReal [32] and JointTuner [3], as all methods utilize the same CogvideoX-5B [38] backbone. As reported in Tab. 4, SA-V2V achieves the lowest peak memory footprint at 19.11 GB. In addition, as a fully training-free framework, SA-V2V reduces the per-video training time to zero, whereas the baselines require 20 to 40 minutes of per-video optimization. While our alternating TFG mechanism introduces a marginal inference overhead of approximately 5–6 seconds, the total wall-clock time to obtain a personalized video is drastically reduced, demonstrating the superior practical efficiency of the proposed SA-V2V.

Table 3: Ablation study on each control scheme. The proposed SA-V2V demonstrates balanced performance across all metrics.

Method	Subject ($\uparrow$)	Background ($\uparrow$)	FVD ($\downarrow$)
w/o temp	0.8252	0.8146	103.82
w/o B-TFG	<u>0.8257</u>	0.8313	97.27
w/o S-TFG	0.8101	0.9041	72.30
SA-V2V (Ours)	0.8312	<u>0.8783</u>	<u>82.84</u>

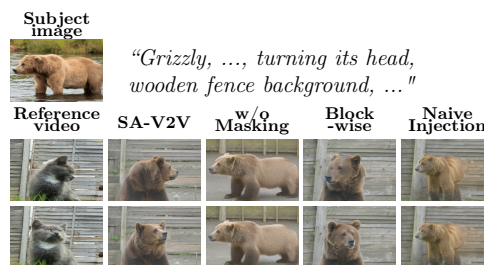


Fig. 7: Ablation study on TFG application scheme. The proposed design of TFG enables stable and effective control of appearance.

Table 4: Computational cost comparison. SA-V2V balances efficiency and quality optimally.

Method	Peak Memory (GB)($\downarrow$)	Training Time (m)($\downarrow$)	Inference Time (s)($\downarrow$)
DualReal	21.72	41.21	119.27
JointTuner	20.38	23.37	120.34
SA-V2V (Ours)	19.11	-	125.38

7 Limitation

SA-V2V is designed for personalizing videos with a single subject, where the appearance of a subject is transferred from the subject image to synthesized video. The current TFG cannot distinguish between multiple objects in the reference video that re-



Fig. 8: Limitation of SA-V2V. SA-V2V struggles to personalize videos with multiple subjects.

quire independent appearance, as shown in Fig. 8. While TFG can be conceptually extended to multi-subject settings by creating the subject-specific masks, practical implementation would require overcoming challenges in attention modulation and potential interference between subjects. Our future direction will explore personalization for multiple subjects, which will lead to improved generalization.

8 Conclusion

In this paper, we present a training-free approach for video personalization that achieves disentangled control over motion dynamics, subject appearance, and background characteristics. Through the systematic analysis of DiT-based VDM components, the proposed method employs selective feature injection for motion guidance and introduces Target-aware Feature Guidance (TFG) for independent appearance control of subject and background. Unlike existing approaches that require per-video optimization or suffer from appearance-motion entanglement, our method enables precise manipulation of individual visual attributes. Extensive experiments demonstrate that our algorithm significantly outperforms existing methods in subject identity preservation, motion coherence, and overall visual quality.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00357197, RS-2024-00335741) and Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Companion Animal Intractable Disease Overcoming R&D Project funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (RS-2026-25529322).

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling

- latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
 3. Chen, F., Zhao, S., Xu, C., Lan, L.: Jointtuner: Appearance-motion adaptive joint training for customized video generation. arXiv preprint arXiv:2503.23951 (2025)
 4. Fan, J., Xue, H., Zhang, Q., Chen, Y.: Refdrop: Controllable consistency in image or video generation via reference feature guidance. *Advances in Neural Information Processing Systems* **37**, 33602–33637 (2024)
 5. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
 6. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
 7. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: European Conference on Computer Vision. pp. 330–348. Springer (2024)
 8. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
 9. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
 10. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 2105–2123 (2024)
 11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *ICLR* **1**(2), 3 (2022)
 12. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
 13. Jeong, H., Park, G.Y., Ye, J.C.: 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. pp. 9212–9221 (2024)
 14. Kim, D., Zhang, J., Jin, W., Cho, S., Dai, Q., Park, J., Luo, C.: Subject-driven video generation via disentangled identity and motion. arXiv preprint arXiv:2504.17816 (2025)
 15. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
 16. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. arXiv preprint arXiv:2403.14468 (2024)
 17. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)

18. Lin, K.H., Mo, S., Klingher, B., Mu, F., Zhou, B.: Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *Advances in Neural Information Processing Systems* **37**, 128911–128939 (2024)
19. Ling, P., Bu, J., Zhang, P., Dong, X., Zang, Y., Wu, T., Chen, H., Wang, J., Jin, Y.: Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338* (2024)
20. Liu, Y., Sun, Y., Xing, Z., Gao, J., Chen, K., Pei, W.: Motionshot: Adaptive motion transfer across arbitrary objects for text-to-video generation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11861–11871 (2025)
21. Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 22911–22921 (2025)
22. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
24. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024)
25. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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*. pp. 22500–22510 (2023)
26. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2022. URL <https://github.com/cloudfans/simlora> **2**, 5 (2022)
27. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: *European conference on computer vision*. pp. 402–419. Springer (2020)
28. Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., Atzmon, Y.: Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)* **43**(4), 1–18 (2024)
29. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
30. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025)
31. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023)
32. Wang, W., Huang, M., Tu, Y., Mao, Z.: Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16565–16575 (2025)
33. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems* **36**, 7594–7611 (2023)
34. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: *ACM SIGGRAPH 2024 Conference Papers*. pp. 1–11 (2024)

35. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6537–6549 (2024)
36. Wei, Y., Zhang, S., Yuan, H., Gong, B., Tang, L., Wang, X., Qiu, H., Li, H., Tan, S., Zhang, Y., et al.: Dreamrelation: Relation-centric video customization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12381–12393 (2025)
37. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)
38. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
39. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
40. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J.W., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. In: European Conference on Computer Vision. pp. 273–290. Springer (2024)
41. Zhou, Y., Zhou, D., Cheng, M.M., Feng, J., Hou, Q.: Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems* **37**, 110315–110340 (2024)