Shaping a Stabilized Video by Mitigating Unintended Changes for Concept-Augmented Video Editing

Mingce Guo[1], **Jingxuan He**[1], **Yufei Yin**[2], **Zhangye Wang**[1], **Shengeng Tang**[3], **Lechao Cheng**[3✉]
[1]Zhejiang University
[2]Hangzhou Dianzi University
[3]Hefei University of Technology
guomingce@zju.edu.cn, jingxuan.he@zju.edu.cn, yinyf@hdu.edu.cn,
zywang@cad.zju.edu.cn, tangsg@hfut.edu.cn, chenglc@hfut.edu.cn

Abstract

Text-driven video editing powered by generative diffusion models holds significant promise for applications spanning film production, advertising, and beyond. However, the limited expressiveness of pre-trained word embeddings often restricts nuanced edits, especially when targeting novel concepts with specific attributes. In this work, we present a novel Concept-Augmented Textual Inversion (CATI) framework that flexibly integrates new object information from user-provided concept videos. By fine-tuning only the V (Value) projection in attention via Low-Rank Adaptation (LoRA), our approach preserves the original attention distribution of the diffusion model while efficiently incorporating external concept knowledge. To further stabilize editing results and mitigate the issue of attention dispersion when prompt keywords are modified, we introduce a Dual Prior Supervision (DPS) mechanism. DPS supervises cross-attention between the source and target prompts, preventing undesired changes to non-target areas and improving the fidelity of novel concepts. Extensive evaluations demonstrate that our plug-and-play solution not only maintains spatial and temporal consistency but also outperforms state-of-the-art methods in generating lifelike and stable edited videos. The source code is publicly available at https://guomc9.github.io/STIVE-PAGE/.

1 Introduction

Text-driven video editing, powered by generative diffusion models [Ho *et al.*, 2020], [Song *et al.*, 2020], [Rombach *et al.*, 2021], has emerged as a transformative technology with broad applications in film, art, and advertising [Ho *et al.*, 2022], [Hong *et al.*, 2022], [Blattmann *et al.*, 2023]. Recent advancements, such as Tune-A-Video [Wu *et al.*, 2023], FateZero [Qi *et al.*, 2023], and VideoComposer [Wang *et al.*, 2024], have significantly enhanced the ability to edit objects, backgrounds, and styles in video while preserving overall scene consistency through optimized attention mechanisms and spatiotemporal continuity. Despite these successes, existing methods are constrained by the limited expressiveness of CLIP [Radford *et al.*, 2021] word embeddings, which restricts their ability to perform nuanced edits on targets with specific attributes. Moreover, modifications to the target prompt often disrupt attention mechanisms, leading to inconsistencies in non-target areas before and after editing.

Inspired by Textual Inversion [Gal *et al.*, 2022], a feasible approach is to leverage external concept word embeddings, which are optimized within CLIP text encoder [Radford *et al.*, 2021] while keeping the diffusion model's parameters frozen. This technique allows the model to incorporate user-provided custom images for guided editing. However, the conventional Textual Inversion faces significant limitations when applied to video editing. Specifically, it lacks the ability to capture novel object information from arbitrary concept videos, resulting in word embeddings with insufficient fidelity to accurately describe target objects. Consequently, directly applying Textual Inversion to one-shot video editing often fails to generate satisfactory results for novel concept pairs, highlighting the need for a more robust and adaptive solution.

To this end, we propose **Concept-Augmented Textual Inversion** to enable one-shot flexible video editing based on external word embedding and target video. Specifically, we employ cutting-edge LoRA (Low-Rank Adaptation) modules to fine-tune attention value weights, focusing exclusively on the V (Value) projection (we elaborate on the rationale for tuning only V in subsequent sections). This approach effectively maintains the advantages of low VRAM overhead during tuning while preserving the plug-and-play capabilities of the model. In the context of V projection LoRA, our primary objective is to perform inversion while integrating novel object information from arbitrary concept videos for one-shot video editing. The textual inversion process relies on the pre-trained denoising network's established text-image attention probability distribution to achieve accurate target representation. By fine-tuning only the V weights—rather than both Q (Query) and K (Key)—we enable the direct integration of new feature representations while minimizing disruptions to the pre-trained attention distribution. This selective fine-tuning strategy ensures stable training during the early stages, as it suppresses unnecessary changes to the model's foundational attention mechanisms. In addition, we introduce a **Dual Prior Supervision (DPS)** mechanism, designed to stabilize the generated video by supervising the cross-attention between the source and target prompts. This mechanism addresses the issue of attention dispersion, which often arises when modifications are made to the target prompt. By effectively controlling the attention distribution, DPS significantly enhances the consistency of non-target areas before and after video editing. Furthermore, it enriches the fidelity of the concepts in the edited results, ensuring that the final output maintains both spatial and temporal coherence.

- We propose a novel **Concept-Augmented Textual Inversion (CATI)** approach that reliably captures target attributes from user-provided concept videos, improving the fidelity and flexibility of video editing.
- We introduce a **Dual Prior Supervision (DPS)** mechanism that stabilizes video generation by supervising cross-attention between source and target prompts. DPS prevents attention dispersion caused by target prompt modifications, significantly improving the consistency of non-target areas before and after editing.
- We orchestrate a framework that allows users to extract concepts from custom videos and generate diverse edited videos through concept templates. This approach supports plug-and-play integration with stable diffusion models, enabling efficient and stable video editing.

2 Related Work

Text-Driven Video Editing. Current approaches for text-driven video editing mainly fall into three categories: fine-tuning video generation models [Zhao *et al.*, 2023], [Wang *et al.*, 2024], fine-tuning image generation models extended with temporal modules [Wu *et al.*, 2023], [Qi *et al.*, 2023], and combining NLA [Kasten *et al.*, 2021] with pre-trained image generation models [Bar-Tal *et al.*, 2022], [Lee *et al.*, 2023], [Chai *et al.*, 2023]. Recent advances have demonstrated various innovative approaches in these categories. For example, [Ku *et al.*, 2024] employs a pretrained model for diverse video editing tasks, while GenVideo [Singer *et al.*, 2025] utilizes a target-image-aware approach with InvEdit masks to overcome text-prompt limitations. [Bar-Tal *et al.*, 2022], [Lee *et al.*, 2023], [Chai *et al.*, 2023] extract layered neural atlases from video to edit atlases which are further processed to synthesize videos; however, generating a neural atlas demands considerable computational time. Recently, Tune-A-Video [Wu *et al.*, 2023] achieves one-shot video editing with improved inter-frame coherency by updating self-attention with sparse causal attention. FateZero [Qi *et al.*, 2023] further proposes self-attention blending and incorporates attention control [Hertz *et al.*, 2023] to enhance the ability of editing objects, background, and styles while maintaining scene consistency. For temporal consistency specifically, VidToMe merges self-attention tokens across frames, while [Geyer *et al.*, 2023] leverages inter-frame correspondences to propagate features. In spatial editing, approaches like [Ceylan *et al.*, 2023], [Cohen *et al.*, 2024], [Liu *et al.*, 2024] improve results using spatial or temporal attention features in diffusion models. For editing targets with specific attributes, it becomes necessary to introduce external word embeddings. Our method supports the incorporation of external concept word embeddings. Furthermore, inspired by Tune-A-Video [Wu *et al.*, 2023] and FateZero [Qi *et al.*, 2023], we introduce a dual prior supervision mechanism between video latents and word embeddings to enhance scene consistency before and after video editing based on attention control methods. Compared to existing approaches, our method focuses on attention supervision and control mechanisms and operates on a one-shot video editing paradigm.

Textual Inversion. [Gal *et al.*, 2022] proposes a textual inversion method that optimizes newly added concept word embeddings in the CLIP [Radford *et al.*, 2021] text encoder, supervised by the latent variable distribution of specific images in the diffusion model. However, using a pre-trained diffusion model for self-supervised text inversion may lead to under-fitting for some specific images due to the finite latent space. Although it's feasible to optimize full parameters of the denosing network in diffusion model with a smaller learning rate simultaneously, or to train it with frozen concept embeddings in the next stage, this process faces issues of easy over-fitting and high storage costs. Our method, building upon textual inversion [Gal *et al.*, 2022], attempts to add LoRA [Hu *et al.*, 2022] to the denosing network, optimizing them simultaneously with concept words at a smaller learning rate, to enhance the text editing capabilities of concept words.

Cross Attention Control and Supervision. Prompt-to-Prompt [Hertz *et al.*, 2023] proposes three attention control methods for stable text-driven image editing based on diffusion models: word swap, refinement, and reweighting. By applying the cross-attention probability map recorded from the original image latent variables and text to the denoising process of original image latent variables and edited text, it has achieved significant success in stable text-driven image editing [Avrahami *et al.*, 2022], [Avrahami *et al.*, 2023]. Additionally, [Qi *et al.*, 2023] proposed self-attention blend effectively transfers the stability of text-driven image editing to video editing. Our method, built upon this foundation, introduces external concept words to support editing with higher degrees of freedom. Inspired by the work of [Yang and Tang, 2022], we introduce an attention supervision mechanism to address the issue of dispersed attention in editing words.

3 Method

3.1 Preliminaries

Textual Inversion. Textual inversion [Gal *et al.*, 2022] learns new embeddings for user-provided visual concepts in the textual embedding space, associating them with pseudo-words for use in new sentences for text-to-vision editing. The process uses a latent diffusion model, typically with a pretrained autoencoder and a noise prediction network: the encoder $\mathcal{E}$ maps image x to latent $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ reconstructs $x \approx \mathcal{D}(z)$. Textual inversion employs a CLIP [Radford *et al.*, 2021] text encoder c_θ with added concept words to encode conditional text y. The optimization

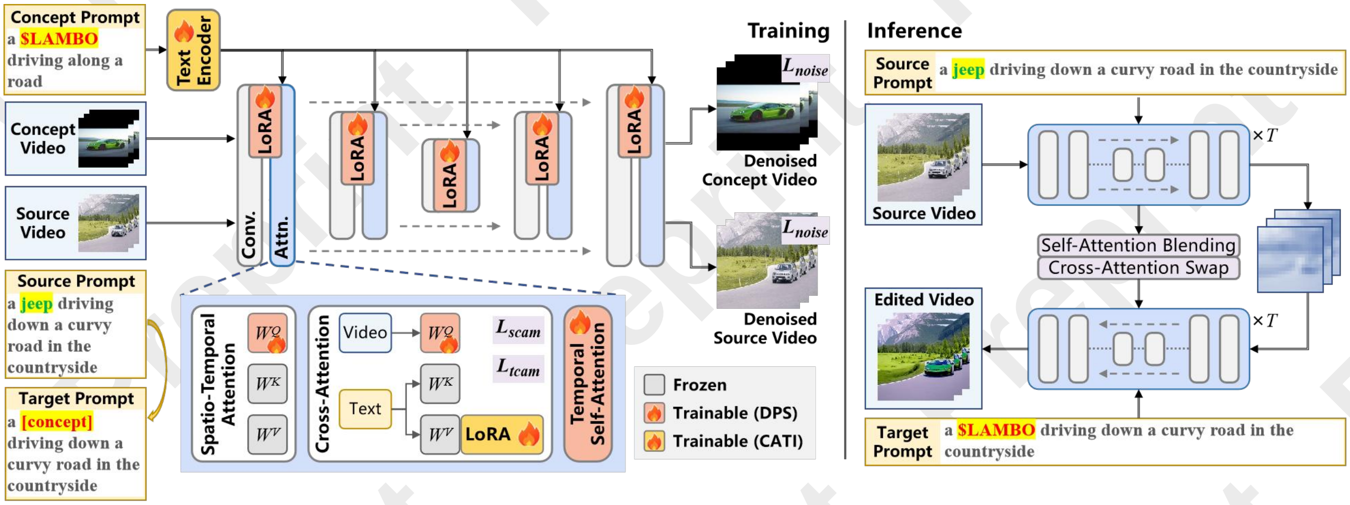

Figure 1: **Overview of our training and inference pipelines.** During the training stage, we first adapt the diffusion model to new visual concepts using our introduced Concept-Augmented Textual Inversion (CATI), and then we tune the temporally extended diffusion model with our proposed Dual Prior Supervision (DPS) mechanism to prevent unintended changes in edited videos. During the inference stage, we blend self-attention matrices (Self-Attention Blending) and swap cross-attention matrices (Cross-Attention Swap) to achieve stable video editing.

objective is:

$$\mathcal{L}_{noise} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2], \quad (1)$$

where z_t is the noised latent at time step t, ϵ is the noise, and ϵ_θ is the noise prediction network.

Low-Rank Adaption. [Hu *et al.*, 2022] proposes an efficient fine-tuning scheme based on matrix low-rank decomposition. For the pre-trained weight $\boldsymbol{W}_0 \in \mathbb{R}^{d\times k}$ in the original model, it updates the weight as $\boldsymbol{W} = \boldsymbol{W}_0 + \Delta\boldsymbol{W}$, where $\Delta\boldsymbol{W} = \boldsymbol{B}\boldsymbol{A}$, $\boldsymbol{B} \in \mathbb{R}^{d\times r}$, $\boldsymbol{A} \in \mathbb{R}^{r\times k}$, and $r \ll min(d, k)$. During the fine-tuning process, the pre-trained weight $\boldsymbol{W}_0$ is frozen, while $\boldsymbol{A}$ and $\boldsymbol{B}$ are trainable parameters. For the forward computation of the original weight $\boldsymbol{h} = \boldsymbol{W}_0\boldsymbol{x}$, the updated forward computation becomes:

$$LoRA(\boldsymbol{h}) = \boldsymbol{W}_0\boldsymbol{x} + \Delta\boldsymbol{W}\boldsymbol{x}. \quad (2)$$

Video Diffusion Models with Temporal Extensions. Tune-A-Video [Wu *et al.*, 2023] introduces Spatio-Temporal Attention (ST-Attn) to replace the original Self-Attention [Vaswani, 2017] in the 2D UNet. When calculating the keys $\boldsymbol{K}$ and values $\boldsymbol{V}$, ST-Attn concatenates latent variables of the first and former frames of the video, leading to the attention result where the current frame attends to both the first and former frames. The specific operations for replacing $\boldsymbol{K}, \boldsymbol{V}$ in Self-Attention are as follows:

$$\boldsymbol{K} = \boldsymbol{W}^K[\boldsymbol{z}_{v_1}; \boldsymbol{z}_{v_{i-1}}], \boldsymbol{V} = \boldsymbol{W}^V[\boldsymbol{z}_{v_1}; \boldsymbol{z}_{v_{i-1}}], \quad (3)$$

where W^K and W^V are projection matrices for key and value respectively, z_{v_i} denotes the latent variable of the i-th frame of the video to the current attention layer, and $[\cdot]$ denotes concatenation.

3.2 Stabilized Text-Driven Video Editing

Our training and inference pipelines are shown in Fig. 1. We use a UNet initialized from Stable Diffusion's pre-trained 2D UNet [Rombach *et al.*, 2021] as the noise predictor. To handle 3D video inputs, we replace spatial self-attention layers with ST-Attn (Eq.3). Following FateZero [Qi *et al.*, 2023], we add LoRA-based temporal convolution layers after spatial convolutions, and temporal self-attention with zero-initialized linear output after cross-attention. These new modules are residually connected to the originals.

Our approach for stabilized text-driven video editing has two learning phases. In the first phase, we introduce Concept-Augmented Textual Inversion (CATI) to adapt the diffusion model to new visual concepts. In the second phase, we tune partial parameters of the temporally extended diffusion model to suppress unintended changes in edited videos by calibrating cross-attention results.

Concept-Augmented Textual Inversion. Textual inversion [Gal *et al.*, 2022] learns to represent a specific set of user-provided images with pseudo-words in the latent space, offering an intuitive way for natural language-guided image editing. We incorporate this technique into our framework to facilitate video editing. However, due to the self-supervised nature within the limited latent space of the pre-trained diffusion model, the vanilla textual inversion often results in varied performance in terms of quality and efficiency for different image sets, requiring meticulous adjustments for learning rates and iteration counts.

To alleviate this issue, we draw inspiration from existing parameter-efficient fine-tuning techniques and propose adding LoRA modules [Hu *et al.*, 2022] to the value projection parameters in the cross-attention layers of the UNet. Consequently, the values $\boldsymbol{V}$ are updated to $LoRA(\boldsymbol{V})$ according to Eq. (2). The rationale behind our approach is that we aim to enhance the expressiveness of the pre-trained diffusion model by slightly adjusting its capacity to accommodate new visual concepts while preserving its original generation capability. Besides, inserting LoRA modules not only

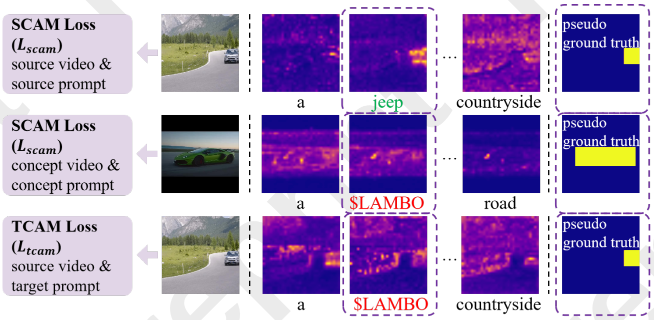
SCAM Loss ($\mathcal{L}_{scam}$) source video & source prompt
a
jeep
...
countryside
pseudo ground truth
SCAM Loss ($\mathcal{L}_{scam}$) concept video & concept prompt
a
$LAMBO
...
road
pseudo ground truth
TCAM Loss ($\mathcal{L}_{tcam}$) source video & target prompt
a
$LAMBO
...
countryside
pseudo ground truth

Figure 2: **Visualization of the dual prior supervision mechanism.** Each row displays a video frame, a set of cross-attention maps between this video frame and prompt words, and a pseudo ground truth mask. The *scam* loss and *tcam* loss are computed between relevant words and pseudo masks to reduce unintended changes.

augments textual inversion with low storage overhead but also maintains a plug-and-play characteristic during inference. Textual inversion process relies on the pre-trained denoising network established text-image attention probability distribution to achieve accurate target representation. In this context, fine-tuning only the $\boldsymbol{V}$ weights instead of $\boldsymbol{Q}$ and $\boldsymbol{K}$ allows new feature representations to be directly integrated while suppressing changes towards the pre-trained attention distribution to enable stable training during the early stages.

We train the concept-word embeddings of textual inversion and the weight parameters of LoRA modules in an end-to-end manner (see orange blocks in Fig. 1), where the learning rate for LoRA parameters is relatively smaller than that for concept-word embeddings to avoid over-fitting. Denote the noise prediction network with LoRA modules loaded on value projection parameters as ϵ_{θ_L}, the optimization objective of concept-augmented textual inversion is then updated from Eq. (1) to the following:

$$\mathcal{L}_{noise} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_{\theta_L}(z_t,t,c_\theta(y))\|_2^2]. \quad (4)$$

Model Tuning with Dual Prior Supervision. After learning concept-augmented textual inversion, we adapt and tune the video diffusion model for text-driven video editing in line with the paradigm of few-shot learning. Specifically, we learn the LoRA-structured temporal convolution layers, the query projection weights within spatio-temporal attention layers and cross-attention layers, and the temporal self-attention layers (see red blocks in Fig. 1). These parameters are selected for updates during training due to their strong relevance to the temporal modeling of 3D videos. To attain more stable and higher quality editing results, we tried directly integrating existing attention control techniques [Hertz *et al.*, 2023] in an early attempt; however, we found that when applying text-driven video editing types such as word swap, the dispersion phenomenon of cross-attention between text embeddings and video latents leads to reduced stability in editing results, which is shown in Fig 5. To address this challenge, we propose a dual prior supervision mechanism, which includes a source cross-attention mask (*scam*) loss and a target cross-attention mask (*tcam*) loss.

The *scam* loss is designed to reduce the attention influence of the words to be replaced in the source prompt on irrelevant frame areas (see the first row in Fig. 2). It is also applied to modulate attention between concept words and concept videos (see the second row in Fig. 2). Specifically, for K cross-attention layers in the UNet, we record cross-attention matrices $\mathbf{M}_s$ between the words and the video frame latents in each cross-attention layer. To obtain ground truth for optimization, we use an off-the-shelf object detection network OWL-ViT [Minderer *et al.*, 2022] to localize objects in video frames and generate corresponding binary pseudo-labels $\mathbf{M}_s^{\text{gt}}$. We further apply max pooling to generate K pseudo-labels, each with a designated resolution P_k. The loss is then calculated as the mean absolute loss on irrelevant areas:

$$\mathcal{L}_{scam} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{P_k}\left[\|\mathbf{M}^{\text{gt}}_{s,k,i} - \mathbf{M}_{s,k,i}\| \cdot (1 - \mathbf{M}^{\text{gt}}_{s,k,i})\right]. \quad (5)$$

The *tcam* loss is introduced to diminish the attention influence of the target words in the edited prompt to further promote the consistency of irrelevant areas before and after video editing (see the third row in Fig. 2). Similar to the *scam* loss, we obtain cross-attention matrices $\mathbf{M}_t$ and pseudo-labels $\mathbf{M}_t^{\text{gt}}$ between the target words in the edited prompt and the video frame latents. The loss is computed as:

$$\mathcal{L}_{tcam} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{P_k}\left[\|\mathbf{M}^{\text{gt}}_{t,k,i} - \mathbf{M}_{t,k,i}\| \cdot (1 - \mathbf{M}^{\text{gt}}_{t,k,i})\right]. \quad (6)$$

Let the trainable parameters during the model tuning phase be denoted as ϵ_{θ_T}. The noise prediction loss $\mathcal{L}_{noise}$ is then obtained by substituting ϵ_θ in Eq. (1) with ϵ_{θ_T}. Given α and β as the weighting coefficients for our proposed *scam* loss and *tcam* loss, respectively, the total loss for model tuning with dual prior supervision is formulated as:

$$\mathcal{L} = \mathcal{L}_{noise} + \alpha\mathcal{L}_{scam} + \beta\mathcal{L}_{tcam}. \quad (7)$$

Inference. As shown in Fig.1, the pipeline consists of an inversion stage with the source prompt and an editing stage with the modified prompt. During inversion, we cache self- and cross-attention matrices at each step, which are later used to control attention in editing. Specifically, we blend self-attention matrices to preserve semantic layout [Qi *et al.*, 2023], and swap cross-attention matrices for changed words and video latents [Hertz *et al.*, 2023].

4 Experiments

4.1 Settings and Datasets

Our experiments are conducted on a machine equipped with an NVIDIA GeForce RTX 4090. During the concept augmented textual inversion stage, we set the learning rate for CLIP [Radford *et al.*, 2021] word embeddings to 1×10^{-3}, and the learning rate for LoRA modules inserted into the UNet to 1×10^{-5}, with the number of training steps set to Additionally, we randomly sample frame numbers within the range $[4, 8]$ from the concept video during training, to prevent the inversion process from over-fitting to a fixed frame number. For the video diffusion model fine-tuning stage, we empirically set $\alpha = 0.1$ and $\beta = 0.1$ in Eq. (7). The

Methods	M-PSNR ↑	Concept Cons. ↑	Frame Cons. ↑
Tune-A-Video	14.70	0.6982	0.9399
FateZero	17.08	0.6822	0.9413
MotionDirector	12.73	0.7222	0.9452
RAVE	17.39	0.6990	0.9379
Ours	**19.71**	**0.7642**	**0.9472**

Table 1: Quantitative results of video editing ***w/*** concept video.

Methods	M-PSNR ↑	Frame Cons. ↑
Tune-A-Video	15.72	0.9397
FateZero	19.42	0.9246
MotionDirector	16.86	0.9403
RAVE	16.20	0.9306
Ours	**22.10**	**0.9405**

Table 2: Quantitative results of video editing ***w/o*** concept video.

training steps above all use AdamW [Loshchilov and Hutter, 2017] optimizer. In the inference stage of video editing, the guidance scale is set to 12.5, the number of DDIM Inversion steps is $T = 50$, and the self-attention blending and cross-attention swap steps are within the interval $[0, 0.7T]$. To evaluate our proposed method, we used a portion of the DAVIS [Caelles *et al.*, 2019] dataset and clip videos from the internet to construct video editing pairs, either with or without concept videos.

4.2 Metrics

Frame Consistency. To compare the coherence of the video frames $\mathbb{F}$, we refer to the metric used in [Wu *et al.*, 2023], [Hessel *et al.*, 2021], which calculates the average cosine distance d between features $(\boldsymbol{v}_i, \boldsymbol{v}_j)$ of each two different frames $(\boldsymbol{f}_i, \boldsymbol{f}_j)$ encoded by the CLIP visual encoder [Radford *et al.*, 2021], as Eq. (8). Here, $\boldsymbol{f}_i, \boldsymbol{f}_j \in \mathbb{F}$, $\boldsymbol{f}_i \neq \boldsymbol{f}_j$, and $\mathbb{D}$ denotes the set of the vector pairs $(\boldsymbol{v}_i, \boldsymbol{v}_j)$.

$$d = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{v}_i, \boldsymbol{v}_j) \in \mathbb{D}} \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_i\| \|\boldsymbol{v}_j\|}. \tag{8}$$

Masked Peek-Signal-Noise Ratio. To compare the stability of the video non-target areas before and after target editing, we design a Masked Peak Signal-to-Noise Ratio (**M-PSNR**) metric. We use the OWL-ViT [Minderer *et al.*, 2022] open-vocabulary object detection model with text pseudo-labels to estimate the bounding box mask $\boldsymbol{M}$ of the edited target. We then compare the average peek-signal-noise ratio of the original video frames and the edited video frames after applying this mask. The calculation formula for the specific function f for the Mean Squared Error (MSE) used as input is as follows, where $\boldsymbol{M} \in \mathbb{R}^{H \times W}$, $\boldsymbol{I}^s \in \mathbb{R}^{H \times W \times C}$, and $\boldsymbol{I}^e \in \mathbb{R}^{H \times W \times C}$ refer to the mask value, the frame pixel value of video before and after editing, respectively.

$$f(\boldsymbol{I}^s, \boldsymbol{I}^e, \boldsymbol{M}) = \frac{1}{C} \frac{\sum_{k \in C} \sum_{i \in H} \sum_{j \in W} (I^s_{i,j,k} - I^e_{i,j,k})^2 (1 - M_{i,j})}{\sum_{i \in H} \sum_{j \in W} (1 - M_{i,j})}. \tag{9}$$

Concept Consistency. We employ a multi-step approach to evaluate the correlation between the video editing results guided by the concept video and the concept video itself while minimizing interference in non-target areas. First, we use a pre-trained OWL-ViT [Minderer *et al.*, 2022] model in conjunction with pseudo-label prediction to generate object masks for both videos. We then extract pixel segments of the target objects from both the edited video and the concept video. Finally, we leverage the CLIP model to predict visual encoding vectors for these extracted segments and calculate the average cosine similarity between them.

4.3 Comparisons with Existing Methods

Quantitative Evaluation. As illustrated in Tab. 1 and Tab. 2, we assess text-driven video editing results in three aspects. Compared with existing methods that extend and fine-tune the Stable Diffusion model, including Tune-A-Video [Wu *et al.*, 2023], FateZero [Qi *et al.*, 2023], RAVE [Kara *et al.*, 2024], and MotionDirector [Zhao *et al.*, 2023], our approach demonstrates superior inter-frame coherence in terms of the Frame Consistency Metric. To evaluate the consistency of unrelated areas before and after video editing, we employ M-PSNR as a reference metric, and our method achieves the highest score by a large margin. Concretely, our method outperforms MotionDirector [Zhao *et al.*, 2023] by a noticeable 6.98 M-PSNR in editing with concept video. This is attributed to our proposed prior supervision mechanism, which effectively reduces the editing noise in non-target areas for both source and concept videos. Furthermore, to evaluate the target fidelity in concept and edited videos, we utilize Concept Consistency as a reference metric, and our method demonstrates greater fidelity compared to others.

Qualitative Evaluation. Fig.3 shows visual comparison results of video editing with and without concept video guidance. Our method maintains content consistency in non-target areas before and after editing. With concept videos, it effectively introduces visual concepts from the concept video into the edited result. For instance, in Fig.3 (Setting I), our method successfully replaces man with '$OPTIMUS', while others fail to preserve background or transfer the complete target shape. Other approaches commonly face instability in non-target areas. Tune-A-Video [Wu *et al.*, 2023] encounters dispersed cross-attention issues due to fine-tuning with only one video-text pair. While FateZero [Qi *et al.*, 2023] and RAVE [Kara *et al.*, 2024] mitigate this through cross-attention manipulation or noise shuffling, their direct concept-driven editing compromises non-target consistency and concept fidelity. MotionDirector [Zhao *et al.*, 2023] extracts targets via its trainable spatial path, but coupled spatial-temporal paths provide unstable guidance, causing non-target inconsistencies. Our CATI and DPS effectively maintain non-target content consistency while accurately capturing user-provided concept attributes.

4.4 Ablation Study

Concept Augmentation Alleviates Under-Fitting of Textual Inversion. In this work, we draw on the idea of Textual Inversion (TI) from text-to-image generation and apply

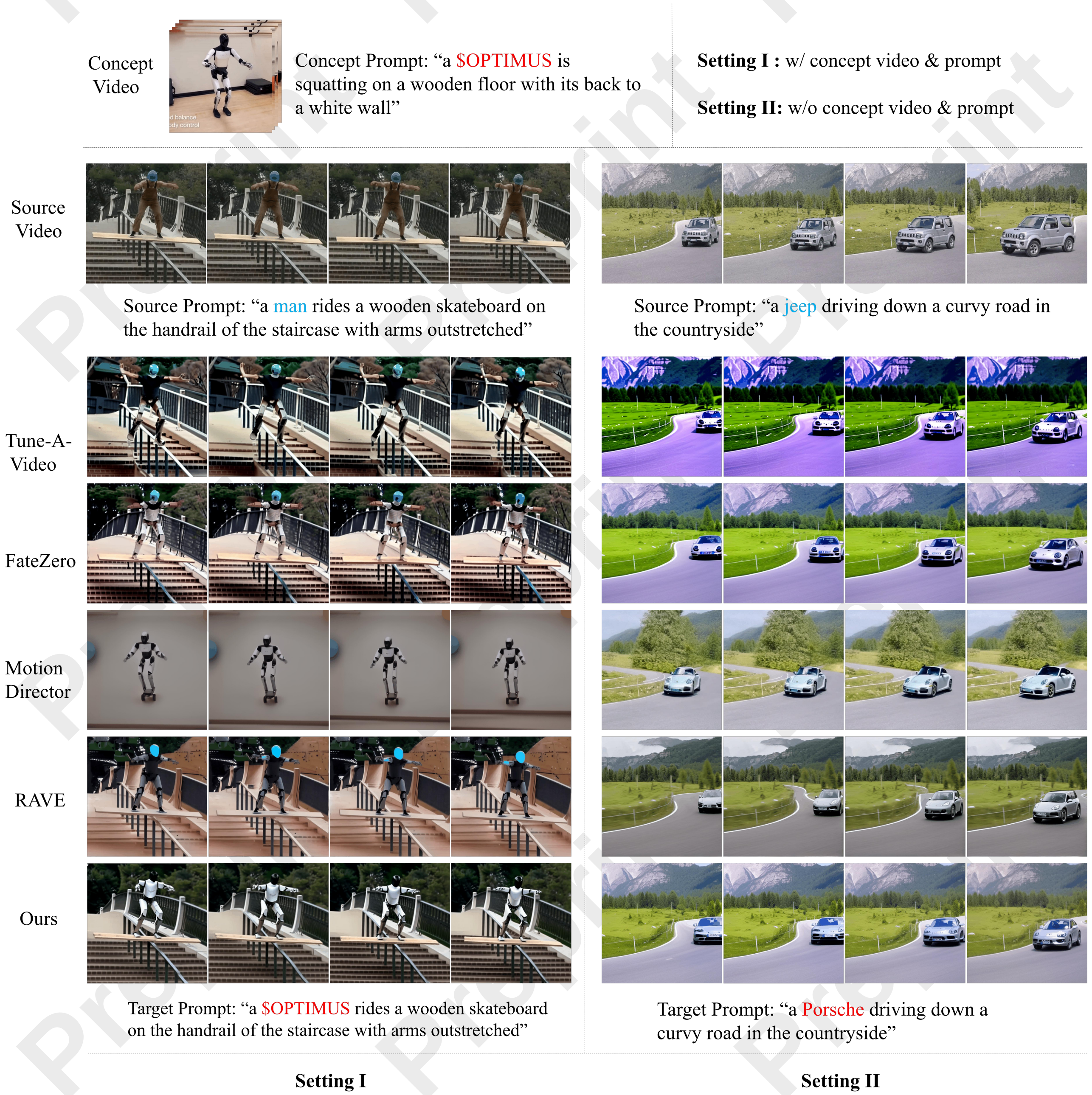
Concept Video
Concept Prompt: "a $OPTIMUS is squatting on a wooden floor with its back to a white wall"
Setting I : w/ concept video & prompt
Setting II: w/o concept video & prompt
Source Video
Source Prompt: "a man rides a wooden skateboard on the handrail of the staircase with arms outstretched"
Source Prompt: "a jeep driving down a curvy road in the countryside"
Tune-A-Video
FateZero
Motion Director
RAVE
Ours
Target Prompt: "a $OPTIMUS rides a wooden skateboard on the handrail of the staircase with arms outstretched"
Target Prompt: "a Porsche driving down a curvy road in the countryside"
Setting I
Setting II

Figure 3: **Video generation with (Setting I) and without (Setting II) concept pairs.** The top row of the figure contains the concept video with its prompt. The second row is the source video frames coupled with prompts that need to be edited. The rows below show the editing results of the source video using the editing prompt for [Wu *et al.*, 2023], [Qi *et al.*, 2023], [Zhao *et al.*, 2023], [Kara *et al.*, 2024] and our method, respectively, in which words with "$" ahead mean concept words, and the same for subsequent results.

it to text-driven video editing to address the embedding of external concept words. However, simply applying TI may lead to under-fitting, resulting in a lack of realism. For instance, in the results shown in Fig. 4(a) and Fig. 4(c), where the keywords 'jeep' are altered to '$LAMBO' and '$CYBERTRUCK', although some attributes (e.g., shape) of the target concepts are partially retained, the results appear to "drift" due to insufficient inductive bias. In contrast, the concept-augmented textual inversion (CATI) can effectively capture the color, shape, and other attributes, as demonstrated in Fig. 4(b) and Fig. 4(d). CATI provides more detailed features for editing, significantly improving inversion fidelity.

Dual Prior Supervision Improves Stability and Fidelity. In this work, we propose a Dual Prior Supervision (DPS) strategy, which consists of two main components (See Sec. 3.2): *scam* loss and *tcam* loss. Both components play crucial roles in maintaining the stability of the target gener-

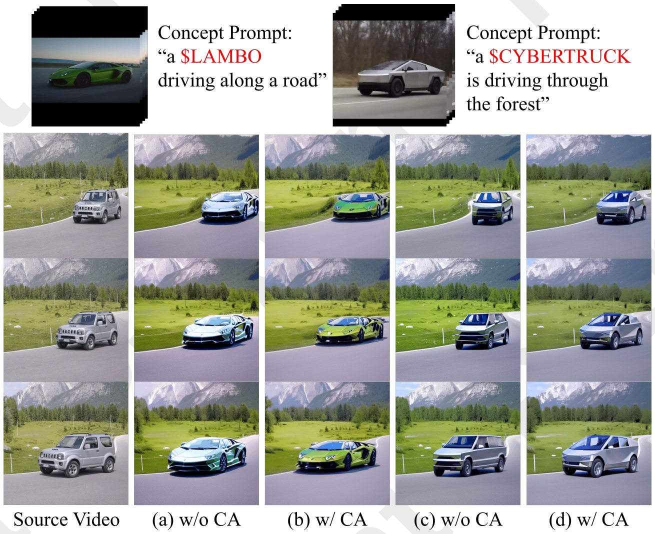

Figure 4: **Comparison of whether to use Concept Augmentation (CA) for textual inversion.** Compared the text inversion results without and with concept augmentation for pairs (a), (b): ‘jeep’ → ‘$LAMBO’; and (c), (d): ‘jeep’ → ‘$CYBERTRUCK’, respectively, from the same source prompt “a jeep driving down a curvy road in the countryside”.

ation. By comparing the attention regions in Fig. 5 (a) (w/o *tcam*, w/o *scam*), Fig. 5 (b) (w/o *tcam*, w/ *scam*), and Fig. 5 (c) (w/ *tcam*, w/o *scam*), we can conclude that both *scam* and *tcam* (Fig. 5 (d)) significantly reduce background disturbances and improve stability. However, the generated video results reveal that using either component alone cannot effectively capture attributes of the target object, such as the color of the car. DPS combines both components, not only enhancing the stability of the background in the target results but also capturing the target object’s attributes more accurately, thereby improving the fidelity of the edited concept target.

Tuning w/ Concept Video Produces Stylized Results. Recall that we construct the target videos in this work by templating the concept pairs to make the editing process more flexible. To explore the impact of the concept video in **Setting I** (Fig. 3), we conduct a simple experiment as shown in Fig. 6. As shown in Fig. 6(a) and Fig. 6(b), tuning models with both concept video and concept prompt produces more stylized videos. The possible explanation lies in that the concept video alleviates the overfitting issue.

5 Limitations and Future Work

Mismatch when Significant Deformation. Our method effectively mitigates the inconsistency in non-target areas caused by attention dispersion in video editing methods using attention replacement mechanisms, it may struggle when a single concept video guides target replacement in cases of significant deformation in the source video, such as running people. For instance, there may be insufficient detailed correspondences between the internal parts of the replacing and replaced targets during deformation, such as moving arms and legs. Potential solutions include ControlNet [Zhang *et al.*, 2023], OpenPose [Cao *et al.*, 2019] and Sign-D2C [Tang *et al.*, 2025] which utilize motion conditions, like human pose or sign language, to guide the video editing process.

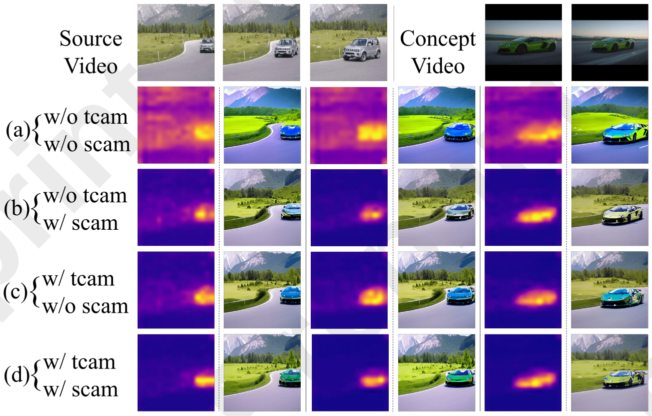

Figure 5: **The impact of dual prior supervision.** From the first to the last row, using the editing example in Fig. 1, we compare the average cross-attention maps and the editing results with and without the supervision mechanism of *scam* and *tcam*. Each case contains three pairs, and each pair consists of an average cross-attention map on the left and an edited frame on the right.

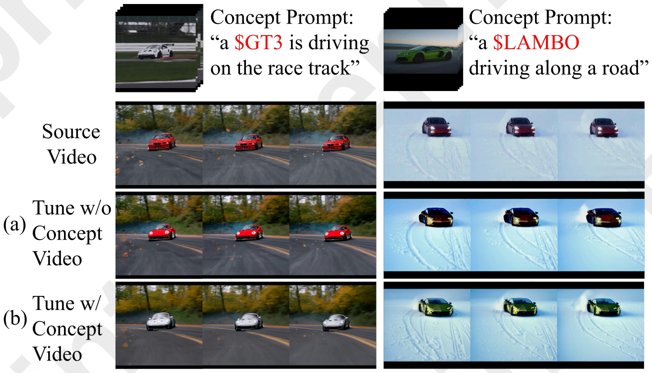

Figure 6: **Comparison of whether to tune with the concept video.** Compared the video editing results without and with tuning concept video for the left part: ‘car’ → ‘$GT3’; and the right part: ‘car’ → ‘$LAMBO’, from the source prompt “a car is drifting around a curve road with the background of a forest” and “a car is drifting in the snow”, respectively.

6 Conclusion

In this paper, we present an improved concept-augmented video editing approach that flexibly produces diverse, stable target videos by leveraging abstract conceptual pairs. Specifically, we introduce Concept-Augmented Textual Inversion (CATI) to capture user-defined target concepts, enabling a plug-and-play, stable diffusion pipeline for more stylized editing. We further propose a Dual Prior Supervision (DPS) mechanism to align cross-attention between source and target prompts, preventing unintended changes in non-target regions. Experimental results show that our method significantly enhances flexibility, consistency, and stability in text-driven video editing.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62472139, by the Anhui Provincial Natural Science Foundation, China (Grant No. 2408085QF191), the Fundamental Research Funds for the Central Universities (Grants No. JZ2024HGTA0178, JZ2023HGQA0097), by the Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2403), Zhejiang University.

References

[Avrahami *et al.*, 2022] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, June 2022.

[Avrahami *et al.*, 2023] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), jul 2023.

[Bar-Tal *et al.*, 2022] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022.

[Blattmann *et al.*, 2023] 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.

[Caelles *et al.*, 2019] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019.

[Cao *et al.*, 2019] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[Ceylan *et al.*, 2023] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023.

[Chai *et al.*, 2023] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *arXiv preprint arXiv:2308.09592*, 2023.

[Cohen *et al.*, 2024] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. *arXiv preprint arXiv:2405.12211*, 2024.

[Gal *et al.*, 2022] 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, 2022.

[Geyer *et al.*, 2023] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

[Hertz *et al.*, 2023] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

[Hessel *et al.*, 2021] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

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

[Ho *et al.*, 2022] 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.

[Hong *et al.*, 2022] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[Hu *et al.*, 2022] 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. In *International Conference on Learning Representations*, 2022.

[Kara *et al.*, 2024] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M. Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Kasten *et al.*, 2021] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021.

[Ku *et al.*, 2024] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

[Lee *et al.*, 2023] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14317–14326, 2023.

[Liu *et al.*, 2024] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Minderer *et al.*, 2022] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022.

[Qi *et al.*, 2023] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535*, 2023.

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

[Rombach *et al.*, 2021] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[Singer *et al.*, 2025] Uriel Singer, Amit Zohar, Yuval Kirstain, Shelly Sheynin, Adam Polyak, Devi Parikh, and Yaniv Taigman. Video editing via factorized diffusion distillation. In *European Conference on Computer Vision*, pages 450–466. Springer, 2025.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Tang *et al.*, 2025] Shengeng Tang, Jiayi He, Lechao Cheng, Jingjing Wu, Dan Guo, and Richang Hong. Discrete to continuous: Generating smooth transition poses from sign language observation. In *CVPR*, 2025.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

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

[Wu *et al.*, 2023] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023.

[Yang and Tang, 2022] Gene-Ping Yang and Hao Tang. Supervised attention in sequence-to-sequence models for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7222–7226. IEEE, 2022.

[Zhang *et al.*, 2023] 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.

[Zhao *et al.*, 2023] 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.