EVA: Zero-shot Accurate Attributes and Multi-Object Video Editing

Xiangpeng Yang[1], Linchao Zhu[2], Hehe Fan[2], and Yi Yang[2]

[1] ReLER Lab, University of Technology Sydney
[2] Zhejiang University

https://knightyxp.github.io/EVA/

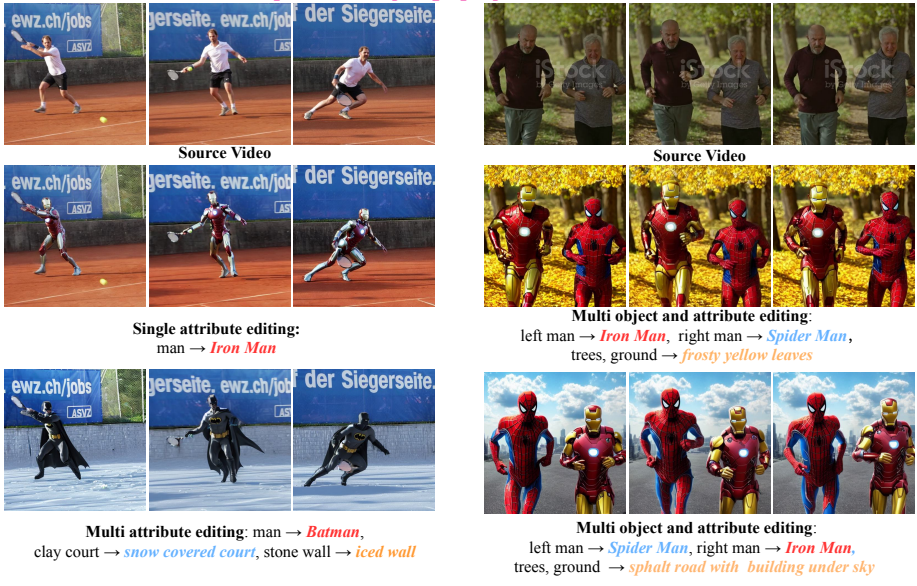
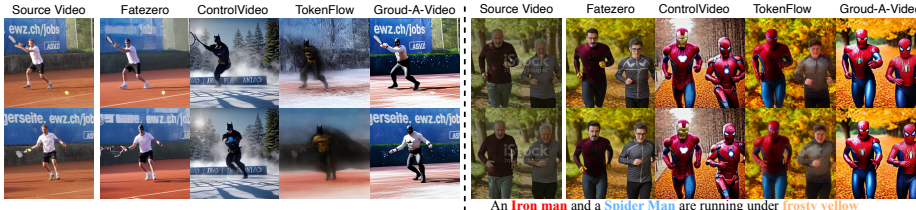


Fig. 1: EVA achieves multi-attribute editing for both single and multi-object scenarios, adhering to the source video's layout and faithfully preserving motion information.

Abstract. Current diffusion-based video editing primarily focuses on local editing (*e.g.,* object/background editing) or global style editing by utilizing various dense correspondences. However, these methods often fail to accurately edit the foreground and background simultaneously while preserving the original layout. We find that the crux of the issue stems from the imprecise distribution of attention weights across designated regions, including inaccurate text-to-attribute control and attention leakage. To tackle this issue, we introduce EVA, a **zero-shot** and **multi-attribute** video editing framework tailored for human-centric videos with complex motions. We incorporate a Spatial-Temporal Layout-Guided Attention mechanism that leverages the intrinsic positive and negative correspondences of cross-frame diffusion features. To avoid attention leakage, we utilize these correspondences to boost the attention scores of tokens within the same attribute across all video frames while limiting interactions between tokens of different attributes in the self-attention layer. For precise text-to-attribute manipulation, we use discrete text embeddings focused on specific layout areas within the cross-attention layer. Benefiting from the precise attention weight distribution, EVA can be easily generalized to multi-object editing scenarios and achieves accurate identity mapping. Extensive experiments demonstrate EVA achieves state-of-the-art results in real-world scenarios. Full results are provided at project page.

1 Introduction



(a) Edit results of previous methods in single-object multi-attribute scene **(b) Edit results of previous methods in multi-object multi-attribute scene**

Fig. 2: Previous methods failed results are displayed in single/multi-object scenes. EVA's successful edit result is shown in the third row of Fig 1 left and the second row of Fig 1 right.

Text-to-video (T2V) editing, which aims to change the visual appearance of a video according to a given prompt, is an emerging field that harnesses strong generation ability from text-to-image/video models [29–32]. Previous works have employed dense correspondences, such as depth/edge maps [41, 42], optical flow [8, 14, 38] and attention maps [27], for local attribute or global style editing, often compromising fidelity.

In this paper, we focus on multi-attribute editing because it enables us to finely adjust local attributes while maintaining the original video's layout and background intact, resulting in more authentic edits. Previous works have encountered many challenges in multi-attribute editing (Fig 2). The main issues include: (1) overlooking or distorting edits of individual attributes, with FateZero [27] unable to edit the object and ControlVideo [41] failing to preserve the background unchanged (Fig 2 (a)); (2) the mixing of different attributes, where ControlVideo leads to the blending of textures between "Iron Man" and "Spider-Man", and TokenFlow [10] incorrectly associates the identities of the two characters (Fig 2 (b)).

Ground-A-Video [15] is a recent approach to multi-attribute editing, which employs a cross-frame gated attention mechanism with word-to-bounding box control. Yet, it still has the aforementioned limitations. The word-to-bounding box control lacks the necessary precision, leading to the loss of fine-grained details, such as the racket illustrated in Fig 2 (a). Furthermore, when bounding boxes are overlapping, it leads to the mixing of textures in adjacent areas.

We identify the imprecise distribution of attention weights as the root cause of these challenges, including inaccurate text-to-attribute control and attention leakage. To ensure precise attention weight distribution, we introduce a Spatial-Temporal Layout-Guided Attention (ST-Layout Attn) mechanism.

First, to achieve accurate text-to-attribute control, we extract the corresponding text embedding for each attribute from the global prompt. These discrete text embeddings are applied to each attribute's corresponding layout areas within the cross-attention layer. For each layout area, we use masks as spatially disentangled information, leveraging the inherent characteristic that masks do not overlap. To further keep the fine-grained details, we perform latent blend [1] to preserve the undesired edit areas.

Second, to avoid attention leakage, our goal is to ensure the mutual exclusivity of different attributes while enhancing the correlation of the same attribute across frames. We leverage the cross-frame diffusion feature similarity (DIFT [33]), which reveals the intrinsic correlations across inter/intra attributes along a spatial-temporal axis. As

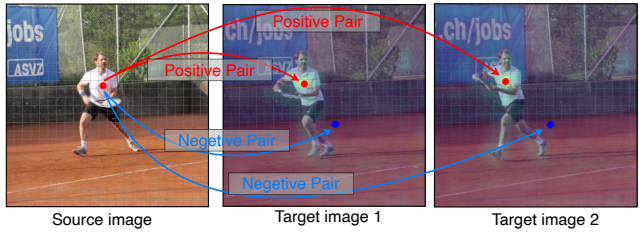


Fig. 3: Intrinsic Cross-frame DIFT [33] feature correspondence. We randomly select a "red point" in the source image, extract its DIFT feature, and compute cosine similarity with the target image. The target's "red point" marks the highest similarity, and "blue point" is the lowest, showing the potential to unsupervised identify intra/inter attributes correspondence.

demonstrated in Fig 3, for each token, we identify its corresponding positive pair in other frames (sharing the same attribute) by maximizing the cosine DIFT similarity. Similarly, we determine its negative pair (across different attributes) by minimizing this similarity. Thus, leveraging this intrinsic correspondence, we assign positive and negative values to each token across various layouts on a spatial-temporal axis. Consequently, we enhance the attention scores for tokens within the same attribute and limit interactions between tokens of different attributes across frames, thus significantly mitigating attention leakage.

Benefiting from precise attention-weight distribution and text-to-attribute control brought by our attention mechanism, we realize accurate identity mapping and background editing in the multi-object scenario. Such as swapping identity while editing the background (Fig 1 right).

Our key contributions can be summarized as follows:

- We propose EVA, a general framework for accurate attributes and multi-object video editing, which realizes accurate weight distribution and identity mapping.
- Leveraging intrinsic cross-frame DIFT correspondence, we introduce ST-Layout Attn for accurate text-to-attribute control and to avoid attention leakage.
- Without tuning any parameters, we achieve state-of-the-art results on existing benchmarks and real-world videos both qualitatively and quantitatively.

2 Related Work

2.1 Text-to-Image Editing/Generation

In the realm of single attribute text-to-image editing, various approaches have been explored, from manipulating attention maps in Pix2Pix-Zero [23] and Prompt2Prompt [13] to employing masks in DiffEdit [9] and Latent Blend [1, 2] for foreground modifications while preserving the background.

For multi-attribute editing, efforts such as Attention and Excite [6] and DPL [36] focus on maximizing the attention scores for each subject token and reducing attention leakage. Recently, in single image generation, [17] adjusted modulate attention value according to layout masks and dense caption. [25] proposed attention refocus loss for regularization. However, guiding the attention map solely with single-frame layout masks and dense captioning is inadequate in video editing, as it fails to maintain the original video's integrity and temporal consistency.

2.2 Text-to-Video Editing

Video Editing based on Image Diffusion Models Tune-A-Video (TAV) [37] is the first work to extend latent diffusion models to the spatial-temporal domain and encode the source motion implicitly by one-shot tuning but still fails to preserve local details. Fatezero [27] is a prompt2prompt [13] based editing method, fusing self- and cross-attention maps for temporal consistency. However, it requires extensive RAM usage and suffers from layout preservation even when equipping TAV for local object editing. [5] and [22], following the Neural Atlas [16] or dynamic Nerf's deformation field [21, 26], struggle with non-grid human motion. Subsequent methods like Rerender-A-Video [38], Flatten [8] ControlVideo [41, 42] achieve strict temporal consistency via optical-flow, depth/edge maps, but failed in multi-attribute editing while preserving original layouts. Tokenflow [10] enforces a linear mix of nearest key-frame features to ensure consistency but results in detail loss. Ground-A-VIDEO [15] leverages groundings for multi-attribute editing, but it suffers from attention leakage when bounding boxes overlap, even with dense guidance such as optical flow.
Video Editing based on Video Diffusion Models Previous video editing work primarily utilized text-to-image SD model [30]. Recent advancements in video foundation models [3, 12, 35, 39] have led efforts like MotionDirector [43] and VideoSwap [11] to employ temporal priors for customized motion transfer. Yet, current video foundation models are limited to fixed views and struggle with complex human motions. Additionally, these editing methods require tuning parameters, which poses a challenge for real-time video editing applications. In contrast, our EVA method requires no parameter tuning, enabling zero-shot, multi-object and multi-attribute video editing.

3 EVA

In this section, we start by analyzing the prerequisite for multi-attribute editing and the necessity for precise attention weight distribution in section 3.1. Subsequently, we present an overview of the proposed EVA pipeline in section 3.2. Following that, we detail the Spatial-Temporal Layout Attention mechanism in 3.3. Notice that, our EVA is a general zero-shot framework for editing both single and multiple objects, as well as backgrounds, in human-centric videos with complex motion.

3.1 What is the Key to Multi-Attribute Video Editing?

Accurate Text-to-Attribute Control To accurately edit multiple attributes in a video, it is essential to ensure the model's capability for precise editing of each individual attribute. The previous method, FateZero [27], addresses this challenge by implementing word-swap in prompt-to-prompt editing [13], making the editing of multiple attributes simultaneously. It also merges attention maps from the inversion process to preserve the original motion and layout information. However, as depicted in Fig 4 (a), FateZero struggles to edit "man" to "Iron Man" and "clay court" to "snow covered court" in video with complex motion, even with pose guidance from ControlNet [40]. This issue stems from the misalignment of attribute weights with their appropriate spatial regions. Specifically, the weight of "man" is concentrated only on the head, and the weight of "snow" is wrongly assigned to "man." Additionally, "court"'s weight doesn't correctly cover its area, spreading around the person instead.

Edit prompt: An Iron Man is playing tennis on snow covered court

(a) Inaccurate text-to-attribute control

Edit prompt: An Batman is playing tennis on snow covered court before an iced wall

(b) Attention leakage

Fig. 4: *Left*: FateZero [27] fails in text-to-attribute control, incorrectly allocating weights to "snow" and not fully covering "man." *Right*: Although Ground-A-Video [15] attempts to ground each attribute individually, it still suffers from attention leakage, leading to texture blending on "Batman's" upper body and imprecise edits of "court" and "wall."

This challenge inspires us to employ spatially disentangled semantic masks as layout information, leveraging their natural ability to capture each attribute's shape features. Moreover, we aim to ensure that each text embedding exclusively focuses on its corresponding attribute's spatial region, thus ensuring accurate text-to-attribute control.

Avoiding Attention Leakage However, relying solely on spatially disentangled information is insufficient. Previous work Ground-A-Video [15] introduced gated attention from [19] for text-to-bounding box control in multi-attribute editing. Yet, in complex motion, [15] still faces challenges in accurately editing the man, ground, and walls.

As illustrated in Fig 4 (b), our visualization of the cross-attention map reveals a phenomenon termed "attention leakage," where weights from other attributes, such as "court" and "wall," leak onto "Batman," causing his upper body to appear white. To address this, we introduce negative example awareness among different attributes to ensure the mutual exclusivity of each attribute weight, thereby avoiding attention leakage. Additionally, the incomplete weight distribution across Batman's body underscores the necessity to strengthen the correlation within each attribute. Cross-frame DIFT features inherently exist this kind of correspondence. For a query token, cross-frame DIFT feature similarity identifies positive pairs within the same attribute across frames and negative pairs in different attributes by calculating max/min similarity, as shown in Fig 3. Leveraging this correspondence, we introduce ST-Layout Attn mechanism. Our ST-Layout Attn not only ensures that each text embedding concentrates on its respective attribute, but also enhances the internal coherence within attributes and keeps the exclusivity of attention weights among different attributes. Through this approach, we effectively achieve accurate text-attribute control and prevent attention leakage.

3.2 Overall Framework

Our framework aims to edit the source video $V^{1:N}$ according to a textual prompt Δ_τ which contains a series of desired local attribute edits $\{\tau_1 \rightarrow \tau_{1'}, \tau_2 \rightarrow \tau_{2'}, \cdots\}$. Following previous work [27,37], we inflated the original StableDiffusion [30] (SD) along the temporal axis to adapt for 3-dimension video input.

For human-centric complex motion, we want to decouple human motion from object identity. Thus, we directly utilize the human pose as sparse motion information from the source video object. Following [41,42], we employ ControlNet's [40] pose guidance to promote temporal consistency.

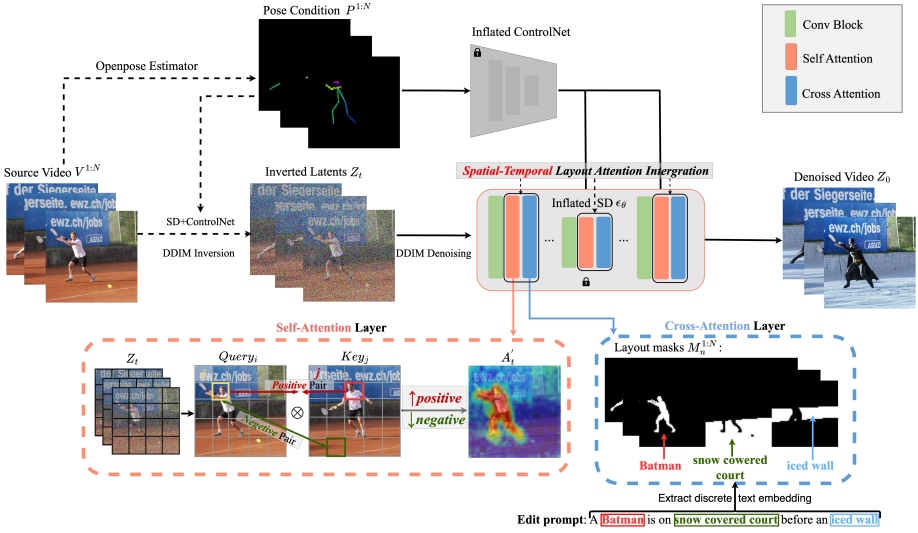


Fig. 5: EVA pipeline. We integrate the ST-Layout Attn within the frozen SD in the denoising process. In the self-attention layer, we compute the positive/negative value of each query token in different attributes from a spatial-temporal perspective, This allows us to augment the attention scores for tokens within the same attribute and reduce them for tokens in different attributes. In the cross-attention layer, we extract each attribute's text embeddings from the edit prompt, ensuring they focus only on corresponding layouts across frames.

Fig 5 illustrates the overall pipeline of our EVA:

(1) Firstly, we obtain layout masks $M_n^{1:N}$ (n denotes the layout or attribute classes) corresponding to each attribute through user-interactive Segment-and-Tracking anything [7], which provides crucial layout information. We obtain pose condition $P^{1:N}$ through the OpenPose estimator [4] to encode source complex motion information. Furthermore, we extract text embeddings for each attribute from the edit prompt, setting the stage for their subsequent use in the cross-attention layer of ST-Layout Attn.

(2) Then, the input videos undergo DDIM inversion in the latent space of Stable Diffusion (SD [30]) and ControlNet [40] to enhance the fidelity of the generated video.

(3) Finally, the inverted Latents Z_t are fed into inflated SD and ControlNet during the DDIM denoising process. In the denoising process, we incorporate ST-Layout Attn to ensure accurate attention weight distribution in a zero-shot manner. In the self-attention layer, based on the cross-frame DIFT similarity, we boost the attention scores of tokens in the same attribute and restrict communication between different attributes across all frames, avoiding attention leakage. In the cross-attention layer, we utilize each attribute's text embedding to enable direct text-to-attribute control.

3.3 Spatial-Temporal Layout-Guided Attention

DenseDiffusion [17] proposed modulating intermediate attention maps according to layout mask guidance for single image generation. However, in the context of dynamic video scenes, the correlation between attributes across different frames con-

stantly changes. Therefore, enhancing intra-attribute similarity while reducing inter-attribute interaction becomes crucial in scenes with complex motion.

Notations We define a set of layout masks $m_L^i = [m_1^i, m_2^i, \cdots, m_l^i]$ for the i^{th} frame and multi attributes $\tau_L = [\tau_1, \tau_2, \cdots, \tau_l]$,where L denotes total classes of layout attributes, and each pair (τ_l, m_l) correspond to a single region.

Identify the correlations of intra/inter attributes Consider the original SD is trained on the large-scale images, and lacks a built-in temporal module in the pretraining process. To incorporate temporal information effectively, we treat the full video frames as "a larger picture". Specifically, for each query Q at frame i, the key K or value V is computed from the concatenated latents across all frames, this process can be formulated as:

$$Q = W^Q z_t^i, \quad K = W^K z_t^N, \quad V = W^V z_t^N, \tag{1}$$

where W^Q,W^K,W^V project z_t into query, key and value. $z_t^N = [z_t^1, \cdots, z_t^n]$ denotes the concatenation of each frame latent state and n represents the total video frames.

Continually, we need to find each attribute's correlations across different frames. As illustrated in Fig 3, the maximum value in cross-frame DIFT feature similarity indicates the strongest response among tokens within the same attribute, whereas the minimum similarity points to the relationship between tokens of different attributes. To discern the relationship of each query token with the same and different attributes throughout the video, we identify the spatial-temporal positive/negative value for each query on the spatial-temporal axis as follows:

$$\begin{aligned} M_{\text{pos}}^i &= \max(Q^i[K^1, \cdots, K^n]^\top) - Q^i[K^1, \cdots, K^n]^\top) \\ M_{\text{neg}}^i &= Q^i[K^1, \cdots, K^n]^\top - \min(Q^i[K^1, \cdots, K^n]^\top) \end{aligned} \tag{2}$$

These spatial-temporal positive/negative values represent the relationships within the same/different attributes, respectively, allowing us to enhance attention scores among tokens of the same attribute and reduce them among tokens of different attributes to avoid attention leakage.

Modulate Spatial-Temporal Attention Value We follow [17], and modulate the attention map A_i to A_i' for each frame i based on the spatial-temporal positive/negative value, this can be formulated as:

$$\begin{aligned} A_i' &= \text{softmax}(\frac{Q^i[K^N]^\top + M}{\sqrt{d}}), \\ M &= \lambda_t \cdot R_{\text{st}}^i \odot M_{\text{pos}}^i \odot (1 - S_{\text{st}}^i) - \lambda_t \cdot (1 - R_{\text{st}}^i) \odot M_{\text{neg}}^i \odot (1 - S_{\text{st}}^i), \end{aligned} \tag{3}$$

where $R_{\text{st}}^i \in \mathbb{R}^{|\text{queries}| \times |\text{keys}|}$ indicates the query-key pair condition map at frame i, manipulating whether to increase or decrease the attention score for a particular pair. For the tokens in the same attribute across different frames, which will be viewed as a positive pair, leading to an increase in their attention score. In contrast, when the tokens are from different attributes (layouts) in the video, they constitute a negative pair, resulting in a reduced attention score. λ_t is a regularization parameter for timestep t,

controlling modulation function intensity. S_{st}^{i} represents the spatial-temporal regularization for each attribute size. We calculate each attribute class area proportions across video frames, enabling dynamic attention weight adjustments to layout size variations.

Regularize Self Attention Map Beyond Spatial. In the self-attention layer, we aim to avoid attention leakage by increasing attention scores for tokens within the same attribute, while restricting interactions between tokens in different attributes within the same frame or across various frames. Consequently, our query-key condition map is defined as:

$$R_{\text{st}}^{(i),\text{self}} := \begin{cases} 0, \forall j \in [1:N], \text{if } m_l^{(i)}[a] \neq m_l^{(j)}[b], \\ 1, \text{otherwise} \end{cases} \tag{4}$$

where a and b are token indexes of the query and key in the condition map, respectively. i, j are frame indices, and m_l represents a binary map for a single attribute. If tokens belong to different attributes across frames, the value is zero.

Discrete Text control in Cross-Attention Layers In the cross-attention layer, to achieve precise text-to-attribute control, we employ discrete text embeddings for each attribute focused on corresponding layouts. Based on layout masks m_L^i, we tailor the cross-attention query-key condition maps for textual cues to target specific regions:

$$R_{\text{st}}^{(i),\text{cross}} := \begin{cases} 0, & \text{if } k[b] = 0 \\ m_{k[b]}^{i}, & \text{otherwise} \end{cases} \tag{5}$$

where $m_{k[b]}^{i}$ is a binary map fitting the spatial resolution at i^{th}frame. $k[b] \in \mathbb{R}^{|\text{keys}|}$ maps the b_{th} text token to its attribute index, with zero indicating no association. Take the phrase "An Iron Man on a snow covered court": we have two attributes, with $\tau_1 =$ man and $\tau_2 =$ court. The value of $k[0, 3, 4]$ is zero for unrelated tokens, $k[1, 2] = 1$ for "Iron Man", and $k[5, 6, 7] = 2$ for "snow covered court".

4 Experiments

4.1 Experimental Settings

Datasets We validate our EVA model on a dataset comprising 26 videos, sourced from DAVIS [24], TGVE[3], and the Internet[4]. This dataset includes 14 single-object and 12 multi-object human-centric complex motion videos. For each video, we manually annotate the descriptions of the source video and create 3 creative textual prompts, encompassing single-attribute, multi-attribute, multi-object and background editing. Ultimately, this process results in the construction of 78 video-text pairs. Each video is cropped and resized to a resolution of 512x512, containing 16-32 frames.

Metrics Following [27, 37], we assess the video quality using five metrics: **Frame Acc** measures frame-wise editing accuracy, which computes the percentage of frames with higher CLIP similarity to the target prompt than the source, following [27]. **CLIP-T** is the average cosine similarity between the input prompt and all video frames, which

[3] https://sites.google.com/view/loveucvpr23/track4

[4] https://www.istockphoto.com/

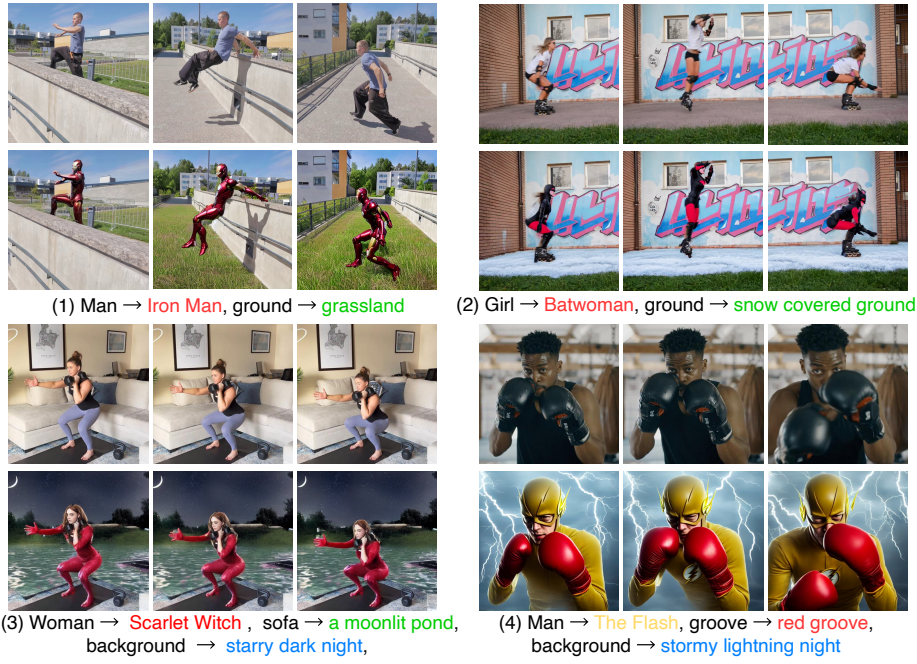


Fig. 6: Single-object multi-attribute editing results. We refer the reader to our webpage for more examples and full-video results.

is used to measure textual alignment. We also follow [8, 10] to measure the temporal consistency by CLIP-F and Warp-Error [18]. **CLIP-F** measures the average cosine similarity between all pairs of consecutive frames, indicating global-level temporal consistency. **Warp-Err** calculates the pixel-level difference by warping the edited video frames according to the estimated optical flow of the source video, extracted by RAFT-Large [34]. This metric provides a more detailed measure of temporal consistency at the pixel level. Assessing editing performance solely with these metrics may not offer a holistic view, as unedited videos could still yield low Warp-Err or high CLIP-F scores. Therefore, following [8], we adopt **Q-edit** = CLIP-T/Wrap-Err as a comprehensive score for video editing quality. For brevity, we scale up Frame Acc/CLIP-F/CLIP-T/Warp-Err all by 100.

Implementation details For our implementation, we inflate a pretrained 2D Stable Diffusion [30] v1.5 model along with ControlNet [40] as the pretrained model. We employ the user-interactive mode of SAM-Track [7] for layout condition, which allows users to specify the areas they wish to edit by clicking to create masks. PCA & clustering or thresholding from cross-attention maps falls short in accurately isolating tiny objects such as “tennis ball" and “racket" due to their limited resolution. To enhance the consistency of edited videos, we adopt DDIM inversion. Our DDIM inversion and denoising steps are all set to 50. To improve efficiency, we have implemented slice attention within ST Layout Attn, which further saves memory usage. We apply ST Layout Attn in the initial 15 denoising steps and set other hyper-parameters the same as [17]. All the experiments are done with one NVIDIA A40 GPU.

(1) Back man → Iron Man, front man → Batman
bridge, ground → snow covered bridge and ground

(2) Left man → Iron Man, rightman → Batman
ground → yellow floor

(3) Left man → Iron Man, right man → Batman
leaves → crimson maple trees,
road→ snow covered road

(4) Man → Batman, woman → Batwoman
ground → rain soaked ground
red wall → stormy lightning night

Fig. 7: Multi-object multi-attribute editing results. Our EVA supports accurate identity mapping in complex motion videos.

Baselines We compare with 4 state-of-the-art video editing methods: (1) Fatezero [27] preserves layout information using source video attention maps. (2) ControlVideo [41] is a training-free method conditioned on ControlNet [40]. (3) Tokenflow [10] samples keyframes and performs linear combinations of features for visual consistency. (4) GroundVideo [15] uses a word-to-bounding box approach for multi-attribute control. For fairness, all baselines are equipped with ControlNet pose guidance.

4.2 Results

Single Object Multi-Attribute Editing In Fig 6 and 8 top, we showcase EVA's editing results in single-object editing. Our method maintains the original layout and critical local details like "railing" in Fig 6 (1) and "mountains" in Fig 8 top. By decoupling object motion and identity, edited objects seamlessly follow the original movements, even in complex scenarios with view changes. Additionally, EVA can edit backgrounds that contrast with the original video's style, such as "a pond under moonlight" in Fig 6 (3) or "a stormy lightning night" in Fig 6 (4).

Multi Object and Attribute Editing Fig 7 displays EVA's multi-object editing outcomes. Our method omits the need for detailed object descriptions. Simple phrases like

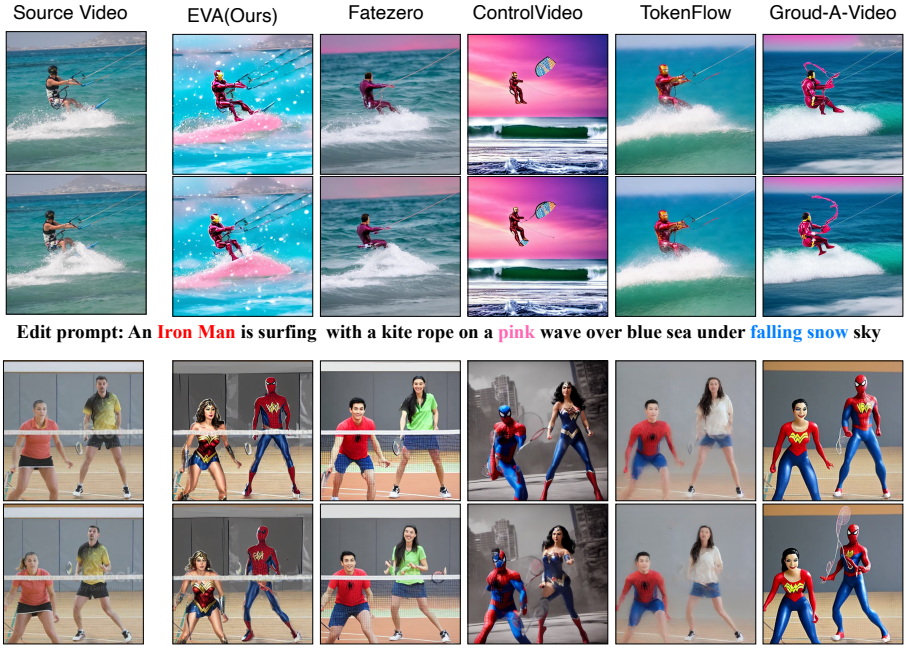


Fig. 8: Qualitative comparisons to the existing video editing methods. The top figure shows single-object editing results, and the bottom displays multi-object editing results. We refer the reader to our project page for full-video comparisons.

"a man and another man" suffice to define source objects' identities. More importantly, our approach enables identity swapping in multi-object scenes, as shown in Fig. 1 right and Fig. 7 (3) (4). This is enabled by discrete text embedding control over each attribute.

4.3 Qualitative and Quantitative Comparisons

Qualitative Comparison Fig 8 compares our editing results with other baseline methods on single/multi-object videos. (1). In single-object editing (Fig 8 top), FateZero [27] failed to edit the object and mistakenly edited the pink wave onto the sky. This error arose because attention weights were not precisely aligned with each attribute's words before the word swap. ControlVideo [41] also incorrectly modified the pink wave onto the sky and could not preserve the layout of the source video. TokenFlow [10] edits the object into "Ironman" but erases the background mountains and is unable to edit the background. Ground-A-Video [15], using word-to-bounding box control, confuses the "kite rope" with the "Iron Man" and fails to edit the "snow sky" and "pink wave". It struggles with preserving local details within the bounding box and lacks awareness of negative examples for adjacent layout weights, which should be mutually exclusive. (2). In multi-object editing (Fig 8 bottom), FateZero, ControlVideo, and Tokenflow all mistakenly confused the "man" and "woman" subjects due to a lack of text-to-attribute control. Ground-A-Video edits the "man" into "Spiderman" but suffers from attention leakage, where textures of "Spiderman" leak onto the "woman," and it fails to retain the

local details of the source video, such as “badminton rackets" and “nets." For additional comparison, please refer to the project page.

Method	Frame Acc ↑	CLIP-F ↑	CLIP-T ↑	Warp-Err ↓	Q_{edit} ↑
FateZero	73.68	95.75	33.78	3.08	10.96
ControlVideo	95.03	**97.71**	34.41	4.73	7.27
TokenFlow	89.26	96.48	34.59	2.82	12.28
Ground-A-Video	95.03	95.17	35.09	4.43	7.92
EVA(ours)	**98.92**	96.09	**36.56**	**2.73**	**13.39**

Table 1: Quantitative comparison with other methods, the best results are **bolded**

Quantitative Comparison (1) Automatic Metrics Table 1 presents a quantitative comparison with other methods. Our EVA, with precise text-to-attribute control, achieves the highest frame edit accuracy and the best CLIP-T scores. Although our frame consistency on CLIP-F is slightly lower than ControlVideo [41], we significantly outperform it in Warp-error and CLIP-T scores. This stems from the fact that ControlVideo keeps the background static. Evaluating temporal consistency with pixel-level optical flow provides a more accurate measure than calculating a global temporal score with CLIP [28]. Moreover, our method achieves the highest overall editing score Q_{edit}. In general, our EVA demonstrates superior performance on all evaluation metrics.

(2)User Study While automatic metrics provide a general comparison, they often fail to align well with human perception [20] and cannot accurately verify the accurate editing of each local attribute or the preservation of layout and undesired editing areas. Therefore, we conducted a user study for a more detailed comparison. We evaluated the quality of edited videos from four aspects: (1). Subject edit accuracy (accuracy of each attribute’s editing), (2). Layout preservation (accuracy of preserving undesired editing areas and overall layout), (3). Motion Alignment, and (4). Overall Preference.

Method	Subject Edit Acc ↑	Layout Preservation ↑	Motion Alignment ↑	Overall Preference ↑
FateZero	2.99	3.37	3.93	2.98
ControlVideo	2.66	2.04	2.50	2.18
TokenFlow	2.27	2.65	2.52	1.99
Ground-A-Video	3.45	3.64	3.60	3.16
EVA(ours)	**4.42**	**4.21**	**4.25**	**4.15**

Table 2: User study comparison with other methods, The number denotes the average score on a scale from 1 to 5 (worst to best). The best results are **bolded**.

We invited 20 participants to rate 78 video-text pairs on a scale of 1 to 5 across these four criteria. Table 2 shows that our method significantly outperformed FateZero, Tokenflow, and ControlVideo in subject edit accuracy, layout preservation, and motion alignment. Furthermore, our approach exceeded competing related work Ground-A-Video [15] in four human evaluation metrics.

4.4 Ablation Study

To assess the contributions of different components in our proposed EVA framework, we conducted ablation studies with the following designs:

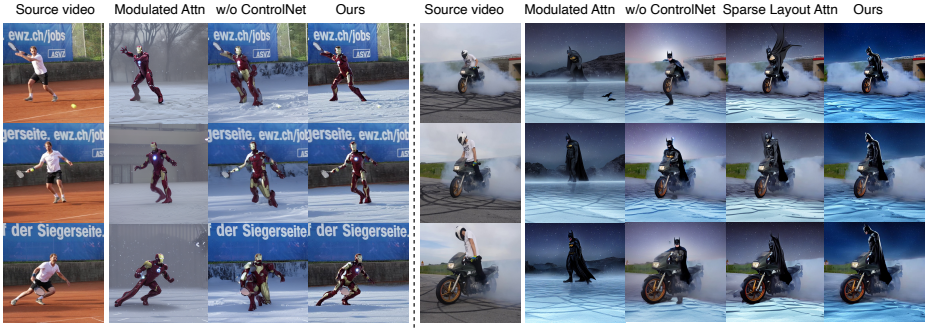


Fig. 9: Comparison of Modulated Attention [17], the absence of ControlNet, and EVA in complex motion, with Sparse Layout Attn results for video object size changes displayed on the right.

Method	CLIP-F ↑	CLIP-T ↑	Warp-Err ↓	Q_{edit} ↑
w/o ControlNet	94.72	36.00	2.75	13.09
w/o Latent Blend	**97.02**	32.56	2.88	11.31
w/o Layout guidance	95.31	35.37	3.11	11.37
Sparse-casual Layout Attn	95.75	35.63	2.83	12.59
EVA(ours)	96.09	**36.56**	**2.73**	**13.39**

Table 3: Quantitative ablation of key components of EVA.

Latent Blend In the second and last column of Fig 9 (1) and (2), we compare the original modulated attention in DenseDiffusion [17] with our method. For fairness, we equip it with ControlNet pose guidance. Our findings show that the modulated attention fails to maintain the source background, resulting in varied backgrounds across frames under the same random seed. Furthermore, using [17] alone struggles to preserve details like the “tennis racket" in Fig 9 (1) and “motorcycle" ,“smoke” in Fig 9 (2).

ControlNet Next, we ablate the use of ControlNet-Pose, showcased in the third column of Fig 9. It is evident that the edited result’s posture does not match the source human posture. Therefore, in human-centric complex motion videos, employing pose conditions for intra-object structure information is necessary.

Spatial-Temporal Layout-Guided Attention Fig 10 contrasts three conditions: without ST-Layout Attn (second row), Sparse Casual Layout-guided Attention (SC-Layout Attn, third row), and with ST-Layout Attn (fourth row). As shown in the second row, the absence of our ST-Layout Attn leads to incorrect identity mapping. For instance, the left man was supposed to be “Iron Man," and the right is “Batman," but their identities were swapped in the second row of Fig 10 left. This underlines the effectiveness of our ST-Layout Attn on accurate identity mapping in multi-object scenes.

Also, sparse layout (the first and the preceding frame) guidance exhibits several limitations, notably: (1) Limited Receptive Field for Negative Values: The sparse method’s reduced receptive field for query tokens positive/negative value selection across different layouts. The unsuitable selection of negative values results in attention leakage, manifesting as a yellow head of “Batman” in Fig 10 left third row and disordered web-

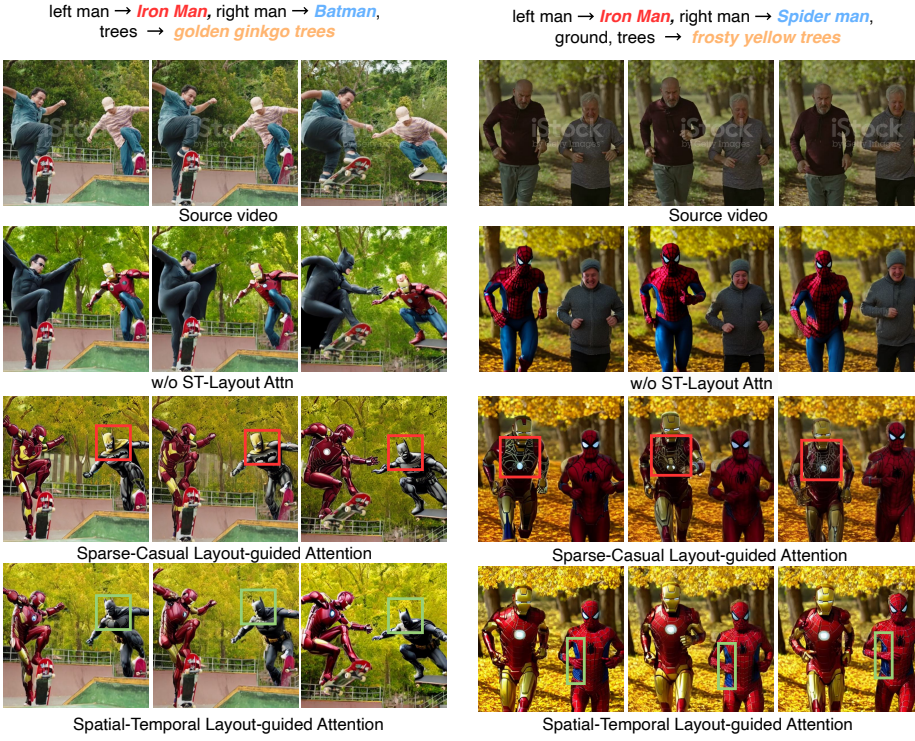


Fig. 10: Qualitative comparison of results without ST-Layout Attn, Sparse-Casual Layout-guided Attention (SC-Layout Attn) and our Spatial-Temporal Layout-guided Attention (ST-Layout Attn). Our method results in accurate identity mapping and distinct local details without attention leakage.

like textures in Fig 10 right across Iron Man's chest (red box). (2) Reduced Interaction Across Full Frames: A lack of interaction across the entire video frames results in the loss of local details, such as the distinctive blue sides of Spider-Man (green box in Fig 10 right). Moreover, this limited interaction contributes to an overall duller color tone. The quantitative results in Table 3 further confirm the effectiveness of ST-Layout Attn.

5 Conclusion

In scenarios of complex human-centric motion, we propose EVA, a general framework for multi-attribute and multi-object video editing. We introduce a Spatial-Temporal Layout-Guided Attention mechanism, which leverages the intrinsic positive and negative correspondences of cross-frame diffusion features; our ST-Layout Attn not only ensures that each text embedding concentrates on its respective attribute, but also enhances the internal coherence within attributes and keeps the exclusivity of attention weights among different attributes. Benefiting from precise attention weighting, EVA can be extended to editing in multi-object scenes. We demonstrate EVA's superior performance in multi-attribute and multi-object editing through extensive experiments.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023) 2, 3
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022) 3
3. 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) 4
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields (2018). arXiv preprint arXiv:1812.08008 (1812) 6
5. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23040–23050 (2023) 4
6. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023) 3
7. Cheng, Y., Li, L., Xu, Y., Li, X., Yang, Z., Wang, W., Yang, Y.: Segment and track anything. arXiv preprint arXiv:2305.06558 (2023) 6, 9
8. Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S.: Flatten: optical flow-guided attention for consistent text-to-video editing. arXiv preprint arXiv:2310.05922 (2023) 2, 4, 9
9. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022) 3
10. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023) 2, 4, 9, 10, 11
11. Gu, Y., Zhou, Y., Wu, B., Yu, L., Liu, J.W., Zhao, R., Wu, J.Z., Zhang, D.J., Shou, M.Z., Tang, K.: Videoswap: Customized video subject swapping with interactive semantic point correspondence. arXiv preprint arXiv:2312.02087 (2023) 4
12. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023) 4
13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022) 3, 4
14. Hu, Z., Xu, D.: Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. arXiv preprint arXiv:2307.14073 (2023) 2
15. Jeong, H., Ye, J.C.: Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. arXiv preprint arXiv:2310.01107 (2023) 2, 4, 5, 10, 11, 12
16. Kasten, Y., Ofri, D., Wang, O., Dekel, T.: Layered neural atlases for consistent video editing. ACM Transactions on Graphics (TOG) **40**(6), 1–12 (2021) 4
17. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: ICCV (2023) 3, 6, 7, 9, 13
18. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018) 9
19. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. CVPR (2023) 5

20. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. arXiv preprint arXiv:2310.11440 (2023) 12
21. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 4
22. Ouyang, H., Wang, Q., Xiao, Y., Bai, Q., Zhang, J., Zheng, K., Zhou, X., Chen, Q., Shen, Y.: Codef: Content deformation fields for temporally consistent video processing. arXiv preprint arXiv:2308.07926 (2023) 4
23. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 3
24. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016) 8
25. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023) 3
26. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021) 4
27. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023) 2, 4, 5, 8, 10, 11
28. 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) 12
29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 2
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 4, 5, 6, 9
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 2
32. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022) 2
33. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=ypOiXjdfnU 2, 3
34. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020) 9
35. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023) 4
36. Wang, K., Yang, F., Yang, S., Butt, M.A., van de Weijer, J.: Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. arXiv preprint arXiv:2309.15664 (2023) 3

37. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, 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. arXiv preprint arXiv:2212.11565 (2022) 4, 5, 8
38. Yang, S., Zhou, Y., Liu, Z., , Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. In: ACM SIGGRAPH Asia Conference Proceedings (2023) 2, 4
39. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023) 4
40. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (October 2023) 4, 5, 6, 9, 10
41. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023) 2, 4, 5, 10, 11, 12
42. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: Controlvideo: Adding conditional control for one shot text-to-video editing. arXiv preprint arXiv:2305.17098 (2023) 2, 4, 5
43. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023) 4