

Generative Video Diffusion for Unseen Novel Semantic Video Moment Retrieval

Dezhao Luo¹, Shaogang Gong^{1*}, Jiabo Huang², Hailin Jin³, and Yang Liu^{4,5*}

¹Queen Mary University of London

²Sony AI

³Adobe Research

⁴WICT, Peking University

⁵State Key Laboratory of General Artificial Intelligence, Peking University

{dezhao.luo, s.gong}@qmul.ac.uk, raymond.huang@sony.com, hljin@adobe.com, yangliu@pku.edu.cn

Abstract

Video moment retrieval (VMR) aims to locate the most likely video moment(s) corresponding to a text query in untrimmed videos. Training of existing methods is limited by the lack of diverse and generalisable VMR datasets, hindering their ability to generalise moment-text associations to queries containing novel semantic concepts (unseen both visually and textually in a training source domain). For model generalisation to novel semantics, existing methods rely heavily on assuming to have access to both video and text sentence pairs from a target domain in addition to the source domain pairwise training data. This is neither practical nor scalable. In this work, we introduce a more generalisable approach by assuming *only* text sentences describing new semantics are available in model training *without* having seen any videos from a target domain. To that end, we propose a Fine-grained Video Editing framework, termed FVE, that explores generative video diffusion to facilitate fine-grained video editing from the seen source concepts to the unseen target sentences consisting of new concepts. This enables generative hypotheses of unseen video moments corresponding to the novel concepts in the target domain. This fine-grained generative video diffusion retains the original video structure and subject specifics from the source domain while introducing semantic distinctions of unseen novel vocabularies in the target domain. A critical challenge is how to enable this generative fine-grained diffusion process to be meaningful in optimising VMR, more than just synthesising visually pleasing videos. We solve this problem by introducing a hybrid selection mechanism that integrates three quantitative metrics to selectively incorporate synthetic video moments (novel video hypotheses) as enlarged additions to the original source training data, whilst minimising potential detrimental noise or unnecessary repetitions in the novel synthetic videos harmful to VMR learning. Experiments on three datasets demonstrate the effectiveness of FVE to unseen novel semantic video moment retrieval tasks.

Introduction

Given an untrimmed video and a sentence query, video moment retrieval (VMR) aims to locate the most relevant video moment(s) semantically corresponding to the query. This

task is challenging because it requires extracting semantic associations between visual and textual data with precise time locations. Annotating VMR datasets requires indexing temporal video moments with corresponding sentences and distinguishing them from contextual moments within the video, which is a more intricate and less scalable process compared to labelling image-text or video-text pairs.

Due to the lack of large-scale video moment-text datasets, VMR models are struggling to learn generalisable *novel* moment-text associations beyond the training source domains, resulting in an inferior cross-domain adaptation where training and testing data display biases. In contrast to the biases in the moment location or length (Hao et al. 2022), we tackle a more intricate challenge: semantic biases across domains, where the semantic in the testing domain is *novel* to the training domain. By ‘novel’, it means novel text vocabularies and their corresponding video moments both unseen in source domain training. Comparing to previous methods employing co-training strategies on both source and target datasets (Cai, Huang, and Gong 2022), and other methods (Nam et al. 2021) creating pseudo moment-text associations by generating textual descriptions for a target video, we propose a more scalable and accessible approach to learning generalisable VMR to unseen novel semantics by exploring target domain sentences describing new semantics only, without any videos from the target domain.

Recent successes in generative diffusion models (Rombach et al. 2021; Wang et al. 2023b) have demonstrated the power of hypothesising new holistic videos with text prompts. To benefit novel semantic VMR, a potential approach is to generate videos conditioned on both source video moments and a target sentence of novel semantics/concepts. However, there are several non-trivial challenges in synthesising a meaningful visual hypothesis for a VMR video displaying the same subject performing different actions in a similar environment (background). The first challenge lies in regulating the generation of a video moment based on novel semantics referenced in a target domain sentence while retaining other contextual information intact. Existing text conditioned video generation (Wang et al. 2023b) or editing methods (Wu et al. 2023) lack specific constraints or image conditioned methods (Jiang et al. 2024), leading towards inaccurate subject details (shown in Table 6, Fig. 3 and the Supplementary). This highlights the

*Corresponding authors

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

need for an *instance-preserving video action editing* method constrained by accurate subject specifics (fine-grained details). The second challenge is that existing generative video editing techniques (Wu et al. 2023; Liu et al. 2023) heavily depend on manual intervention to choose appropriate editing text prompts to ‘guide’ credible outcomes. Automated video generation poses a risk of producing implausible or trivial repetitive videos, not meaningful and potentially detrimental to the generalisability of a VMR model if trained with such data. An unsolved critical problem of existing methods is how to select video hypothesis generation that can optimise VMR model learning to novel semantic concepts.

To address these challenges, we introduce a Fine-grained Video Editing framework (FVE) which explores *fine-grained* generative video diffusion to finely edit videos of seen semantics from the source domain, guided by target sentences of new unseen semantics, thus hypothesising more meaningful unseen target video moments featuring these novel semantics for VMR training. To address the first challenge of moment generation by involving only ‘sentence-referred’ local visual variations while maintaining the background and the subject details from the original source video, we design a 2-stage video editing model for simultaneous accurate subject preservation and fine-grained detail change guided by unseen novel concepts. Specifically, we first train an image diffusion model to align a text token with instances present across a set of video frames, ensuring precise visual-textual alignment. Second, we treat video frames as a sequence and introduce a temporal layer within the image diffusion model to learn the video motions. To tackle the challenge of minimising potential noisy and/or trivial repetitions in synthesising towards VMR training, we formulate three quantitative metrics aimed at filtering out implausible samples while selecting beneficial data for training the VMR model. Firstly, we introduce a cross-modal relevance metric to assess the semantic relevance between the target prompt and the generated video moment, thereby ensuring the quality of the moment-text association. Secondly, a uni-modal structure metric is introduced to evaluate the visual similarity of video moments between the source and generated moments before their utilisation in VMR training, which provides insights into video fidelity. Lastly, we introduce the model performance disparity metric, emphasising the importance of enriching VMR training by selecting more diverse synthesised hypotheses rather than duplicating similar visual content to the source domain videos. In practice, a synthetic video moment is incorporated into training only if it exhibits inferior retrieval performance, i.e. selection by a VMR discriminative constraint measured by a synthetic video’s model performance disparity being high.

We make three contributions: (1) Instead of collecting moment-query pairs with novel semantic associations for VMR training in every new target domain in order to tackle the semantic bias across domains, we propose to *only* leverage target domain sentences containing unseen new semantics without any videos from the target domain. (2) We propose a Fine-grained Video Editing framework (FVE) to adapt a source domain video according to a target sentence of novel concepts as a controller for editing source do-

main videos to synthesise target moment-text associations for VMR training. Specifically, we enhance generation control with an instance-preserving video diffusion model, addressing the limitations of existing methods in maintaining video subjects while altering action details. Additionally, we propose a hybrid data selection strategy to curate the most beneficial simulated videos for VMR training. (3) We demonstrate the effectiveness of FVE on both VMR and action editing tasks under diverse datasets.

Related Works

Cross-Domain Video Moment Retrieval. To solve the problem of lacking an extensive video dataset to train an effective generalisable VMR model, CanShuffle (Hao et al. 2022) proposed a data augmentation strategy to solve the temporal bias problem by sacrificing the temporal semantics (Cai et al. 2024), lacking the ability to understand novel semantics inherited in a video moment and its corresponding description. For broader semantic understanding, previous methods (Luo et al. 2023; Zheng et al. 2024) applied large-scale vision-language models like CLIP (Radford et al. 2021), InternVideo (Wang et al. 2022) or BLIP2 (Li et al. 2023), but they still lack the ability to localise fine-grained moment-text associations due to the pre-training on coarse (holistic broad-strokes weakly-supervised) image-text or video-text pairs. Although unsupervised methods (Nam et al. 2021; Liu et al. 2022a; Zheng et al. 2023) were proposed to generate pseudo moment-text associations from unlabelled videos and use them to train fully supervised VMRs, they are inherently limited by both the quantities of videos available and error-propagation from self-labelling. In light of these challenges, contemporary cross-domain solutions are either impractical due to insufficient training videos or suboptimal in capturing fine-grained associations. Our method is more scalable and optimised for learning novel target semantic associations from only target text query without any videos from the target domain.

Video Diffusion Models. Current video diffusion models have demonstrated promising outcomes in the domains of video synthesising (Hong et al. 2022; Wu et al. 2023; Feng et al. 2023). We primarily focus on two specific perspectives: video generation and video editing.

Video Generation. Due to the difficulties in collecting high-quality video data, existing video generation methods leverage both images and videos in model training (Singer et al. 2022; Blattmann et al. 2023; Wang et al. 2023b). Moreover, predominantly trained on images (2.3B images v.s. 10M videos), they are incapable of generating fine details in human-centred videos of complex dynamics. Further, they rely solely on text prompts and are ‘blind’ to visual controls from each subject instance of an action and the scene context (background environment).

To generate videos with better dynamics, motion customisation methods (Zhao et al. 2023; Wei et al. 2024) concentrated on learning a specific motion with given samples, which requires additional labelled text-video pairs for each action. To generate videos with specified subjects, Videobooth (Jiang et al. 2024) proposed to input image

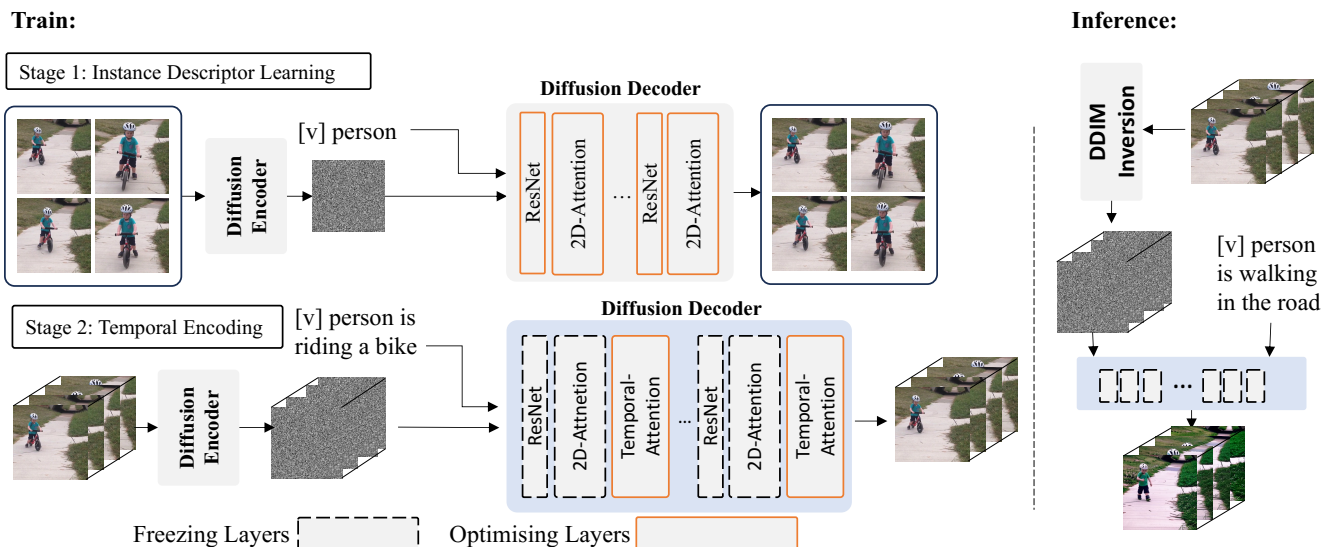


Figure 1: Our designed instance-preserving action editing model. We first take the video as a set of images and train an image diffusion model to align a special text token with the instance shared between those frames. Subsequently, we take those frames as a sequence and freeze the layers in the image diffusion model, and append a temporal layer to capture the video motions.

as a prompt, however, it is still insufficient to handle human instances with rich details (shown in the Supplementary). Dreamix (Molad et al. 2023) introduced a subject-driven action generation by a mixed reconstructing strategy of subject-driven image generation (Ruiz et al. 2023; Kumari et al. 2023; Chen et al. 2023) and video generation (Wu et al. 2023). However, the mixed training of reconstructing images and actions (Molad et al. 2023) is likely to entangle the two features in latent embedding, resulting in inaccurate specific information or overfitting on the source action (Chen et al. 2023). This highlights the need to design a better subject-preserving video editing method.

Video Editing. Existing methods (Wu et al. 2023; Jeong and Ye 2023; Yan et al. 2023; Liew et al. 2023) have demonstrated proficient object editing capabilities by manipulating the associated textual descriptions. To maintain the integrity of the background and prevent alterations to regions not in the focus on change, plug-and-play techniques (Qi et al. 2023; Liu et al. 2023) employed a decoder with cross-attention derived masks to protect unrelated areas, necessitating users to interactively discern and selectively specify the interchangeable parts between the source and target prompts. Other methods (Geyer et al. 2023; Lu et al. 2023; Yang et al. 2023) were proposed to eliminate the training process on a specific video and to edit video objects directly using priors from image diffusion models (Rombach et al. 2021). However, current video editing methods fall short in the realm of VMR video simulation, particularly in crafting moments with a consistent subject engaging in various actions. These methods lack the necessary controls to edit actions while preserving the distinctive features of the subject. VMR tasks typically require distinguishing subtle differences between matching and non-matching video moments, and they often feature the same subjects, backgrounds, and

video styles. Therefore, the aforementioned shortcomings of existing video editing methods restrict inherently generating meaningfully diverse and visually plausible data for learning novel unseen concepts in VMR.

Overall, existing generative models (Liu et al. 2023; Qi et al. 2023) heavily rely on designing a delicate generation control, e.g., through an interactive selection of a target prompt, to produce plausible videos through trial-and-error. How to design effective automatic controls for generating target moment-text associations capable of optimising novel semantic cross-domain video moment retrieval model learning has not been studied, nor it is straightforward.

Method

Our aim is to simulate video moment retrieval (VMR) training data with fine-grained moment-text associations, and autonomously regulate the video generation process using a collection of sentences, without any human intervention by interactive text prompts or reliance on target exemplar videos. In this section, we first recap diffusion models, then present our Fine-grained Video Editing framework (FVE) with an instance-preserving action editing model (Fig. 1) and an automatic video generation and hybrid selection pipeline (Fig. 2).

Latent Diffusion and DDIM Inversion

Latent Diffusion Models (LDMs). LDMs are introduced to diffuse and denoise data (Sohl-Dickstein et al. 2015) within a compressed latent space. For an image x , the process starts with the encoding of x into a latent representation (Kingma and Welling 2013): $z = \mathcal{E}(x)$. Gaussian noise is then added to this representation to create z_t at timestep t . A denoising autoencoder is subsequently trained to predict the Gaussian

noise in the latent representation, aiming to reverse the noise addition. The objective is defined as:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0, I), t} \left[\|\epsilon - \epsilon_{\theta}(z_t, t, p)\|_2^2 \right], \quad (1)$$

where ϵ_{θ} is a U-Net (Ronneberger, Fischer, and Brox 2015) architecture conditioned on a timestep t and a text prompt embedding p and the visual input z_t .

DDIM Inversion. DDIM Inversion (Song, Meng, and Ermon 2020) maps a clean latent representation z to its noisy counterpart using a sequence of reverse timesteps from $t = T - 1$ to 1. The DDIM iterative process is defined by:

$$\hat{z}_t = \sqrt{\alpha_t} \hat{z}_{t-1} + \left(\sqrt{1 - \alpha_t} - \sqrt{\frac{1 - \alpha_t}{\alpha_{t-1}}} \right) \epsilon_{\theta}, \quad (2)$$

where \hat{z}_t denotes the estimated noisy latent state at timestep t and α_t represents the variance schedule at timestep t .

Instance-Preserving Action Editing

To simulate a VMR video involving the same subject executing different actions, we aim to address the limitations of current video editing methods (Wu et al. 2023; Liu et al. 2023), which lack controls for constraining subject specifics. Additionally, we aim to overcome the inaccuracies in instance information learned by subject-driven video generation methods (Molad et al. 2023), attributed to their mixed subject-specific and motion training strategy. Specifically, we separate the learning of subject-specific instance information and the video motion (Fig. 1). We first train an image model to align a text token ‘[v] person’ with the visual information of the shared subject instance across frames. This alignment is achieved in the 2D-attention layer containing self-attention (visual) and cross-attention (textual-visual) layers. Subsequently, we *freeze* the learned textual-visual alignment and introduce a temporal layer following the 2D-attention layer to capture the video motion.

Stage1: Instance Descriptor Learning. In the first stage, we take the video as a set of unordered frames for instance descriptor learning. For effective learning, we select a subset of frames from the whole frame set, aiming to maximize diversity whilst ensuring frame clarity by minimising noise such as motion blur. Mathematically, given a frame f_i and its immediate neighbouring frames f_{i-1} and f_{i+1} , we seek frames to maximize the following function:

$$\Phi(f_i) = \delta(f_i, f_{i-1}) + \delta(f_i, f_{i+1}) + \chi(f_i), \quad (3)$$

where δ denotes the dissimilarity measure between frames using histogram, and χ signifies the frame’s clarity, determined with the Laplacian operator to evaluate the visual sharpness. We select 10 frames with higher $\Phi(f_i)$ scores.

For the training of the instance descriptor, we leverage the Dreambooth strategy (Ruiz et al. 2023) to train an image diffusion model to reconstruct the selected video frames from a text token ‘[v] person’, while simultaneously reconstructing images of other instances based on the prompt ‘a person’. The model’s goal is to embed the visual instance within the

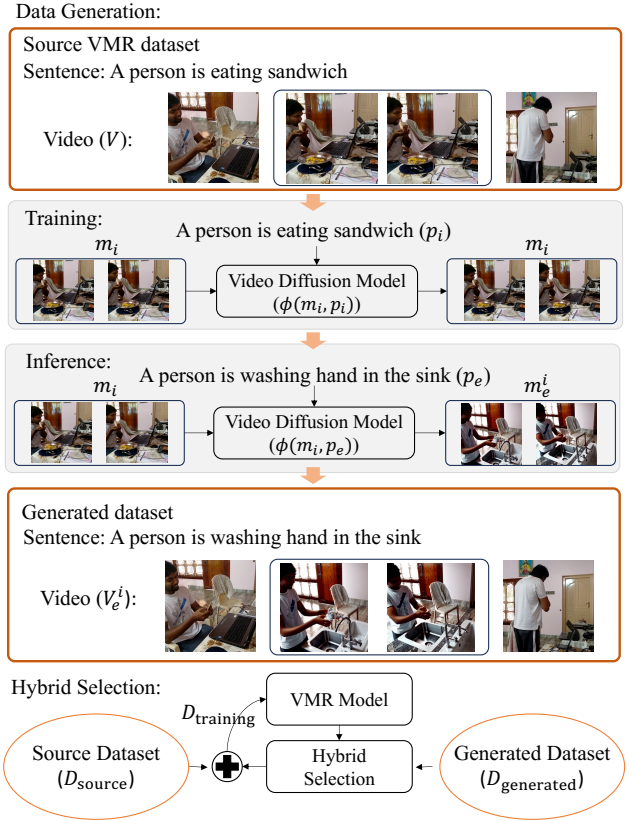


Figure 2: Data generation and hybrid selection. For data generation, we first train the video diffusion model ϕ to align moment m_i with a sentence p_i , then we use an editing prompt p_e to edit the moment to m_e^i . The hybrid selection strategy includes a cross-modal relevance and unimodal structure score to select high-quality generation, as well as a model performance disparity to select beneficial data for VMR training.

output domain specified by the text token ‘[v] person’. The objective L_{IDL} is formulated as:

$$L_{IDL} = \mathbb{E}_{\mathcal{E}(x_{inst}), \epsilon \sim \mathcal{N}(0, I), t} \left[\|\epsilon - \epsilon_{\theta}(z_t, inst, t, p_{inst})\|_2^2 \right], \\ + \mathbb{E}_{\mathcal{E}(x_{class}), \epsilon \sim \mathcal{N}(0, I), t} \left[\|\epsilon - \epsilon_{\theta}(z_t, class, t, p_{class})\|_2^2 \right], \quad (4)$$

where x_{inst} and x_{class} denote the input of instance image and class images (other instances in the same class), p_{inst} and p_{class} denote their corresponding prompts. After training, the text token is aligned with instance visual information, enabling a generalisable generation by combining it with other concepts.

Stage 2: Temporal Encoding. When editing a moment based on a target sentence, it is crucial to preserve the original unrelated motions not referenced by the sentence. This includes maintaining subject-specifics, background, and video style intact. As shown in Fig. 1, we append a temporal attention layer after each 2D-attention layer. In order to fix the alignment of ‘[v] person’ with the visual content, in

contrast to the methods employed in Dreamix (Molad et al. 2023), we freeze previously optimised layers and only train the temporal layer. The training loss L_{TE} is formulated as:

$$\begin{aligned} L_{TE}(m, p) &= \|m - \phi(m, p)\| \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathcal{E}(f_i), \epsilon \sim \mathcal{N}(0, I), t} \left[\|\epsilon - \epsilon_\gamma(z_{t,i}, t, p)\|_2^2 \right], \end{aligned} \quad (5)$$

where n is the frame number in moment m , and each frame is denoted as f_i . ϕ is our model consisting of a U-Net architecture (ϵ_γ) with a temporal layer attached after each 2D-attention layer. p is the prompt describing the action.

Data Generation and Hybrid Selection

Data Generation. Considering a video V , a VMR dataset (Gao et al. 2017) will provide a list of moments in the video with their sentence descriptions, denoted as $\{(m_i, p_i)\}_{i=1}^a$, where m_i denotes the i^{th} moment, p_i denotes the corresponding description and a is the number of moments. An editing sentence is also provided, represented as p_e . As depicted in Fig. 2, our model involves a training and inference stage: we first train our video diffusion model on each moment-text pair as $L_{TE}(m_i, p_i)$, aligning the moment m_i with its textual description p_i . Then we modify the specific moment m_i in the video V using the sentence p_e , as:

$$m_e^i = \phi(m_i, p_e), \quad (6)$$

where m_e^i denotes the p_e -edited version of moment m_i . Then m_e^i replaces the original moment m_i to create a new variant of the video, denoted as V_e^i . For the video V with a moments, we generate a different video variants $\{V_e^1, V_e^2, \dots, V_e^a\}$, where V_e^i contains all original moments except for the i^{th} moment, which is replaced by its edited version m_e^i , resulting in a set of videos where each variant showcases a unique modification at a distinct moment.

Hybrid Selection. Without a delicate selection of the editing prompt, automatic VMR data generation may result in noisy or repetitive videos. In order to select high-quality and beneficial data for VMR training, we design a hybrid selection strategy with three quantitative metrics: cross-modal relevance, uni-modal structure, and model performance disparity. For cross-modal relevance, we notice that a lack of semantic relevance between the source video and the target sentence might result in implausible outcomes where the generated video content does not align well with the provided text. Training a VMR model using noisy pseudo moment-query pairs misleads the model to learn inaccurate moment-text associations. To this end, we introduce a cross-modal relevance score to evaluate the coherence between the target prompt and the generated video moment:

$$s_c(p_e, m_e) = \frac{1}{N} \sum_{i=1}^N \cos(\text{VLM}(p_e), \text{VLM}(f_{m_e}^i)), \quad (7)$$

where s_c denotes the cross-modal relevance score, p_e the editing prompt, $f_{m_e}^i$ the i^{th} frame in the generated (edited) moment m_e , and N the frame number in m_e . VLM denotes a vision-language model pre-trained on large-scale datasets.

For uni-modal structure score, we evaluate the visual consistency between the source and generated video moment:

$$s_u(m_s, m_e) = \frac{1}{N} \sum_{i=1}^N \cos(\text{VM}(f_{m_s}^i), \text{VM}(f_{m_e}^i)), \quad (8)$$

where s_u is the uni-modal structure score, $f_{m_s}^i$ is the i^{th} frame in the source moment m_s . VM is a visual encoder pre-trained on a large-scale vision dataset, which can predict the general structure, including the environment and object features, of an image. Successful and high-quality editing is the generation that both matches the text prompt and maintains the original video structure, so we integrate the two metrics using a harmonic score:

$$s_{cu}(p_e, m_e, m_s) = \frac{2 \times s_c(p_e, m_e) \times s_u(m_s, m_e)}{s_c(p_e, m_e) + s_u(m_s, m_e)}. \quad (9)$$

The selection process is represented as follows:

$$D_{cu} = \text{TOP}_k(\{(d, s_{cu}(p_e, d, m_s)) \mid d \in D_{\text{generated}}\}), \quad (10)$$

where $D_{\text{generated}}$ is the generated dataset, D_{cu} comprises k samples with top s_{cu} .

In addition to the high-quality moment selection, we underscore the significance of enriching VMR training by diverse data that is both visually plausible and semantically meaningful for enhancing VMR novel concept generalisation, rather than simply duplicating existing content from the source domain. To address this, we introduce the third metric, called model performance disparity. This metric measures the degree to which the model’s predictions differ from the ground truth or desired outcomes across various samples. Higher model error rates in certain samples indicate instances where the model struggles to accurately capture the relationship between moments and queries. These samples are earmarked for further analysis or refinement in the training process. In practice, we evaluate VMR on the previously filtered dataset D_{cu} and incorporate only those samples of high disparity in training:

$$\begin{aligned} D_{\text{mpd}} &= \text{TOP}_l(\{(d, -\text{VMR}(d)) \mid d \in D_{cu}\}), \\ D_{\text{training}} &= D_{\text{source}} \cup D_{\text{mpd}}, \end{aligned} \quad (11)$$

where D_{training} comprises the source data D_{source} and additional D_{mpd} data with a length of l , selected with low VMR performance. This enables us to identify and select cases that are not adequately handled by the existing model.

Experiments

To evaluate the effectiveness of our Fine-grained Video Editing framework (FVE), we validate on both video moment retrieval and video action editing tasks.

Video Moment Retrieval

Data Setup. To assess FVE for novel semantic VMR, we employed the ‘novel-word’ split (Li et al. 2022) on Charades-STA (Gao et al. 2017), where the testing split

¹These methods are not reported on the same split.

Method	Year	Target	Charades-STA			
			R1@0.5	R1@0.7	mIoU	
EVA ¹	2022	Video	40.21	18.77	-	
MMCDa ¹	2022	& Text	54.80	35.77	-	
I3D feature						
TMN	2018	No	9.43	4.96	11.23	
TSP-PRL	2020		14.83	2.61	14.03	
2D-TAN	2020		29.36	13.12	28.47	
LGI	2020		26.48	12.47	27.62	
VSLNet	2020		25.60	10.07	30.21	
VISA	2022		42.35	20.88	40.18	
VDI [†]	2023		46.19	26.19	40.95	
FVE (Ours)	2025		48.51	28.48	42.67	
Slowfast feature						
M-DETR	2021		No	43.45	21.73	38.37
QD-DETR	2023	48.20		26.19	43.22	
UVCOM [†]	2024	48.63		28.57	42.65	
MESM [†]	2024	51.08		29.78	44.16	
FVE (Ours)	2025	Text	52.37	31.94	44.59	

Table 1: Novel-word testing on Charades-STA. The ‘Target’ column indicates the information required from the target domain. Symbol ‘†’ indicates our implementation with the author-released code.

contains ‘novel-words’ not seen in the training split. For QVHighlights (Lei, Berg, and Bansal 2021) and TaCoS (Regneri et al. 2013), we sample sentences from the standard training split and exclude them from the training set. In our implementation, we selected 50/300/300 sentences separately from each dataset for data generation. Selection details are shown in the Supplementary.

Implementational Details. For each target sentence, we created 100/50/50 videos for each dataset. This resulted in a total of 5,000/15,000/15,000 generated videos, ready to be chosen to support the training of the VMR model. For hybrid selection, we used CLIP (Radford et al. 2021) to compute the cross-modal relevance score and DINO (Caron et al. 2021) for the uni-modal structure score. We set k to 500, 1500 and 1500 respectively for the three datasets. For the model performance disparity metric, we set l to be 100, 500 and 500 respectively for each dataset. We adopted R1@ μ , mAP@ μ , mIoU, and mAPavg as the evaluation metrics.

Comparisons. We compare our FVE with the following methods: EVA (Cai, Huang, and Gong 2022), MMCDa (Fang et al. 2022), TMN (Liu et al. 2018), TSP-PRL (Wu et al. 2020), 2D-TAN (Zhang et al. 2020b), LGI (Mun, Cho, and Han 2020), VSLNet (Zhang et al. 2020a), VISA (Li et al. 2022), VDI (Luo et al. 2023), M-DETR (Lei, Berg, and Bansal 2021), QD-DETR (Moon et al. 2023), UVCOM (Xiao et al. 2024), MESM (Liu et al. 2024), UMT (Liu et al. 2022b), UniVTG (Lin et al. 2023), MH-DETR (Xu et al. 2023), EaTR (Jang et al. 2023), CBLN (Liu et al. 2021), RaNet (Gao et al. 2021), SeqPAN (Zhang et al. 2021), SMIN (Wang et al. 2021) and MS-DETR (Wang et al. 2023a).

For Charades-STA, we apply our method on two different feature extractors, I3D (Carreira and Zisserman 2017) and

Method	Target	QVHighlights					
		R1		mAP			
		@0.5	@0.7	@0.5	@0.75	avg	
M-DETR	Video & Text	52.89	33.02	54.82	29.40	30.73	
UMT		56.23	41.18	53.38	37.01	36.12	
UniVTG		58.86	40.86	57.60	35.59	35.47	
MH-DETR		60.05	42.28	60.75	38.13	38.38	
QD-DETR		63.06	45.10	63.04	40.10	40.19	
EaTR		61.36	45.79	61.86	41.91	41.74	
MESM		62.78	45.20	62.64	41.45	40.68	
UVCOM		63.55	47.47	63.37	42.67	43.18	
MESM [†]		No	61.95	45.03	60.23	38.94	39.03
UVCOM [†]			61.39	45.45	60.43	40.38	40.30
FVE (Ours)	Text	63.35	47.16	62.17	42.00	41.33	

Table 2: VMR results on QVHighlights. The ‘Target’ column and the symbol ‘†’ denotes the same as Table 1.

Method	Target	TaCoS		
		R1@0.3	R1@0.5	mIoU
VSLNet	Video & Text	29.61	24.27	24.11
2D-TAN		37.29	25.32	-
CBLN		38.98	27.65	-
RaNet		43.34	33.54	-
SeqPAN		31.72	27.19	25.86
SMIN		48.01	35.24	-
MMN		39.24	26.17	-
MS-DETR		47.66	37.36	35.09
MESM		52.69	39.52	36.94
MESM [†]		No	44.01	29.39
FVE (Ours)	Text	48.09	31.92	31.61

Table 3: VMR results on TaCoS. The ‘Target’ column and the symbol ‘†’ denotes the same as Table 1.

Slowfast (Fan et al. 2020) using VDI (Luo et al. 2023) and MESM (Liu et al. 2024) as the baseline separately. As shown in Table 1, with a collection of 50 out of 703 sentences, we improve the performance for Charades-STA from 29.78% to 31.94% on R1@0.7 and we reach the SOTA on all metrics. For QVHighlights, we take Slowfast as the feature extractor and UVCOM as the baseline. As shown in Table 2, we obtain gains in all metrics over the baseline model UVCOM. Also, with only a collection of text, we reach comparable performance to those requiring video&text pairs (row 8 vs. row 11 on R1@0.5 and R1@0.7). For TaCoS, we apply C3D (Tran et al. 2015) as the feature extractor and MESM as the baseline. As shown in Table 3, we also obtain gains in all metrics on over the baseline model MESM. More comparisons are given in the Supplementary.

Ablation Study. In this study, we first eliminate the effect of data volume and then highlight the importance of data selection. We conduct ablation studies using I3D features on the Charades-STA dataset and model VDI. As shown in Table 4, increasing the data volume by adding replicated samples does not enhance performance (row 1 vs. row 2). Moreover, in comparison to randomly sampling data from

Method	Datasize	R1@0.5	R1@0.7	mIoU
No	3533	46.19	26.19	40.95
Concat	4033	46.23	26.51	40.17
R_1		47.19	27.19	41.05
R_2	4033	45.61	26.33	40.40
R_3		44.03	25.47	40.00
FVE (Ours)	4033	48.51	28.48	42.67

Table 4: Ablation on the effect of data volume. ‘No’ denotes no use of generated data, ‘Concat’ video generation through the random concatenation of existing videos. ‘ R_1 ’-‘ R_3 ’ a random sampling with different random seeds, ‘Ddatasize’ denotes the number of videos for each method.

s_{cu}	k	s_{mpd}	l	R1@0.5	R1@0.7	mIoU
×	1500	×	-	43.88	25.04	38.50
	1500			44.89	26.19	40.79
✓	1000	×	-	43.60	24.46	40.05
	500			46.64	27.16	41.21
	100			46.91	28.78	41.14
			200	45.75	26.04	40.88
✓	500	✓	100	48.51	28.48	42.67
			50	46.19	25.47	41.35
✓	100	✓	50	47.34	28.49	42.24
			10	47.48	27.05	41.89

Table 5: Ablation of the hybrid selection. s_{cu} and k denote the score of combining cross-modal relevance and uni-modal structure and its selecting numbers. s_{mpd} denotes model performance disparity selection with a number l . $s_{cu} = \times$ denotes a random sampling from the generated data. $s_{mpd} = \times$ denotes s_{mpd} is not applied and we use all the samples selected by s_{cu} .

the generated pool (‘ R_1 ’-‘ R_3 ’) into the training set, with an equal volume of data, FVE selects effectively those generated videos that enhance VMR model training.

Table 5 shows an ablation study on the two hyperparameters for the number selected from: the harmonic score between cross-modal relevance and the uni-modal structure score (k) and the model performance disparity (l). We observe the best combination is $k=500$ and $l=100$. More ablation studies including the combination of s_c and s_u scores, the ablation of frame selection, the ablation of the location to replace the frames, and ablations on other datasets are presented in the Supplementary.

Action Editing

For action editing, we compare with Tune-A-Video (Wu et al. 2023), Video-P2P (Liu et al. 2023), Fatezero (Qi et al. 2023) and Dreamix (Molad et al. 2023). We collect a 10-video dataset with videos from Charades-STA and videos in the wild. We carry out comparisons on both quantitative and qualitative evaluations.

For quantitative evaluation, we evaluate the result using cross-modal relevance (s_c) and uni-modal structure scores (s_u). We define a successful generation as a video that optimizes both metrics quantified by their harmonic score s_{cu} .

Method	s_c	s_u	s_{cu}
Tune-A-Video	0.2910	0.4939	0.3641
Video-P2P	0.2895	0.5802	0.3862
Fatezero	0.2621	0.7061	0.3822
Dreamix	0.2672	0.6826	0.3841
FVE (Ours)	0.2722	0.7086	0.3933

Table 6: Quantitative comparisons. s_c denotes the cross-modal relevance score, s_u the uni-modal structure score and s_{cu} their harmonic score.



Figure 3: Qualitative comparisons. The first and last frames of the video are presented.

As shown in Table 6, previous methods either fail to maintain the video structure (TAV, Video-P2P) or show inferiority in cross-modal relevance (Dreamix, Fatezero), FVE demonstrates the best result with the harmonic score that combines the impact of both metrics. We present the qualitative evaluation of a single video due to space limitations. As depicted in Fig. 3, FVE generates videos that adeptly preserve instance appearance and generalise to new actions. More visualisations are presented in the Supplementary.

Conclusion

In this work, we addressed the problem of unseen novel semantic video moment retrieval (VMR) cross domains without having seen any target videos in training. Given the aim of learning a target domain by only text sentences describing new concepts, we proposed a Fine-grained Video Editing framework (FVE) to edit source videos automatically controlled by target sentences to simulate target domain training data. To control the generation process and select both visually plausible and semantically meaningful fine-grained video hypotheses for VMR training, we formulated an instance-preserving video diffusion model and a hybrid data selection strategy. Experimental results on three datasets demonstrated the effectiveness and generality of our method improving performance on VMR. Evaluation of video editing further demonstrated the ability of our method to change the action in a video and maintain the subject information. Future directions could explore long-temporal video editing conditioned on complex sentence prompts.

Acknowledgements

This work was partially supported by Adobe, Veritone, NSFC (62372014), CSC, and Queen Mary University of London’s Apocrita HPC facility from QMUL RESEARCH-IT.

References

- Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 22563–22575.
- Cai, W.; Huang, J.; and Gong, S. 2022. Hybrid-Learning Video Moment Retrieval across Multi-Domain Labels. *BMVC*.
- Cai, W.; Huang, J.; Hu, J.; Gong, S.; Jin, H.; and Liu, Y. 2024. Semantic Video Moment Retrieval by Temporal Feature Perturbation and Refinement. In *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*, 1–7. IEEE.
- Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.
- Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.
- Chen, H.; Zhang, Y.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023. DisenBooth: Disentangled Parameter-Efficient Tuning for Subject-Driven Text-to-Image Generation. *arXiv preprint arXiv:2305.03374*.
- Fan, H.; Li, Y.; Xiong, B.; Lo, W.-Y.; and Feichtenhofer, C. 2020. PySlowFast. <https://github.com/facebookresearch/slowfast>.
- Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*.
- Feng, R.; Weng, W.; Wang, Y.; Yuan, Y.; Bao, J.; Luo, C.; Chen, Z.; and Guo, B. 2023. CCEdit: Creative and Controllable Video Editing via Diffusion Models. *arXiv preprint arXiv:2309.16496*.
- Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*, 5267–5275.
- Gao, J.; Sun, X.; Xu, M.; Zhou, X.; and Ghanem, B. 2021. Relation-aware video reading comprehension for temporal language grounding. *arXiv preprint arXiv:2110.05717*.
- Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*.
- Hao, J.; Sun, H.; Ren, P.; Wang, J.; Qi, Q.; and Liao, J. 2022. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *European Conference on Computer Vision*, 130–147. Springer.
- Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.
- Jang, J.; Park, J.; Kim, J.; Kwon, H.; and Sohn, K. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, 13846–13856.
- Jeong, H.; and Ye, J. C. 2023. Ground-A-Video: Zero-shot Grounded Video Editing using Text-to-image Diffusion Models. *arXiv preprint arXiv:2310.01107*.
- Jiang, Y.; Wu, T.; Yang, S.; Si, C.; Lin, D.; Qiao, Y.; Loy, C. C.; and Liu, Z. 2024. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, 6689–6700.
- Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*, 1931–1941.
- Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 34: 11846–11858.
- Li, J.; Dongxu, L.; Silvio, S.; and Hoi, S. 2023. Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, 3032–3041.
- Liew, J. H.; Yan, H.; Zhang, J.; Xu, Z.; and Feng, J. 2023. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*.
- Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In *ICCV*, 2794–2804.
- Liu, B.; Yeung, S.; Chou, E.; Huang, D.-A.; Fei-Fei, L.; and Niebles, J. C. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*, 552–568.
- Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, 11235–11244.
- Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI*, volume 36, 1683–1691.
- Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; and Jia, J. 2023. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*.
- Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022b. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, 3042–3051.
- Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *AAAI*, volume 38, 3855–3863.
- Lu, T.; Zhang, X.; Gu, J.; Xu, H.; Pei, R.; Xu, S.; and Wu, Z. 2023. Fuse Your Latents: Video Editing with Multi-source Latent Diffusion Models. *arXiv preprint arXiv:2310.16400*.
- Luo, D.; Huang, J.; Gong, S.; Jin, H.; and Liu, Y. 2023. Towards Generalisable Video Moment Retrieval: Visual-Dynamic Injection to Image-Text Pre-Training. In *CVPR*, 23045–23055.

- Molad, E.; Horwitz, E.; Valevski, D.; Acha, A. R.; Matias, Y.; Pritch, Y.; Leviathan, Y.; and Hoshen, Y. 2023. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*.
- Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, 23023–23033.
- Mun, J.; Cho, M.; and Han, B. 2020. Local-global video-text interactions for temporal grounding. In *CVPR*, 10810–10819.
- Nam, J.; Ahn, D.; Kang, D.; Ha, S. J.; and Choi, J. 2021. Zero-shot natural language video localization. In *ICCV*, 1470–1479.
- Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.
- Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics*, 1: 25–36.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752*.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 22500–22510.
- Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2256–2265. PMLR.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 4489–4497.
- Wang, H.; Zha, Z.-J.; Li, L.; Liu, D.; and Luo, J. 2021. Structured multi-level interaction network for video moment localization via language query. In *CVPR*, 7026–7035.
- Wang, J.; Sun, A.; Zhang, H.; and Li, X. 2023a. MS-DETR: Natural Language Video Localization with Sampling Moment-Moment Interaction. *arXiv preprint arXiv:2305.18969*.
- Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023b. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.
- Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; et al. 2022. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. *arXiv preprint arXiv:2212.03191*.
- Wei, Y.; Zhang, S.; Qing, Z.; Yuan, H.; Liu, Z.; Liu, Y.; Zhang, Y.; Zhou, J.; and Shan, H. 2024. DreamVideo: Composing Your Dream Videos with Customized Subject and Motion. In *CVPR*.
- Wu, J.; Li, G.; Liu, S.; and Lin, L. 2020. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, volume 34, 12386–12393.
- Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 7623–7633.
- Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; and Li, X. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *CVPR*, 18709–18719.
- Xu, Y.; Sun, Y.; Li, Y.; Shi, Y.; Zhu, X.; and Du, S. 2023. Mh-detr: Video moment and highlight detection with cross-modal transformer. *arXiv preprint arXiv:2305.00355*.
- Yan, H.; Liew, J. H.; Mai, L.; Lin, S.; and Feng, J. 2023. MagicProp: Diffusion-based Video Editing via Motion-aware Appearance Propagation. *arXiv preprint arXiv:2309.00908*.
- Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. *arXiv preprint arXiv:2306.07954*.
- Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Parallel attention network with sequence matching for video grounding. *arXiv preprint arXiv:2105.08481*.
- Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based localizing network for natural language video localization. In *ACL*, 6543–6554. Online: Association for Computational Linguistics.
- Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, 12870–12877.
- Zhao, R.; Gu, Y.; Wu, J. Z.; Zhang, D. J.; Liu, J.; Wu, W.; Keppo, J.; and Shou, M. Z. 2023. MotionDirector: Motion Customization of Text-to-Video Diffusion Models. *arXiv preprint arXiv:2310.08465*.
- Zheng, M.; Cai, X.; Chen, Q.; Peng, Y.; and Liu, Y. 2024. Training-free video temporal grounding using large-scale pre-trained models. In *ECCV*, 20–37. Springer.
- Zheng, M.; Gong, S.; Jin, H.; Peng, Y.; and Liu, Y. 2023. Generating Structured Pseudo Labels for Noise-resistant Zero-shot Video Sentence Localization. In *ACL*, 14197–14209.