VE-Bench: Subjective-Aligned Benchmark Suite for Text-Driven Video Editing Quality Assessment

Shangkun Sun[1, 2], **Xiaoyu Liang**[1], **Songlin Fan**[1,2], **Wenxu Gao**[1,2], **Wei Gao**[1,2*]

[1]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, SECE, Shenzhen Graduate School, Peking University,
[2]Peng Cheng Laboratory
{sunshk, liangxiaoyu_yy, gaowx}@stu.pku.edu.cn, {slfan, gaowei262}@pku.edu.cn

Abstract

Text-driven video editing has recently experienced rapid development. Despite this, evaluating edited videos remains a considerable challenge. Current metrics tend to fail to align with human perceptions, and effective quantitative metrics for video editing are still notably absent. To address this, we introduce VE-Bench, a benchmark suite tailored to the assessment of text-driven video editing. This suite includes VE-Bench DB, a video quality assessment (VQA) database for video editing. VE-Bench DB encompasses a diverse set of source videos featuring various motions and subjects, along with multiple distinct editing prompts, editing results from 8 different models, and the corresponding Mean Opinion Scores (MOS) from 24 human annotators. Based on VE-Bench DB, we further propose VE-Bench QA, a quantitative human-aligned measurement for the text-driven video editing task. In addition to the aesthetic, distortion, and other visual quality indicators that traditional VQA methods emphasize, VE-Bench QA focuses on the text-video alignment and the relevance modeling between source and edited videos. It proposes a new assessment network for video editing that attains superior performance in alignment with human preferences. To the best of our knowledge, VE-Bench introduces the first quality assessment dataset for video editing and an effective subjective-aligned quantitative metric for this domain.

Code — https://github.com/littlespray/VE-Bench
Datasets — https://openi.pcl.ac.cn/OpenDatasets
Extended version — https://arxiv.org/pdf/2408.11481v2

Introduction

With the rise of the AI-Generated Content (AIGC) trend, an increasing number of text-driven video editing methods (Ma et al. 2024a,b, 2023, 2024c; Feng et al. 2024a) are gaining momentum and finding widespread application in daily life. However, there remains a lack of a suitable quantitative metric to assess video editing quality. Currently, the predominant evaluation method involves subjective experiments with human participants, which are costly and yield non-reusable results.

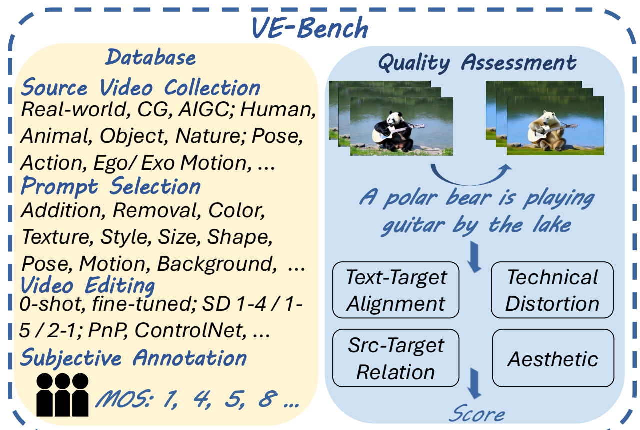


Figure 1: Overview of the proposed VE-Bench.

Currently, quantitative metrics such as CLIP and LPIPS scores (Radford et al. 2021; Zhang et al. 2018), FVD (Unterthiner et al. 2018), and Warp scores (Ceylan, Huang, and Mitra 2023; Kara et al. 2024) are commonly used in recent works. They mainly quantify results based on objective measurements in dimensions like editing quality, temporal consistency, and text consistency. Nevertheless, these existing metrics primarily face two main issues: *(1) The tendency of misalignment with human subjective perceptions. (2) Incomplete evaluation. As they mainly measure results from a single dimension, making it difficult to comprehensively reflect the overall quality of the effects.*

Traditional video quality assessment (VQA) methods (Wu et al. 2023a, 2022; Kou et al. 2023) have been able to align with human perceptions. However, these methods are primarily designed for natural videos and struggle to evaluate AIGC video editing tasks adequately. In this case, these methods overlook the different distortions in AIGC videos (Kou et al. 2024), such as irrational objects and irregular motion patterns. Besides, they do not simultaneously consider text-video alignment and the inner connection between source and edited videos, which are crucial for the assessment of editing results.

To address these challenges, we propose VE-Bench, a specialized suite tailored for text-driven AIGC video edit-

*Corresponding author.

ing, as shown in Figure 1. We first establish VE-Bench DB, a quality assessment database for video editing. VE-Bench DB collects source videos containing real-world videos, CG-rendered videos, and AIGC videos, with multiple different actions, subjects, and scenarios, along with various types of prompts and edited results from different video editing methods. We then assembled 24 human subjects to gather the Mean Opinion Scores (MOS) for each video. To the best of our knowledge, this is the first AIGC VQA dataset for evaluating the quality of edited videos.

Building on this foundation, we introduce VE-Bench QA, a novel multi-modal quality assessment network specifically designed for AIGC video editing. VE-Bench QA evaluates edited videos from various aspects such as source-target video relationship, text-video alignment, and other aspects such as aesthetics and distortion. Detailed experiments demonstrate that VE-Bench QA achieves state-of-the-art alignment with human preferences, surpassing existing advanced metrics and VQA methods.

Our contributions could be summarized as: (1) We collect VE-Bench DB, a diverse dataset with videos featuring various motions and subjects, along with multiple distinct editing prompts, and the corresponding editing results with rich human feedback. To the best of our knowledge, it is the first quality assessment dataset for text-driven video editing. (2) Unlike traditional VQA methods that focus mainly on visual quality indicators, we propose VE-Bench QA, which further emphasizes text-video alignment and relevance modeling between source and edited videos. (3) The proposed VE-Bench QA is tailored to the assessment of the text-driven video editing task, surpassing existing advanced evaluation methods in aligning with human subjective ratings and showcasing the effectiveness of VE-Bench QA in evaluating AIGC video editing quality.

Related Work

Metrics for Video Editing

Currently, metrics commonly used in text-driven video editing include some objective metrics (Radford et al. 2021; Zhang et al. 2018; Kara et al. 2024; Unterthiner et al. 2018), as well as some Video Quality Assessment (VQA) methods (Hartwig et al. 2024; Wu et al. 2023b, 2024a; Qu et al. 2024; Wu et al. 2023e; Xu et al. 2024) aligned with human feedback. CLIP (Radford et al. 2021) has been widely used due to its success in vision-language tasks (Wu et al. 2024b; Jia et al. 2024a,b). CLIP-T calculates the average cosine similarity between each edited frame and the corresponding textual prompt. CLIP-F(Tmp-Con), refers to the average cosine similarity between consecutive edited frames. Fram-Acc (Qi et al. 2023) represents the percentage of edited frames that has a higher similarity to the target text than to the original source text. LPIPS-P (Chai et al. 2023) and LPIPS-T (Chai et al. 2023) denote the LPIPS deviation from the original video frames and deviation between adjacent edited frames, respectively. FVD (Unterthiner et al. 2018) calculates the Fréchet Distance between two videos. Spat-Con (Yang et al. 2024) refers to the average distance between VGG features. OSV (Objective Semantic Variance), proposed by (Liu et al. 2024), uses DINO-ViT (Oquab et al. 2024) to measure semantic consistency and calculates the frame-wise feature variance in the edited region. Warp-MSE and Warp-SSIM (Kara et al. 2024) represent the MSE and SSIM between the edited video and the warped edited video by optical flow models. However, these individual metrics often only assess the editing results from a single dimension. Q_{edit} (Kara et al. 2024) multiplies Warp-SSIM with CLIP-T to obtain a more comprehensive evaluation. S_{edit} (Cong et al. 2024) (CLIP-T / Warp-MSE) combines Warp-MSE with CLIP-T to provide overall the assessment for videos. Nevertheless, these metrics are not aligned with human perceptions. PickScore (Kirstain et al. 2023) calculates the estimated alignment with human preferences via a CLIP-style model fine-tuned on human preference data. FastVQA (Wu et al. 2022) proposes grid mini-patch sampling to evaluate videos efficiently via the consideration of local quality. SimpleVQA (Sun et al. 2022) leverages quality-aware spatial features and motion features to assess videos. DOVER (Wu et al. 2023a) evaluates natural videos from the aesthetic and technical distortion perspective. However, these methods are typically suitable for evaluating single videos, neglecting the inherent relationship between edited results and the source video, and many VQA methods do not model the alignment between text and video. Currently, there is still a lack of a proper metric to evaluate the edited results based on the source video and editing prompts.

Datasets for Video Editing Assessment

In assessing edited videos, a common practice in prior works (Qi et al. 2023; Wu et al. 2023c; Yang et al. 2024; Sun et al. 2024a) has been to assemble human annotators to conduct subjective preion of ference experiments. However, the results of subjective preference experiments are difficult to reproduce, and there is considerable variance in the data and prompts selected when comparing different methods. Recently, some studies (Wu et al. 2023d; Feng et al. 2024b) have curated high-quality video-editing prompt pairs through diverse data collection and prompt design for unified community assessment. Nevertheless, these efforts still face two challenges: (1) These datasets do not include subjective experimental feedback (Mean Opinion Score, MOS) corresponding to the video data, requiring others to still use objective metrics or conduct subjective experiments again when utilizing these datasets. (2) The scenarios covered by these datasets could potentially be expanded further. For instance, TGVE (Wu et al. 2023d) has only collected 76 videos, and BalanceCC (Feng et al. 2024b) has gathered 100 videos, both of which mainly focus on real-world scenes. In this work, we introduce VE-Bench DB, a dataset for assessing video editing quality that includes diverse content and human feedback scores. This dataset encompasses a variety of categories including real-world scenes, CG-rendered scenes, and AIGC-generated scenes. It covers various subjects such as people performing different actions, occupations, genders, and ages, as well as different animals, objects, and landscapes. The dataset also includes multiple types of motion, such as ego-motion and exo-motion, along with various editing prompts. This provides a solid foundation for more robust video editing evaluation. Details on the dataset are provided in the following sections. In total, VE-Bench DB comprises 169 different videos edited using 8 different video editing methods, yielding 1,170 edited results after manual screening. After that, 24 human subjects are invited to obtain the corresponding MOS scores. To the best of our knowledge, it is the first VQA dataset for text-driven video editing.

Methods for Video Editing

Recently, with the fast development of diffusion models, lots of video editing methods have emerged (Wu et al. 2023c; Ma et al. 2022; Wang et al. 2024; Zhu et al. 2024; Chen et al. 2024). Different from image editing, one key to video editing is to maintain temporal consistency. Tune-a-video (Wu et al. 2023c) inflates the 2D convolutions in T2I models to pseudo-3D convolutions and fine-tunes the attention matrix with source videos. Text2Video-Zero (Khachatryan et al. 2023) leverages cross-frame attention and introduces latent motion dynamics to keep the global scene and the background consistent. FateZero (Qi et al. 2023) proposes to fuse the attention maps in the inversion process and generation process and utilizes the source prompt's cross-attention map to improve consistency. ControlVideo (Zhang et al. 2024) introduces the interleaved-frame smoother and full-frame attention to keep the temporal consistency and model the temporal relationship in different frames. Flatten (Cong et al. 2024) utilizes optical-flow-guided attention during the diffusion process to improve the visual consistency, which models the similarity of patch trajectories in different clips. RAVE (Kara et al. 2024) leverages the full-frame attention and proposes the grid sampling strategy to maintain the temporal consistency in video sequences. Based on Rerender-a-video (Yang et al. 2023) and the flow-guided attention in (Cong et al. 2024), Fresco (Yang et al. 2024) further develops a group of temporal attention including efficient cross-frame attention, spatial/temporal-guided attention, etc., and fine-tunes the translated feature for better temporal consistency. In this work, we selected a variety of methods for video editing, including early and recent approaches, 0-shot and few-shot techniques, different Stable Diffusion (SD) base models, and various editing methodologies, to ensure the diversity of the generated results.

VE-Bench DB: Subjective-Aligned Dataset for Text-Driven Video Editing

Source Video Collection

To support a more robust quality assessment for video editing, VE-Bench collected a diverse set of source videos that are not limited to real-world scenes but also include some content rendered by computer graphics and text-driven AIGC videos. Notably, considering the current wide range of use cases, real-world scenes still account for a larger proportion. Different from previous works, given concerns about copyright issues, watermarks, and resolution, we did not randomly sample videos from Webvid (Bain et al. 2021). Instead, to cover as many different content subjects, action categories, and scenarios as possible, VE-Bench manually selected 123 videos from four datasets: DAVIS (Pont-Tuset et al. 2017), Kinetics-700 (Kay et al. 2017), Sintel (Butler et al. 2012), and Spring (Mehl et al. 2023). Spring and Sintel are high-resolution datasets rendered by computer graphics. To ensure a diversity of actions and rich content, we did not randomly sample but carefully handpicked the corresponding video content. In addition, we selected 15 different videos from Sora (Brooks et al. 2024) and Kling (Kuaishou 2024) based on the principle of diversity in motion and content. Furthermore, to cover more content and a variety of actions, we selected 31 videos from the internet with the appropriate permissions. For the detailed data collection process, please refer to (Sun et al. 2024b). Ultimately, we collected 169 source videos with diverse content, and their specific sources, contents, and category compositions are shown in Figure 2. All selected videos were resized to have a long side of 768 pixels while maintaining their original aspect ratios. Considering the limited length supported by existing video editing methods, each video was trimmed to 32 frames.

Prompt Selection

Referring to past work (Huang et al. 2024), we classify prompts used for video editing into three major categories: (1) Style editing, which includes the edit on color, texture, or the overall atmosphere. (2) Semantic editing, which includes background editing and local editing such as the addition, replacement, or removal on a certain object. (3) Structural editing, which includes the change in object size, pose, motion, etc. To ensure the specificity and diversity of the prompts, we manually crafted corresponding prompts for each video, and the specific distribution is shown in Figure 3.

Video Editing

We then select 8 video editing methods. To ensure the distribution of edited video quality, in addition to recent top-performance models, we also include some earlier video editing methods. Besides, we select methods with different base models ranging from SD 1-4 to SD2-1 to improve the diversity of edited results. Furthermore, to ensure the diversity of edited content, we choose both 0-shot methods and methods that require fine-tuning. We also select models based on different editing paradigms, including effective editing strategies such as Instruct P2P (Brooks, Holynski, and Efros 2023), PnP (Tumanyan et al. 2023), ControlNet (Zhang, Rao, and Agrawala 2023), etc. The specific details are presented in Table 1.

Subjective Study

According to the ITU standard (Series 2002), the number of participants in subjective experiments should be at least 15 to ensure that the results' variance is within a controllable range. For this experiment, a total of 24 human subjects with diverse backgrounds were recruited. During the experiment, the subjects were asked to consider their subjective impressions and evaluate the text-video consistency, source-target fidelity, and quality of the edited videos in a comprehensive manner. The text-video consistency refers to whether the edited content adheres to the given prompt. The

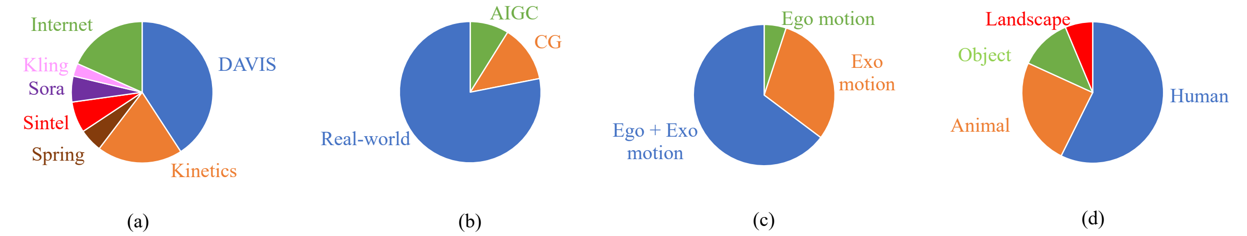


Figure 2: Collection of source videos. (a) Sources of videos. (b) Types of videos. (c) Motion categories. (d) Content categories.

Model	Time	0-shot	Edit.	SD.
Tune-a-video (Wu et al. 2023c)	ICCV'23	✗	Others	1-4
T2V-Zero (Khachatryan et al. 2023)	ICCV'23	✓	Instruct-P2P (Brooks, Holynski, and Efros 2023)	1-5
Fate-Zero (Qi et al. 2023)	ICCV'23	✓	Others	1-4
ControlVideo (Zhang et al. 2024)	ICLR'24	✓	ControlNet (Zhang, Rao, and Agrawala 2023)	1-5
TokenFlow (Geyer et al. 2024)	ICLR'24	✓	PnP (Tumanyan et al. 2023)	2-1
Flatten (Cong et al. 2024)	ICLR'24	✓	Others	2-1
RAVE (Kara et al. 2024)	CVPR'24	✓	Others	1-5
Fresco (Yang et al. 2024)	CVPR'24	✓	ControlNet (Zhang, Rao, and Agrawala 2023)	1-5

Table 1: Collection of the editing models.

source-target fidelity indicates the degree to which the original video and the edited video maintain a certain level of connection. The edited video quality can be assessed from aspects such as temporal and spatial coherence, aesthetics, and technical distortions. When evaluating, all participants rated the videos on a scale from 1 to 10. More information about the annotators is shown in (Sun et al. 2024b). Following previous works (Wu et al. 2023a; Kou et al. 2024; Chai et al. 2023), Z-scores are used to normalize the raw MOS values, which could be formulated as After collecting all raw Mean Opinion Score (MOS) values, we use the Z-score normalization method to eliminate inter-subject differences, which could be formulated as:

$$Z_{m,i} = \frac{X_{m,i} - \mu(X_i)}{\sigma(X_i)}, \tag{1}$$

where $X_{m,i}$ and $Z_{m,i}$ refer to the raw MOS and Z-score of m-th video from i-th participant, respectively. $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation operators, respectively, and X_i is the collection of all MOS from the i-th annotator. Then we apply the screening method in BT.500 (Int.Telecommun.Union 2000) to filter the outliers. The difference between the raw scores and the normalized scores is illustrated in Figure 4.

Dataset Analysis

We further conducted a detailed analysis of the data in the VE-Bench DB and investigated the scores for the editing results of each model on the collected videos, as shown in

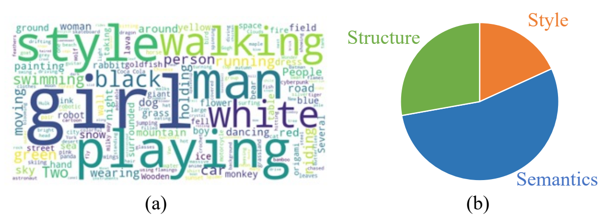
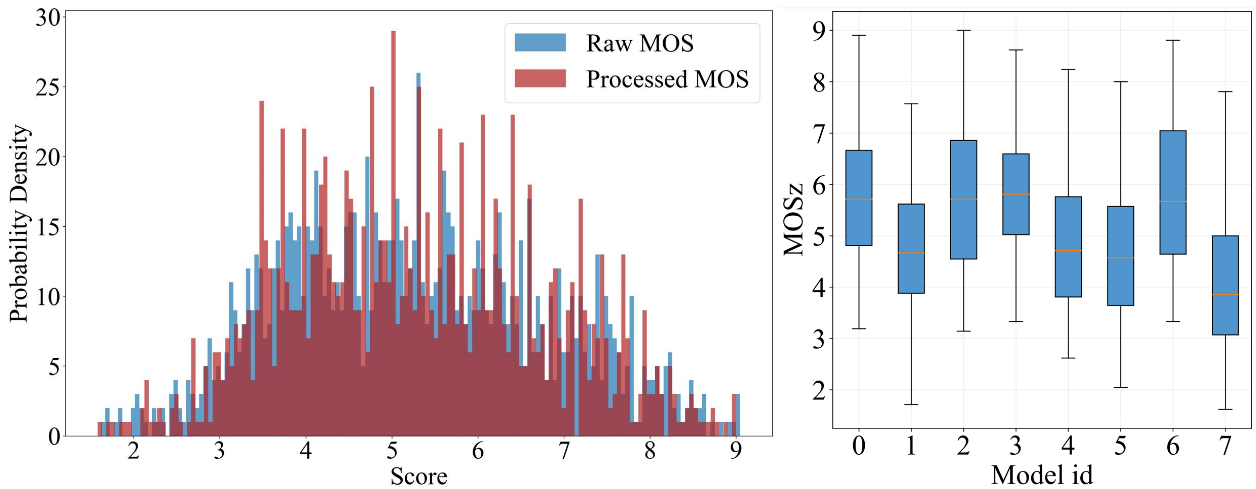


Figure 3: Statistics of VE-Bench DB prompts. (a) Word cloud of VE-Bench DB prompts. (b) Proportion of different types

Fig. 5. We can see that different models have varying capabilities across different types of edits. Furthermore, for stylization prompts, current editing methods generally perform better, whereas for tasks such as changing size or shape, the models' performance tends to degrade significantly. Notably, FateZero (Qi et al. 2023) demonstrates notably superior performance for this type of prompt. Additionally, compared to removal commands, models generally perform better with addition commands, suggesting that removing objects from a video tends to be harder than adding. A more detailed analysis and impacts of motion degree, fps, etc. are shown in (Sun et al. 2024b).

VE-Bench QA: Subjective-Aligned Metric for Text-Driven Video Editing

Based on the VE-Bench DB, we further developed the VE-Bench Quality Assessment (VE-Bench QA) network, which



(a) (b)

Figure 4: Statistics on MOS. (a) The distribution of the raw/Z-score MOS. (b) Z-score MOS distributions of 8 editing methods.

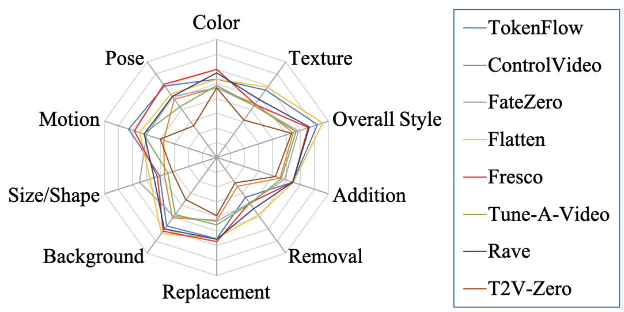


Figure 5: Model performance on different types of prompts.

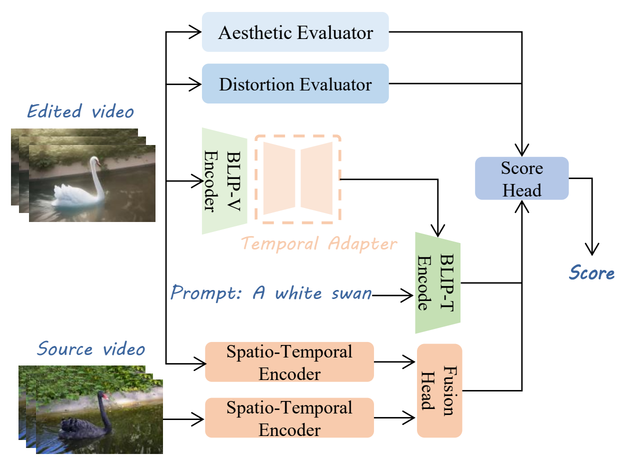


Figure 6: Network architecture of VE-Bench QA.

aligns with human subjective perceptions for evaluating the quality of edited videos, as illustrated in Fig. 6. The VE-Bench QA evaluates the quality of edited videos from three aspects: (1) Alignment between the edited video and the prompt. (2) Relevance between the edited video and the original video. (3) Quality of the edited video. We will elaborate on each component in the following sections.

Video-Text Alignment

Traditional natural video evaluation methods do not need to consider the alignment between the video and text prompt, which is one reason why they tend to fail when directly applied to AIGC video quality assessment (VQA). Therefore, based on the successful VQA method (Wu et al. 2023a), we incorporate the text branch to model the alignment between the generated content and the corresponding text. Inspired by BLIP (Li et al. 2022), we design an effective temporal adapter to extend it to the temporal dimension, as shown in 6, which could be formulated as,

$$e_{bv} = \mathbf{F_{bv}}(V^\star), \tag{2}$$
$$t_{bv} = \mathbf{F_{ta}}(e_{bv}), \tag{3}$$
$$e_{bt} = \mathbf{F_{bt}}(p, \mathbf{F_{ca}}(t_{bv})), \tag{4}$$

where $\mathbf{F_{bv}}$, $\mathbf{F_{bt}}$ refer to the BLIP visual and text encoder, and $\mathbf{F_{ta}}$ represents the temporal adapter. p is the prompt. The derived spatio-temporal feature t_{bv} is then interacted with the text encoder via cross-attention (denoted as $\mathbf{F_{ca}}$).

Source-Target Relationship

Measuring the consistency between the src and dst videos is challenging. There is inherently a connection between the src and dst videos, but there are also significant differences in the pixel space. Therefore, directly using methods like MSE for RGB space measurements has certain limitations. We design an effective spatiotemporal extractor to project the src-dst videos into a latent space. After concatenating them along the dimension, we obtain a reasonable score estimate through a FFN which could be formulated as,

$$f = \mathbf{F}(V), \quad (5)$$

$$f^{\star} = \mathbf{F}^{\star}(V^{\star}), \quad (6)$$

$$o_s = \mathbf{H_s}(Concat(f, f^{\star})), \quad (7)$$

where V, $V^{\star}$ denote the original and edited videos, respectively. o_s is the output vector measuring the relevance between source and edited videos. $\mathbf{H_s}$ is the lightweight feed-forward network. $\mathbf{F}$ and $\mathbf{F}^{\star}$ denote the spatio-temporal encoder for source and target, respectively. In practice, we try different spatio-temporal backbones (Liu et al. 2022b; Meng et al. 2022) and finally choose the Uniformer (Li et al. 2023).

Visual Quality

To assess the quality of the edited video, we start from the perspectives used in the previous top-performance method DOVER (Wu et al. 2023a), which evaluates videos based on aesthetics and technical distortion. DOVER achieves success in some natural video quality assessment datasets such as (Ying et al. 2021; Sinno and Bovik 2018). In practice, the measurement of aesthetics is implemented via the inflated ConvNext (Liu et al. 2022a) pre-trained on AVA (Murray, Marchesotti, and Perronnin 2012), and the distortion is assessed with the Video-Swin (Liu et al. 2022b) backbone pre-trained with GRPB (Wu et al. 2022). At the first stage of training, the backbones from DOVER are frozen and only the parameters of the regression head are updated. In the second stage, all parameters of the visual quality branches (namely, the aesthetic and technical branch) are updated.

Supervision

Following previous works (Wu et al. 2023a, 2022; Qu et al. 2024), we adopt the combination of PLCC (Pearson Linear Correlation Coefficient) loss and rank loss (Gao et al. 2019) with the weight of α as the total loss for all branches of the overall network, which could be formulated as follows.

$$L = L_{plcc} + \alpha \cdot L_{rank}, \quad (8)$$

where α is set to 0.3 in practice.

Experiments

Implementation Details

We build all models via PyTorch and train them via NVIDIA V100 GPUs. Following the 10-fold method (Kou et al. 2023; Wu et al. 2023a; Sun et al. 2022), all models are trained with the initial learning rate of $1e-3$ and the batch size of 8 on VE-Bench DB for 60 epochs. Following DOVER (Wu et al. 2023a), we first fine-tuning the head for 40 epochs with linear probing, and then train all parameters for another 20 epochs. Adam (Kingma and Ba 2014) optimizer and a cosine scheduler are applied during training. Following previous works (Wu et al. 2023a), the aesthetic and technical branches of the evaluator are initialized with pre-trained ConvNext (Liu et al. 2022a) and VideoSwin-Tiny (Liu et al. 2022b) with GRPB (Wu et al. 2022).

Evaluation Metrics

Following previous works (Wu et al. 2023a; Kou et al. 2024, 2023; Sun et al. 2022), we use four metrics as our evaluation metrics: Spearman's Rank Order Correlation Coefficient (SROCC), Pearson's Linear Correlation Coefficient (PLCC), Kendall rank-order correlation coefficient (KRCC), and Root Mean Square Error (RMSE).

Quantitative Results

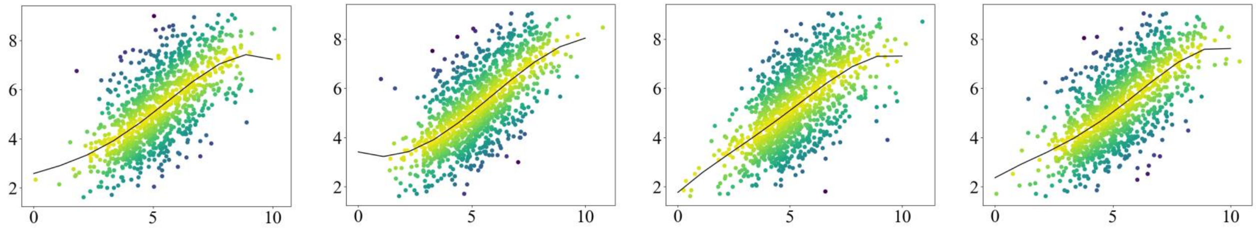
We compare our results with advanced evaluation metrics in video editing, including objective metrics (Radford et al. 2021; Cong et al. 2024; Kirstain et al. 2023) and state-of-the-art human-aligned Video Quality Assessment (VQA) methods (Kou et al. 2023; Wu et al. 2022, 2023a). The results are shown in Table 2. We further collected a validation set generated by other models and tested on more datasets (Kou et al. 2024), which are shown in (Sun et al. 2024b). From these, it can be seen that compared to previous traditional VQA methods (Wu et al. 2023a, 2022; Kou et al. 2023) and commonly used objective metrics, VE-Bench QA achieves significantly superior performance in aligning with human subjective perception, surpassing the second by 7.64%, 8.06%, 8.85%, 13.2% in SROCC, PLCC, KLCC, and RMSE, respectively. Compared with our baseline method DOVER, VE-Bench attains higher performance gain. Compared with learning-based methods, the performances of 0-shot objective measurements are relatively low. Similar situations are quite common in previous works such as (Kou et al. 2024; Qu et al. 2024), as the objective quantitative metric struggles to align with human perceptions.

Qualitative Results

We also plot the difference between the predicted scores after training and the MOS scores, as illustrated in Figure 7. The curves are obtained by a four-order polynomial nonlinear fitting. As the brightness of scatter points grows from low to high, the density goes from low to high. From there, it can be intuitively seen that VE-Bench QA has prediction results more aligned with human perception. We further conducted a qualitative comparison for different score levels in VE-Bench, as illustrated in the supplements, where we present several video examples in VE-Bench DB with varying MOS.

Ablation Study

As shown in Table 3, the settings we adopted in our final model are underlined. We first explore different ways of video-text alignment. Here, we experimented with CLIP and fine-tuned the regression head composed of Feed-Forward Networks, which learn the alignment from the cosine similarity of its visual and text Backbone outputs. Experiments demonstrate that, although both CLIP and BLIP possess rich vision-language prior knowledge and enhance network performance, BLIP achieved more effective improvements. Furthermore, we explored how to effectively model the relevance between the source video and the edited video. We attempted to efficiently extract video features and, through experiments conducted on methods such as Video SwinTransformer (Liu et al. 2022b) (VSwin), Masked Video Distillation (Wang et al. 2023) (MVD), and Uniformer (Li et al.



(a) FastVQA (b) StableVQA (c) DOVER (d) VE-Bench QA

Figure 7: Plots of predicted vs. GT scores. The brightness of scatter points from dark to bright means density from low to high.

Type	Models	VE-Bench DB 10-fold			
		SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
0-shot	CLIP-F (Radford et al. 2021)	0.2284	0.1860	0.1545	4.448
	S_{edit} (Cong et al. 2024)	0.1686	0.1865	0.1135	3.981
	PickScore (Kirstain et al. 2023)	0.2266	0.2446	0.1540	1.786
Fine-tuned	FastVQA (Wu et al. 2022)	0.6333	0.6326	0.4545	1.312
	StableVQA (Kou et al. 2023)	0.6889	0.6783	0.4974	1.262
	DOVER (Wu et al. 2023a)	0.6119	0.6295	0.4354	1.311
Ours	VE-Bench QA	**0.7415**	**0.7330**	**0.5414**	**1.095**

Table 2: Comparison of different methods with VE-Bench QA.

Experiment	Method	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
Baseline	DOVER	0.6119	0.6295	0.4354	1.311
Text	CLIP	0.6379	0.6529	0.4560	1.269
	BLIP	**0.7171**	**0.7094**	**0.5193**	**1.146**
Temporal	None	0.7171	0.7094	0.5193	1.146
	VSwin	0.7187	0.7101	0.5197	1.143
	MVD	0.7228	0.7143	0.5240	1.134
	Uformer	**0.7317**	**0.7252**	**0.5328**	**1.116**
Fusion	None	0.7317	0.7252	0.5328	1.116
	MCA	0.7330	0.7255	0.5341	1.113
	Concat	**0.7415**	**0.7330**	**0.5414**	**1.095**
Param.	w/o	0.7251	0.7174	0.5262	1.130
	w	**0.7415**	**0.7330**	**0.5414**	**1.095**

Table 3: Ablation study of the proposed VE-Bench QA.

2023) (Uformer), we identified suitable feature extractors. Additionally, we explored effective ways to fuse features from the source video and the edited video, as presented in Table 3. MCA denotes the mutli-head cross-attention. We found that concatenation along the dimension is a simple and effective design for the assessment. We further ablate the effect of additional parameter, which demonstrates the improvements are not from more parameters.

Conclusion

In this work, we introduce VE-Bench DB, a subjective-aligned dataset specifically designed for evaluating text-driven video editing, and VE-Bench QA, a novel human-aligned metric for assessing the effects of text-driven video editing. To our knowledge, VE-Bench DB is the first VQA dataset tailored for text-driven video editing. Compared to previous metrics commonly used, VE-Bench QA achieves significantly better alignment with human perceptions.

Acknowledgments

This work was supported by National Science and Technology Major Project (2024ZD01NL00101), Natural Science Foundation of China (62271013, 62031013), Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (2024B1212010006), Guangdong Province Pearl River Talent Program (2021QN020708), Guangdong Basic and Applied Basic Research Foundation (2024A1515010155), Shenzhen Science and Technology Program (JCYJ20240813160202004, JCYJ20230807120808017).

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Butler, D. J.; Wulff, J.; Stanley, G. B.; and Black, M. J. 2012. A naturalistic open source movie for optical flow evaluation. In A.

Fitzgibbon et al. (Eds.), ed., *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, 611–625. Springer-Verlag.

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23206–23217.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23040–23050.

Chen, Q.; Ma, Y.; Wang, H.; Yuan, J.; Zhao, W.; Tian, Q.; Wang, H.; Min, S.; Chen, Q.; and Liu, W. 2024. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*.

Cong, Y.; Xu, M.; Chen, S.; Ren, J.; Xie, Y.; Perez-Rua, J.-M.; Rosenhahn, B.; Xiang, T.; He, S.; et al. 2024. FLATTEN: optical FLow-guided ATTENtion for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations*.

Feng, K.; Ma, Y.; Wang, B.; Qi, C.; Chen, H.; Chen, Q.; and Wang, Z. 2024a. Dit4edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286*.

Feng, R.; Weng, W.; Wang, Y.; Yuan, Y.; Bao, J.; Luo, C.; Chen, Z.; and Guo, B. 2024b. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6712–6722.

Gao, F.; Tao, D.; Gao, X.; and Li, X. 2019. Learning to Rank for Blind Image Quality Assessment. arXiv:1309.0213.

Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2024. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *The Twelfth International Conference on Learning Representations*.

Hartwig, S.; Engel, D.; Sick, L.; Kniesel, H.; Payer, T.; Ropinski, T.; et al. 2024. Evaluating Text to Image Synthesis: Survey and Taxonomy of Image Quality Metrics. *arXiv preprint arXiv:2403.11821*.

Huang, Y.; Huang, J.; Liu, Y.; Yan, M.; Lv, J.; Liu, J.; Xiong, W.; Zhang, H.; Chen, S.; and Cao, L. 2024. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*.

Int.Telecommun.Union. 2000. Methodology for the Subjective Assessment of the Quality of Television Pictures ITU-R Recommendation. *Tech. Rep.*

Jia, M.; Zhao, L.; Li, G.; and Zheng, Y. 2024a. ContextHOI: Spatial Context Learning for Human-Object Interaction Detection. arXiv:2412.09050.

Jia, M.; Zhao, L.; Li, G.; and Zheng, Y. 2024b. Orchestrating the Symphony of Prompt Distribution Learning for Human-Object Interaction Detection. arXiv:2412.08506.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2024. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6507–6516.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15954–15964.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Kou, T.; Liu, X.; Sun, W.; Jia, J.; Min, X.; Zhai, G.; and Liu, N. 2023. Stablevqa: A deep no-reference quality assessment model for video stability. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1066–1076.

Kou, T.; Liu, X.; Zhang, Z.; Li, C.; Wu, H.; Min, X.; Zhai, G.; and Liu, N. 2024. Subjective-Aligned Dateset and Metric for Text-to-Video Quality Assessment. *arXiv preprint arXiv:2403.11956*.

Kuaishou. 2024. Kling. URL:https://kling.kuaishou.com/. Accessed: 2024-08-10.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; and Qiao, Y. 2023. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12581–12600.

Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; and Jia, J. 2024. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8599–8608.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022a. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022b. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024a. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; et al. 2024b. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. *arXiv preprint arXiv:2403.08268*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024c. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.

Mehl, L.; Schmalfuss, J.; Jahedi, A.; Nalivayko, Y.; and Bruhn, A. 2023. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4981–4991.

Meng, L.; Li, H.; Chen, B.-C.; Lan, S.; Wu, Z.; Jiang, Y.-G.; and Lim, S.-N. 2022. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12309–12318.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, 2408–2415.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15932–15942.

Qu, B.; Liang, X.; Sun, S.; and Gao, W. 2024. Exploring aigc video quality: A focus on visual harmony, video-text consistency and domain distribution gap. *arXiv preprint arXiv:2404.13573*.

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 *International conference on machine learning*, 8748–8763. PMLR.

Series, B. 2002. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, 500(13).

Sinno, Z.; and Bovik, A. C. 2018. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2): 612–627.

Sun, K.; Huang, K.; Liu, X.; Wu, Y.; Xu, Z.; Li, Z.; and Liu, X. 2024a. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*.

Sun, S.; Liang, X.; Fan, S.; Gao, W.; and Gao, W. 2024b. VE-Bench: Subjective-Aligned Benchmark Suite for Text-Driven Video Editing Quality Assessment. *arXiv preprint arXiv:2408.11481*.

Sun, W.; Min, X.; Lu, W.; and Zhai, G. 2022. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 856–865.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024. COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. *arXiv preprint arXiv:2406.08850*.

Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Yuan, L.; and Jiang, Y.-G. 2023. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6312–6322.

Wu, H.; Chen, C.; Hou, J.; Liao, L.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2022. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, 538–554. Springer.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023a. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20144–20154.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; and Lin, W. 2024a. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. In *ICLR*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023b. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023c. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Wu, J. Z.; Li, X.; Gao, D.; Dong, Z.; Bai, J.; Singh, A.; Xiang, X.; Li, Y.; Huang, Z.; Sun, Y.; et al. 2023d. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023e. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Wu, Z.; Weng, Z.; Peng, W.; Yang, X.; Li, A.; Davis, L. S.; and Jiang, Y.-G. 2024b. Building an open-vocabulary video CLIP model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, 1–11.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2024. FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8703–8712.

Ying, Z.; Mandal, M.; Ghadiyaram, D.; and Bovik, A. 2021. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14019–14029.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Wei, Y.; Jiang, D.; ZHANG, X.; Zuo, W.; and Tian, Q. 2024. ControlVideo: Training-free Controllable Text-to-video Generation. In *The Twelfth International Conference on Learning Representations*.

Zhu, C.; Li, K.; Ma, Y.; Tang, L.; Fang, C.; Chen, C.; Chen, Q.; and Li, X. 2024. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*.