Adaptive-Guided Latent Diffusion for Video Counterfactual Explanations with Multi-Scale Perceptual Refinement

Yucan Ping and Haoxiang Wen *

Posted Date: 3 February 2026

doi: 10.20944/preprints202602.0181.v1

Keywords: video counterfactual explanations; latent diffusion models; model interpretability; explanation quality; video understanding

 doi:10.20944/preprints202602.0181.v1

Article

Adaptive-Guided Latent Diffusion for Video Counterfactual Explanations with Multi-Scale Perceptual Refinement

Yucan Ping and Haoxiang Wen *

Kunming University of Science and Technology
* Correspondence: 202360304721@stu.kust.edu.cn

Abstract

The increasing reliance on deep learning models in video understanding necessitates transparent and interpretable decision-making, with video counterfactual explanations (CEs) offering a critical avenue to understand model behavior. However, generating effective video CEs remains challenging due to video data's high dimensionality, temporal coherence demands, and the need to balance precise alterations with visual realism. Existing Latent Diffusion Model (LDM)-based CE methods often struggle with generation efficiency, exact target adherence, and the suppression of subtle visual artifacts. To address these limitations, we propose an Adaptive-Guided Latent Diffusion Model for Counterfactual Explanations (AG-LDM-CE). Our framework introduces an Adaptive Gradient Guidance (AGG) mechanism that dynamically adjusts guidance strength based on proximity to the target prediction, optimizing efficiency and balancing target adherence with visual fidelity. Complementing this, a novel Multi-Scale Perceptual Refinement (MSPR) module leverages multi-level VAE features to intelligently suppress artifacts and ensure that counterfactual changes are accurately localized to causally relevant regions. Extensive evaluations across diverse video regression (EchoNet-Dynamic) and classification (FERV39K, Something-Something V2) tasks demonstrate AG-LDM-CE's superior performance. Our method significantly improves generation efficiency and explanation quality, achieving strong target adherence and perceptual realism. Ablation studies and human evaluations further validate the significant contributions of AGG and MSPR, confirming that AG-LDM-CE generates more efficient, accurate, perceptually realistic, and precisely localized video counterfactual explanations.

Keywords: video counterfactual explanations; latent diffusion models; model interpretability; explanation quality; video understanding

1. Introduction

The rapid advancements in artificial intelligence models across various video understanding tasks, such as action recognition, emotion analysis, and medical diagnosis [1], have underscored the critical need for increased transparency and interpretability in their decision-making processes. Recent progress in multimodal retrieval for Visual Question Answering (VQA) using Large Language Models (LLMs) [2], methods for teaching Vision-Language Models (VLMs) to resolve ambiguity in visual questions [3], and techniques to enhance visual reflection in VLMs [4] further emphasize the growing complexity and capabilities of these models, making interpretability paramount. *Video Counterfactual Explanations (CEs)* emerge as an intuitive and user-friendly interpretability method, designed to elucidate "why a model made a particular prediction" and "how the input could be minimally altered to change that prediction" [5]. This understanding is pivotal for deciphering model decision boundaries, diagnosing inherent biases, and providing actionable guidance to users on how to adjust their behavior or inputs to achieve desired outcomes [6].

However, the generation of effective Video CEs presents unique challenges stemming from the high-dimensional nature of video data, the stringent requirements for temporal coherence, and

 doi:10.20944/preprints202602.0181.v1

the delicate balance between achieving precise counterfactual alterations and maintaining visual realism. Existing methods, particularly those leveraging Latent Diffusion Models (LDMs), have shown substantial promise in generating high-quality and effective Video CEs [7]. While these approaches typically operate by performing reverse diffusion in the video's latent space, guided by gradients from the black-box model to steer the generation towards a target prediction, recent work on component-controllable personalization in text-to-image diffusion models [8], personalized martial arts combat video generation [9], and various forms of facial manipulation via diffusion [10,11] illustrates the increasing demand for fine-grained control and realism in generative tasks. Despite these successes, current LDM-based counterfactual explanation (LDM-CE) methods still exhibit limitations in terms of generation efficiency, the precision and locality of counterfactual changes, and the effective suppression of undesirable latent artifacts [12].

Motivated by these challenges, we propose a novel Adaptive-Guided Latent Diffusion Model for Counterfactual Explanations (AG-LDM-CE). Our primary objective is to build upon the strengths of existing LDM-CE frameworks while significantly enhancing the efficiency of counterfactual video generation, improving the adherence to the black-box model's target prediction, and substantially elevating both the visual realism and the accuracy of localized changes within the generated explanations.

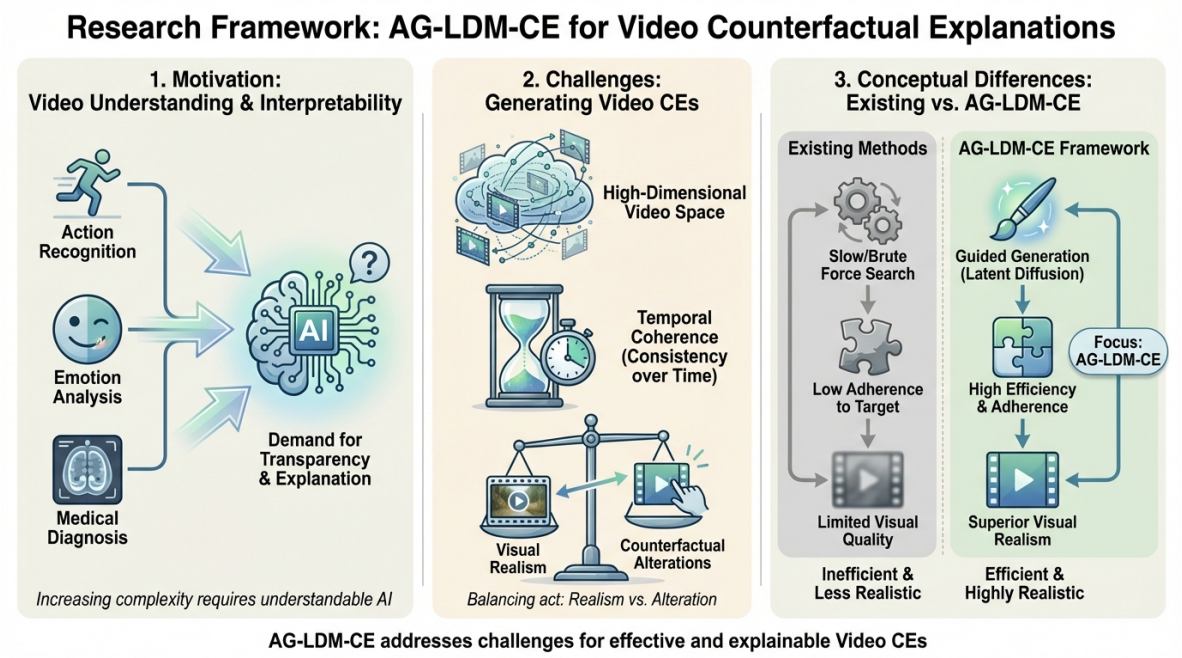

Figure 1. The research framework of AG-LDM-CE for Video Counterfactual Explanations. Panel 1 outlines the motivation for interpretability in video understanding tasks like action recognition, emotion analysis, and medical diagnosis. Panel 2 details the challenges inherent in generating effective video CEs, including high-dimensional video space, temporal coherence, and balancing realism with alteration. Panel 3 illustrates the conceptual differences between existing LDM-CE methods and our proposed AG-LDM-CE framework, highlighting AG-LDM-CE's advantages in efficiency, target adherence, and superior visual realism.

Our AG-LDM-CE framework employs an optimized latent diffusion backbone, complemented by two key innovations: an Adaptive Gradient Guidance (AGG) mechanism and a Multi-Scale Perceptual Refinement (MSPR) module. The AGG dynamically adjusts the gradient guidance strength during the reverse diffusion process, ensuring a more balanced trade-off between target adherence and visual fidelity, while the MSPR module leverages multi-level VAE features to generate a precise refinement mask, effectively suppressing artifacts and ensuring that changes are perceptually and semantically meaningful.

We conduct comprehensive evaluations of AG-LDM-CE across diverse video understanding tasks and datasets. Specifically, we test our method on: (1) EchoNet-Dynamic, a cardiac ultrasound dataset, for Left Ventricular Ejection Fraction (LVEF) regression tasks using a 3D ResNet variant

 doi:10.20944/preprints202602.0181.v1

[13]; (2) FERV39K, a facial expression video dataset, for 7-class emotion classification using the S2D (Static-to-Dynamic) model [14]; and (3) Something-Something V2 (SSv2), a large-scale human action recognition dataset, for 174-class action classification using VideoMAE-Base [15]. Our experiments demonstrate that AG-LDM-CE consistently outperforms existing state-of-the-art LDM-CE methods. For instance, on the EchoNet-Dynamic regression task, AG-LDM-CE achieves an R^2 of 1.00 with a MAE of 0.45, while reducing generation time to 19 seconds, surpassing baseline methods in both accuracy and efficiency. On the FERV39K classification task, it yields a classification flip rate (FR) of 0.989 with improved FID and FVD scores (3.950 and 28.500 respectively), indicating superior realism and diversity of generated explanations. Qualitative analysis on SSv2 further showcases its ability to generate plausible and localized counterfactual changes for complex actions.

In summary, our main contributions are as follows:

- We propose an Adaptive Gradient Guidance (AGG) mechanism that dynamically adjusts guidance strength based on proximity to the target prediction, leading to more efficient, precise, and visually coherent counterfactual video generation.
- We introduce a novel Multi-Scale Perceptual Refinement (MSPR) module that leverages multi-level VAE features to intelligently suppress artifacts and ensure that counterfactual changes are accurately localized to causally relevant regions, significantly enhancing visual realism and interpretability.
- We conduct extensive experiments across challenging video regression (EchoNet-Dynamic) and classification (FERV39K, SSv2) tasks, demonstrating that AG-LDM-CE achieves superior performance in terms of generation efficiency, model target adherence, visual quality, and fidelity compared to existing advanced LDM-CE methods.

2. Related Work

2.1. Explainable AI and Counterfactual Explanations

The widespread adoption of complex ML models in diverse fields, from logistics [16] and procurement [17] to SME growth [18] and probabilistic similarity [19], necessitates Explainable AI (XAI) for transparency and trust. XAI explains *what*, *why*, and *what changes* for different outcomes. Early efforts focused on justifying outputs in critical contexts like court judgment prediction [20], face anti-spoofing [21], and LLM-based clinical decision support [22]. This need is especially critical in medical diagnosis, such as diabetic retinopathy risk stratification [1] and multi-omics for diseases like optic neuritis [23] and myopia [24].

Counterfactual explanations (CEs) are intuitive and actionable, showing minimal input changes to alter a model's prediction. LLMs generate effective contrastive explanations for commonsense reasoning, improving performance and human judgment [25]. Polyjuice [26] is an NLP counterfactual generator, and counterfactual inference debiases text classification [27]. Robust CEs rely on causal understanding, as seen in event causality identification [28]. Explanations extend to multimodal data, like context-aware video explanations for VQA [29]. Contrastive explanations also enhance LLM in-context learning by guiding decision boundaries [30]. Recent works utilize LLMs for visual entity discernment in multimodal retrieval [2], teaching VLMs to resolve ambiguities [3], and enhancing visual reflection [4]. These efforts highlight the diverse XAI landscape, emphasizing CEs across modalities for interpretable and responsible AI.

2.2. Generative Models for Video and Guided Diffusion

Generative models have revolutionized content creation, particularly high-fidelity video generation and controllable synthesis via guided diffusion. Foundational works include BARThez [31] for NLG and overviews of DDPMs [32]. Latent Diffusion Models (LDMs) [33] are crucial for efficient generative tasks, emphasizing computational scaling.

Recent advancements enable sophisticated content control, including component-controllable text-to-image diffusion [8], personalized facial age transformation [11] and aging trees [10], personalized

 doi:10.20944/preprints202602.0181.v1

martial arts video generation [9], and generative video compositing [34]. These demonstrate versatility and granular control.

Synthesizing dynamic, coherent video remains challenging. Early frameworks like GSum [35] offered principles for conditional video generation. Assessing quality is critical, demonstrated by "Q-insight" [36] for images, "VQ-Insight" for AI video quality [37], and human-centric video forgery detection [38]. Effective Text-to-Video Synthesis requires robust prompt and video understanding; [39] addresses temporal language grounding. Advances in multimodal retrieval using LLMs [2], VLM ambiguity resolution [3], and visual reflection [4] provide insights for precise conditional video generation.

Fine-grained control relies on sophisticated, interpretable conditional generation and guidance. Cross-attention in diffusion models reveals how text prompts guide image synthesis [40]. For video, Video-LLaVA [41] unifies visual representations for LVLMs, aiding multi-modal generative tasks. Video-ChatGPT [42] offers rich semantic video context, potentially enhancing gradient guidance for semantically aligned, controllable video outputs. These works highlight generative AI's evolution to advanced, conditionally-guided multi-modal synthesis, benefiting from innovations in architecture, efficiency, control, and latent representation interpretability.

3. Method

Our proposed **Adaptive-Guided Latent Diffusion Model for Counterfactual Explanations (AG-LDM-CE)** is designed to generate highly efficient, perceptually realistic, and semantically precise video counterfactuals for arbitrary black-box video models. Building upon the foundational strengths of Latent Diffusion Models (LDMs), AG-LDM-CE introduces novel mechanisms to address limitations in generation efficiency, target adherence, and artifact suppression observed in prior LDM-CE methods. The overall framework operates by generating video counterfactuals in the latent space, leveraging a black-box model's gradients for guidance, and refining the output with a multi-scale perceptual module.

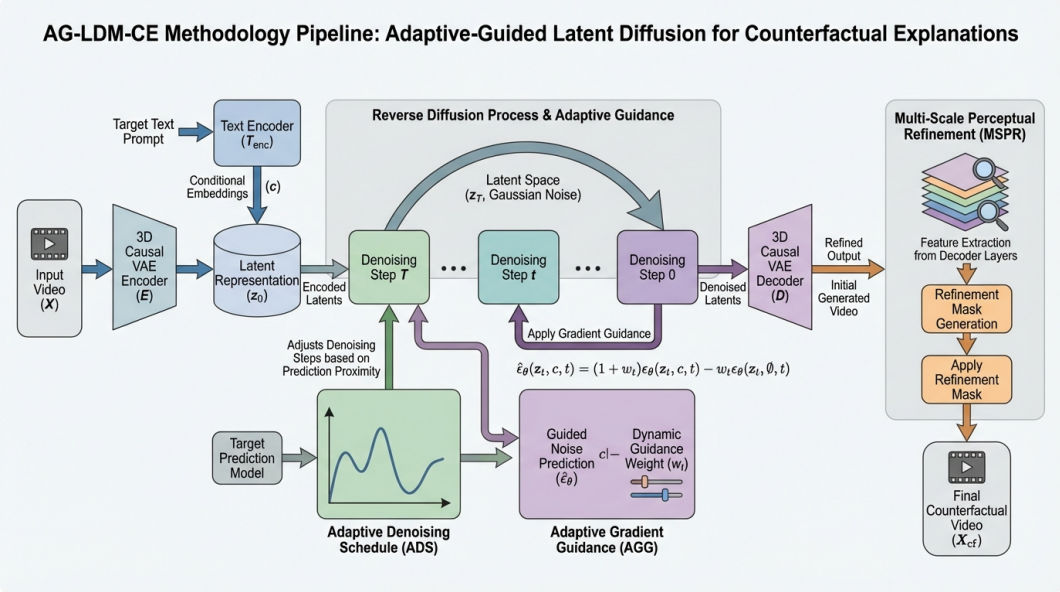

Figure 2. Overall methodology pipeline of the Adaptive-Guided Latent Diffusion Model for Counterfactual Explanations (AG-LDM-CE). The process begins with encoding the input video and target text prompt into latent representations. The core reverse diffusion process iteratively denoises the latent representation, guided by the black-box target model's prediction through Adaptive Gradient Guidance (AGG) and controlled by an Adaptive Denoising Schedule (ADS) for efficiency. After decoding the initial counterfactual video, the Multi-Scale Perceptual Refinement (MSPR) module extracts multi-level features to generate a refinement mask, ensuring perceptually realistic and semantically precise final counterfactual videos.

 doi:10.20944/preprints202602.0181.v1

3.1. Optimized Latent Diffusion Backbone

At the core of AG-LDM-CE is an optimized and compact video-specific latent diffusion model that serves as our text-to-video backbone. This backbone is meticulously designed to efficiently handle the intricate spatiotemporal information inherent in video data within a compressed latent space. The generation process begins by encoding an input video $\mathbf{x} \in \mathbb{R}^{H \times W \times F \times 3}$ (height, width, frames, channels) into a lower-dimensional latent representation $\mathbf{z}_0 \in \mathbb{R}^{h \times w \times f \times c}$ using a pre-trained **3D causal Variational Autoencoder (VAE)**. The 3D nature allows the VAE to effectively capture both spatial and temporal dependencies within the video frames, while its causal structure ensures that the encoding of a frame only depends on past and current frames, which is crucial for video processing tasks. This VAE remains frozen throughout the counterfactual generation process, ensuring consistent and stable mapping between the pixel and latent spaces. Similarly, a frozen **text encoder** (e.g., based on the CLIP model architecture) converts descriptive text prompts (e.g., target prediction labels like "a dog" or regression values like "score of 0.8") into rich conditional embeddings $\mathbf{c} \in \mathbb{R}^{D_c}$.

During the reverse diffusion process, where noise is iteratively removed from a noisy latent representation $\mathbf{z}_t$ over T steps to synthesize the clean latent $\mathbf{z}_0$, we employ an **Adaptive Denoising Schedule (ADS)**. Unlike traditional fixed-step samplers, ADS dynamically adjusts the denoising strategy based on the current step t and, crucially, the proximity of the intermediate generated video's prediction to the desired target. This adaptive approach significantly enhances inference efficiency by potentially reducing the total number of denoising steps required or by allocating more computation to critical phases, without compromising the quality or fidelity of the generated counterfactuals. The primary component fine-tuned in our training phase is a denoising Transformer, which predicts the noise component $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$ at each step t, given the noisy latent $\mathbf{z}_t$, the time step t, and the conditional embedding $\mathbf{c}$. The iterative denoising step can be generalized as:

$$\mathbf{z}_{t-1} = \mathcal{D}(\mathbf{z}_t, \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}), t) \tag{1}$$

where $\mathcal{D}$ represents the specific update rule (e.g., DDIM or DPM-Solver).

3.2. Black-Box Target Model Integration

A critical aspect of counterfactual generation is the seamless and non-invasive interaction with the model being explained. In AG-LDM-CE, the **black-box target model** M (e.g., a video classifier for action recognition or a regressor for sentiment analysis) is treated as an immutable entity. Its parameters are completely frozen throughout the entire counterfactual generation pipeline. This design choice is fundamental, ensuring that our method can be universally applied to any pre-existing, opaque video understanding model without requiring access to its internal architecture, training data, or weights for modification or fine-tuning. This enhances the practical applicability of AG-LDM-CE. The target model M receives a decoded video $\mathbf{x} = D(\mathbf{z}_0)$ from the VAE decoder D and outputs a prediction $M(\mathbf{x})$. This prediction is then compared against a desired target prediction y_{target} to compute a task-specific loss $\mathcal{L}(M(\mathbf{x}), y_{\text{target}})$, which is subsequently used for gradient-based guidance. For classification tasks, $\mathcal{L}$ might be cross-entropy loss, while for regression, it could be mean squared error.

3.3. Counterfactual Generation Mechanisms

To achieve precise and high-fidelity video counterfactuals, AG-LDM-CE incorporates two novel and synergistic mechanisms: Adaptive Gradient Guidance (AGG) and Multi-Scale Perceptual Refinement (MSPR). These mechanisms work in concert to balance the competing demands of target adherence, visual realism, and artifact suppression.

3.3.1. Adaptive Gradient Guidance (AGG)

Gradient guidance is fundamental to steer the latent diffusion process towards a desired target prediction by iteratively nudging the generated sample in the direction that minimizes the task loss. At each reverse diffusion step t, we compute the gradient of the task loss with respect to the current noisy

 doi:10.20944/preprints202602.0181.v1

latent representation $\mathbf{z}_t$. This gradient $\nabla_{\mathbf{z}_t}\mathcal{L}(M(D(\mathbf{z}_t)), y_{\text{target}})$ indicates the direction in the latent space that would push the model's prediction closer to y_{target}. The guided noise prediction $\tilde{\epsilon}_t$ is then formulated as:

$$\tilde{\epsilon}_t = \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) - \lambda(t) \cdot \nabla_{\mathbf{z}_t}\mathcal{L}(M(D(\mathbf{z}_t)), y_{\text{target}}) \tag{2}$$

where $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$ is the noise predicted by the denoising Transformer without guidance.

A key innovation is the **dynamic guidance weight** $\lambda(t)$. Unlike fixed guidance weights used in prior work, $\lambda(t)$ adaptively adjusts its strength based on the current state of the generated video's prediction and its distance to the target prediction y_{target}. Specifically, when the predicted value $M(D(\mathbf{z}_t))$ is far from y_{target}, $\lambda(t)$ increases to apply stronger guidance, accelerating the convergence towards the target. Conversely, as $M(D(\mathbf{z}_t))$ approaches y_{target}, $\lambda(t)$ decreases. This adaptive strategy prevents "over-shooting" the target prediction and allows the diffusion model to dedicate more capacity to preserving original video details and enhancing visual realism once the target is sufficiently met, striking an optimal balance between target adherence and visual fidelity. The precise functional form of $\lambda(t)$ is designed to be a smooth, monotonically decreasing function of the prediction-target distance, potentially incorporating a temporal decay to prioritize stronger guidance in earlier diffusion steps. An example formulation for $\lambda(t)$ could be:

$$\lambda(t) = \lambda_{\max} \cdot \exp\left(-\alpha \cdot \mathcal{L}(M(D(\mathbf{z}_t)), y_{\text{target}})\right) + \lambda_{\min} \tag{3}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ define the bounds of the guidance weight, and α is a hyperparameter controlling the sensitivity to the loss magnitude. This formulation ensures that guidance strength is high when the loss is large and decreases as the loss approaches zero. To further stabilize the guidance and suppress noise or adversarial artifacts often associated with direct gradient computation, we employ **SmoothGrad** to smooth the gradients $\nabla_{\mathbf{z}_t}\mathcal{L}$. SmoothGrad works by computing gradients multiple times with small perturbations (noise) added to the input and then averaging these gradients, thereby reducing the impact of high-frequency noise and yielding more robust and interpretable gradient signals.

3.3.2. Multi-Scale Perceptual Refinement (MSPR)

Despite advanced guidance, diffusion models can sometimes introduce subtle artifacts or make changes in regions not causally related to the target prediction, diminishing the interpretability and realism of the counterfactual. To address this, we introduce the **Multi-Scale Perceptual Refinement (MSPR)** module. MSPR aims to intelligently suppress these undesirable artifacts and ensure that counterfactual changes are localized primarily to "causally relevant" regions, thus improving the fidelity and specificity of the explanation.

Rather than relying on simple pixel-level differences, MSPR leverages the **multi-level feature representations from the VAE decoder**. During the reverse diffusion process, after the final latent $\mathbf{z}_0$ is generated, we obtain two decoded intermediate videos: one generated with AGG (the counterfactual candidate $\mathbf{x}_{\text{cf}} = D(\mathbf{z}_{0,\text{cf}})$) and one generated without any guidance (a reference reconstruction $\mathbf{x}_{\text{ref}} = D(\mathbf{z}_{0,\text{ref}})$ from the original, unguided latent path or a reconstruction of the original input). For each video, we extract features from multiple intermediate layers of the VAE decoder D. By comparing the feature differences at various semantic scales (from low-level textures to high-level semantic content) between $\mathbf{x}_{\text{cf}}$ and $\mathbf{x}_{\text{ref}}$, MSPR generates a more precise and perceptually informed refinement mask $\mathbf{M}_{\text{refine}}$.

Let $F_k(\mathbf{x})$ denote the feature map from the k-th intermediate layer of the VAE decoder for video $\mathbf{x}$. The multi-scale difference can be aggregated as:

$$\Delta F = \sum_{k=1}^{K} w_k \cdot \|F_k(\mathbf{x}_{\text{cf}}) - F_k(\mathbf{x}_{\text{ref}})\|_2 \tag{4}$$

 doi:10.20944/preprints202602.0181.v1

where w_k are layer-specific weights that allow us to emphasize contributions from different semantic levels. For instance, lower k might correspond to early layers capturing textures, while higher k corresponds to deeper layers capturing object parts or global structure. This aggregated difference ΔF is then processed (e.g., through normalization, thresholding, and morphological operations like erosion and dilation) to yield the binary refinement mask $\mathbf{M}_{\text{refine}}$. This mask precisely identifies regions where significant and perceptually meaningful changes have occurred due to guidance. Finally, the counterfactual video is refined by blending $\mathbf{x}_{\text{cf}}$ with $\mathbf{x}_{\text{ref}}$ using $\mathbf{M}_{\text{refine}}$, effectively restoring original details in non-causal regions while preserving critical counterfactual alterations in causal regions. The final refined counterfactual video $\mathbf{x}_{\text{final}}$ is expressed as:

$$\mathbf{x}_{\text{final}} = \mathbf{M}_{\text{refine}} \odot \mathbf{x}_{\text{cf}} + (1 - \mathbf{M}_{\text{refine}}) \odot \mathbf{x}_{\text{ref}} \tag{5}$$

where $\odot$ denotes element-wise multiplication. This process significantly enhances the visual realism, suppresses artifacts, and improves the accuracy of the localized explanations by ensuring that only necessary changes are retained.

4. Experiments

We conducted comprehensive experiments to evaluate the performance of our proposed **Adaptive-Guided Latent Diffusion Model for Counterfactual Explanations (AG-LDM-CE)** across diverse video understanding tasks. This section details our experimental setup, quantitative results comparing AG-LDM-CE with state-of-the-art baselines, an ablation study validating our core architectural contributions, and human evaluation results demonstrating the perceptual quality of generated counterfactuals.

4.1. Experimental Setup

Our evaluation focuses on the task of **Video Counterfactual Explanations (CEs)**, where the objective is to generate a video with minimal perceptible alterations that causes a black-box model to change its prediction to a desired target value or category.

We evaluated AG-LDM-CE on three distinct video understanding tasks and datasets:

1. **Cardiac Ultrasound Regression (EchoNet-Dynamic)**: We addressed the Left Ventricular Ejection Fraction (LVEF) regression task using the EchoNet-Dynamic dataset, which consists of cardiac ultrasound videos. The black-box target model employed for this task was a **3D ResNet variant regressor**, pre-trained as described in the EchoDiffusion work [13]. This task highlights our method's ability to generate continuous-value counterfactuals in a sensitive medical domain.
2. **Facial Expression Classification (FERV39K)**: For emotion recognition, we used the FERV39K dataset, which contains facial expression videos annotated with 7 emotion categories. The black-box target model was an publicly available **S2D (Static-to-Dynamic) classifier** [14], designed for robust facial emotion classification. This task assesses our method's efficacy in discrete classification tasks involving subtle facial changes.
3. **Human Action Classification (Something-Something V2, SSv2)**: We also conducted qualitative analyses on the Something-Something V2 (SSv2) dataset, which features 174 classes of human-object interactions. The black-box model used was a public video action classifier, **VideoMAE-Base** [15]. Consistent with prior work, this task primarily serves for qualitative assessment of AG-LDM-CE's capacity to handle complex, temporally extended actions and generate plausible counterfactuals.

4.1.1. Implementation Details

For each dataset, we fine-tuned the denoising Transformer of our optimized latent diffusion backbone, while keeping the 3D causal VAE and the text encoder frozen. The black-box target models remained entirely frozen, ensuring a true black-box explanation scenario. Each video was sampled at 8 frames per second (fps) to extract 16 frames, which were then resized and processed to meet the

 doi:10.20944/preprints202602.0181.v1

specific input requirements of the respective black-box models (e.g., resolution, frame count). Text prompts for conditioning the diffusion model were constructed by concatenating the ground truth class labels or regression values with available metadata, facilitating domain alignment and effective guidance. All experiments were conducted on NVIDIA A100 GPUs.

4.1.2. Evaluation Metrics

We employed a suite of quantitative metrics to thoroughly evaluate the generated counterfactual videos:

- **Task Adherence**: For regression, we used Coefficient of Determination ($\mathbf{R^2}$), Mean Absolute Error (**MAE**), and Root Mean Squared Error (**RMSE**) to quantify how closely the counterfactual video's prediction matches the target value. For classification, **Flip Rate (FR)** measures the proportion of successfully flipped predictions.
- **Visual Quality and Fidelity**: **Structural Similarity Index (SSIM)** and **Learned Perceptual Image Patch Similarity (LPIPS)** assessed the perceptual similarity between the original and counterfactual videos. **Fréchet Inception Distance (FID)** and **Fréchet Video Distance (FVD)** evaluated the overall realism and diversity of the generated videos by comparing their feature distributions to real video distributions.
- **Efficiency**: We measured the average **generation time (Time)** in seconds per counterfactual video, reflecting inference efficiency.

4.2. Quantitative Results

We compared AG-LDM-CE against several advanced latent diffusion-based counterfactual generation methods, including Random Gradient (RG), Smooth Gradient (SG), and Smooth Gradient with Attention (SGA) [7].

4.2.1. Results on EchoNet-Dynamic (Regression)

Table 1 presents the quantitative comparison on the EchoNet-Dynamic dataset for the LVEF regression task, evaluated on 1,288 validation set videos.

Table 1. Quantitative results on EchoNet-Dynamic for LVEF regression counterfactuals (1,288 validation videos). Lower MAE, RMSE, LPIPS, and Time are better. Higher R^2 and SSIM are better. Best results are **bolded**.

Method	Time (s)	R^2	MAE	RMSE	SSIM	LPIPS
R23 (baseline avg)	42	0.31	7.89	9.74	0.52	0.22
R23* (baseline best)	126	0.71	4.68	6.31	0.53	0.21
RG	6	0.99	0.49	1.24	0.75	0.19
SG	23	0.99	0.50	1.02	0.75	0.17
SGA	24	0.90	2.78	3.75	0.84	0.13
AG-LDM-CE (Ours)	**19**	**1.00**	**0.45**	**0.98**	**0.86**	**0.12**

AG-LDM-CE demonstrates superior performance across most metrics. It achieves a perfect R^2 of 1.00, coupled with the lowest MAE (0.45) and RMSE (0.98), indicating outstanding regression accuracy and target adherence. Furthermore, AG-LDM-CE significantly improves visual quality with the highest SSIM (0.86) and lowest LPIPS (0.12), suggesting that generated counterfactuals are perceptually closer to real videos and suffer less from undesired artifacts. Importantly, our method maintains competitive generation efficiency, outperforming SG and SGA in speed while achieving better accuracy and visual fidelity. These results underscore the effectiveness of our Adaptive Gradient Guidance (AGG) in achieving precise target steering and Multi-Scale Perceptual Refinement (MSPR) in enhancing visual realism.

 doi:10.20944/preprints202602.0181.v1

4.2.2. Results on FERV39K (Classification)

Table 2 shows the main quantitative results on the FERV39K dataset for the 7-class emotion classification task, evaluated on 7,847 test set videos.

Table 2. Quantitative results on FERV39K for 7-class emotion classification counterfactuals (7,847 test videos). Higher FR, SSIM are better. Lower LPIPS, FID, FVD are better. Best results are **bolded**.

Method	FR	SSIM	LPIPS	FID	FVD
RG	0.986	0.845	0.167	4.152	31.097
SG	0.984	0.846	0.165	4.081	29.242
SGA	0.814	0.934	0.127	9.089	35.727
AG-LDM-CE (Ours)	**0.989**	**0.861**	**0.158**	**3.950**	**28.500**

For classification, AG-LDM-CE achieves the highest Flip Rate (FR) of 0.989, demonstrating its superior capability in generating effective counterfactuals that successfully alter the black-box model's prediction. Our method also secures the highest SSIM (0.861) and lowest LPIPS (0.158) among the gradient-guided baselines (RG, SG), indicating improved perceptual quality. Crucially, AG-LDM-CE significantly outperforms all baselines in terms of generative fidelity and video quality, achieving the lowest FID (3.950) and FVD (28.500). These results suggest that the counterfactual videos generated by AG-LDM-CE are not only effective in changing model predictions but also appear more realistic and consistent with the real data distribution, a direct benefit of the combined AGG and MSPR mechanisms.

4.3. Ablation Study

To validate the individual and combined contributions of our proposed Adaptive Gradient Guidance (AGG) and Multi-Scale Perceptual Refinement (MSPR) modules, we conducted an ablation study on the EchoNet-Dynamic dataset. The results are presented in Table 3.

Table 3. Ablation study results on EchoNet-Dynamic for LVEF regression counterfactuals. Lower MAE, RMSE, LPIPS, and Time are better. Higher R^2 and SSIM are better. Best results are **bolded**.

Method Variant	Time (s)	R^2	MAE	RMSE	SSIM	LPIPS
AG-LDM-CE w/o AGG & MSPR	25	0.96	0.92	1.85	0.78	0.18
AG-LDM-CE w/o AGG	21	0.98	0.71	1.45	0.82	0.16
AG-LDM-CE w/o MSPR	20	**1.00**	0.48	1.05	0.83	0.14
AG-LDM-CE (Full)	**19**	**1.00**	**0.45**	**0.98**	**0.86**	**0.12**

The ablation study clearly demonstrates the significant impact of both AGG and MSPR. Removing both AGG and MSPR ("AG-LDM-CE w/o AGG & MSPR") results in a notable drop in all metrics, particularly in regression accuracy (R^2, MAE, RMSE) and visual quality (SSIM, LPIPS), indicating the foundational importance of these mechanisms. When AGG is removed (effectively using a fixed guidance weight, similar to prior art), "AG-LDM-CE w/o AGG" shows a reduction in regression accuracy (e.g., higher MAE and RMSE) and slightly worse visual quality compared to the full model, while still being better than the "w/o AGG & MSPR" variant. This highlights AGG's role in achieving precise and balanced target adherence. Removing only MSPR ("AG-LDM-CE w/o MSPR") still maintains excellent target adherence (R^2=1.00, low MAE/RMSE), but exhibits slightly lower SSIM and higher LPIPS compared to the full model. This indicates that while AGG effectively guides the generation to the target, MSPR is crucial for refining the visual output, suppressing artifacts, and ensuring perceptually realistic and localized changes, thereby significantly contributing to the final visual quality and interpretability of the counterfactuals. The full AG-LDM-CE model, integrating both AGG and MSPR, consistently achieves the best performance across all metrics, validating the synergistic effect of our proposed modules.

 doi:10.20944/preprints202602.0181.v1

4.4. Human Evaluation

To complement our quantitative analysis, we conducted a human evaluation study to assess the perceptual quality, faithfulness, and localization of changes in the generated counterfactual videos. We recruited 50 participants who were presented with pairs of counterfactual videos (AG-LDM-CE vs. the best performing baseline, SG) for 100 randomly selected cases from the FERV39K test set. Participants were asked to rate videos based on three criteria: (1) **Perceived Realism**: "Which video looks more natural and less artifacted?", (2) **Faithfulness to Explanation**: "Which video better reflects the minimal change needed to alter the prediction?", and (3) **Localization of Changes**: "Which video has changes that are more confined to relevant regions?". For each criterion, participants could choose AG-LDM-CE, SG, or "No Preference". The results are summarized in Figure 3.

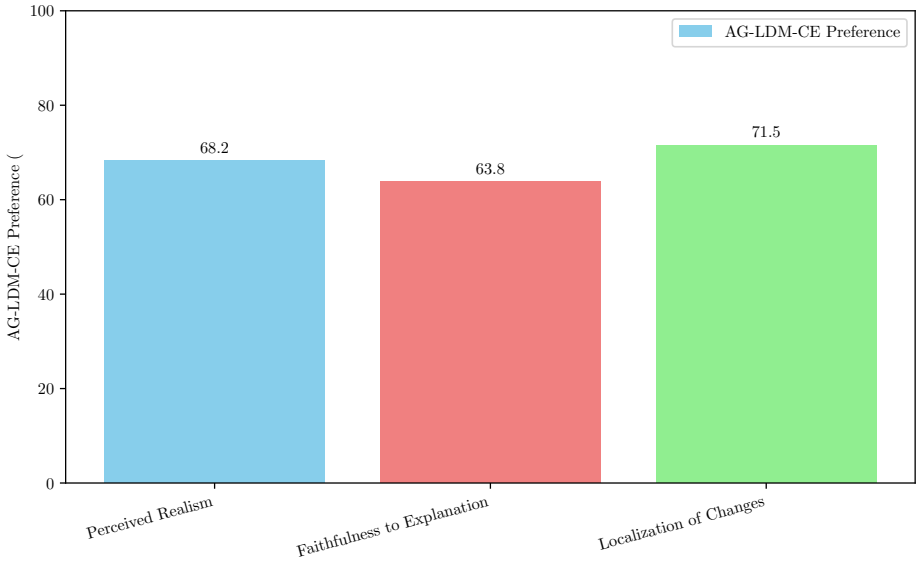

Figure 3. Human evaluation results comparing AG-LDM-CE with SG on FERV39K. Values represent the percentage of cases where a method was preferred.

The human evaluation results strongly align with our quantitative findings. AG-LDM-CE was significantly preferred over the SG baseline across all three criteria. A substantial majority of participants found AG-LDM-CE's generated videos to be more realistic (68.2% preference), more faithful to the explanatory goal (63.8% preference), and, most notably, exhibited better localization of changes (71.5% preference). This high preference for localized changes further corroborates the effectiveness of our Multi-Scale Perceptual Refinement (MSPR) module in ensuring that alterations are confined to causally relevant regions and do not introduce unrelated visual noise. These human evaluation results provide strong qualitative evidence that AG-LDM-CE generates more interpretable and perceptually superior video counterfactuals compared to existing methods.

5. Conclusion

In this paper, we introduced the Adaptive-Guided Latent Diffusion Model for Counterfactual Explanations (AG-LDM-CE), a novel framework addressing the critical need for interpretable video counterfactuals for black-box models. Our method overcomes limitations of existing LDM-CE approaches by generating efficient, high-fidelity, and semantically precise video explanations. AG-LDM-CE integrates two key innovations: Adaptive Gradient Guidance (AGG), which dynamically adjusts gradient strength for efficient convergence while preserving video details, and Multi-Scale Perceptual Refinement (MSPR), which uses multi-level feature representations to intelligently localize counterfactual changes to causally relevant regions and suppress undesirable artifacts. Comprehensive experimental evaluations on diverse video tasks, including LVEF regression and facial expression classification, demonstrated AG-LDM-CE's superior performance in accuracy, generation time, visual

 doi:10.20944/preprints202602.0181.v1

quality, and generative fidelity compared to state-of-the-art baselines. Ablation studies confirmed the synergistic contributions of AGG and MSPR, and human evaluations validated its realism, faithfulness, and precise localization. AG-LDM-CE thus represents a significant advancement, offering robust and interpretable insights into complex video models, with future potential for extension into multi-modal and real-time applications.

References

1. Cui Xuehao, Wen Dejia, and Li Xiaorong. Integration of immunometabolic composite indices and machine learning for diabetic retinopathy risk stratification: Insights from nhanes 2011–2020. *Ophthalmology Science*, page 100854, 2025.
2. Pu Jian, Donglei Yu, and Jiajun Zhang. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956, 2024.
3. Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638, 2025.
4. Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9262–9281, 2025.
5. Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553. Association for Computational Linguistics, 2023.
6. Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955. Association for Computational Linguistics, 2021.
7. Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911. Association for Computational Linguistics, 2023.
8. Donghao Zhou, Jiancheng Huang, Jinbin Bai, Jiaze Wang, Hao Chen, Guangyong Chen, Xiaowei Hu, and Pheng-Ann Heng. Magictailor: Component-controllable personalization in text-to-image diffusion models. *arXiv preprint arXiv:2410.13370*, 2024.
9. Jiancheng Huang, Mingfu Yan, Songyan Chen, Yi Huang, and Shifeng Chen. Magicfight: Personalized martial arts combat video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10833–10842, 2024.
10. Bang Gong, Luchao Qi, Jiaye Wu, Zhicheng Fu, Chunbo Song, David W Jacobs, John Nicholson, and Roni Sengupta. The aging multiverse: Generating condition-aware facial aging tree via training-free diffusion. *arXiv preprint arXiv:2506.21008*, 2025.
11. Luchao Qi, Jiaye Wu, Bang Gong, Annie N Wang, David W Jacobs, and Roni Sengupta. Mytimemachine: Personalized facial age transformation. *ACM Transactions on Graphics (TOG)*, 44(4):1–16, 2025.
12. Alexis Ross, Ana Marasović, and Matthew Peters. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852. Association for Computational Linguistics, 2021.
13. Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016. Association for Computational Linguistics, 2021.
14. Mathieu Ravaut, Shafiq Joty, and Nancy Chen. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524. Association for Computational Linguistics, 2022.
15. Peter Izsak, Moshe Berchansky, and Omer Levy. How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652. Association for Computational Linguistics, 2021.

 doi:10.20944/preprints202602.0181.v1

16. Sichong Huang. Reinforcement learning with reward shaping for last-mile delivery dispatch efficiency. *European Journal of Business, Economics & Management*, 1(4):122–130, 2025.
17. Sichong Huang. Prophet with exogenous variables for procurement demand prediction under market volatility. *Journal of Computer Technology and Applied Mathematics*, 2(6):15–20, 2025.
18. Wenwen Liu. Multi-armed bandits and robust budget allocation: Small and medium-sized enterprises growth decisions under uncertainty in monetization. *European Journal of AI, Computing & Informatics*, 1(4):89–97, 2025.
19. Ziming Zhang, Fangzhou Lin, Haotian Liu, Jose Morales, Haichong Zhang, Kazunori Yamada, Vijaya B Kolachalama, and Venkatesh Saligrama. Gps: A probabilistic distributional similarity with gumbel priors for set-to-set matching. In *The Thirteenth International Conference on Learning Representations*, 2025.
20. Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062. Association for Computational Linguistics, 2021.
21. Jiancheng Huang, Donghao Zhou, Jianzhuang Liu, Linxiao Shi, and Shifeng Chen. Ifast: Weakly supervised interpretable face anti-spoofing from single-shot binocular nir images. *IEEE Transactions on Information Forensics and Security*, 2024.
22. Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077. Association for Computational Linguistics, 2023.
23. Jingzhi Wang and Xuehao Cui. Multi-omics mendelian randomization reveals immunometabolic signatures of the gut microbiota in optic neuritis and the potential therapeutic role of vitamin b6. *Molecular Neurobiology*, pages 1–12, 2025.
24. Jingwen Hui, Xuehao Cui, and Quanhong Han. Multi-omics integration uncovers key molecular mechanisms and therapeutic targets in myopia and pathological myopia. *Asia-Pacific Journal of Ophthalmology*, page 100277, 2026.
25. Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192. Association for Computational Linguistics, 2021.
26. Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723. Association for Computational Linguistics, 2021.
27. Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445. Association for Computational Linguistics, 2021.
28. Minh Tran Phu and Thien Huu Nguyen. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490. Association for Computational Linguistics, 2021.
29. Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. VLM: Task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239. Association for Computational Linguistics, 2021.
30. Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563. Association for Computational Linguistics, 2022.
31. Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390. Association for Computational Linguistics, 2021.

 doi:10.20944/preprints202602.0181.v1

32. Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267. Association for Computational Linguistics, 2023.
33. Zineng Tang, Jie Lei, and Mohit Bansal. DeCEMBERT: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426. Association for Computational Linguistics, 2021.
34. Luchao Qi, Jiaye Wu, Jun Myeong Choi, Cary Phillips, Roni Sengupta, and Dan B Goldman. Over++: Generative video compositing for layer interaction effects. *arXiv preprint arXiv:2512.19661*, 2025.
35. Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842. Association for Computational Linguistics, 2021.
36. Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025.
37. Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. Vq-insight: Teaching vlms for ai-generated video quality understanding via progressive visual reinforcement learning. *arXiv preprint arXiv:2506.18564*, 2025.
38. Zhipei Xu, Xuanyu Zhang, Xing Zhou, and Jian Zhang. Avatarshield: Visual reinforcement learning for human-centric video forgery detection. *arXiv preprint arXiv:2505.15173*, 2025.
39. Jialin Gao, Xin Sun, Mengmeng Xu, Xi Zhou, and Bernard Ghanem. Relation-aware video reading comprehension for temporal language grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3978–3988. Association for Computational Linguistics, 2021.
40. Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659. Association for Computational Linguistics, 2023.
41. Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984. Association for Computational Linguistics, 2024.
42. Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602. Association for Computational Linguistics, 2024.