CoT-Edit: Let CoT Guide Instruction Video Editing

Sen Liang[1*] Fengbin Guan[1*] Youliang Zhang[3] Xin Li[1†] Zhibo Chen[1,2†]
[1]University of Science and Technology of China
[2]Zhongguancun Academy [3]Tsinghua University
{liangsen, guanfb}@mail.ustc.edu.cn zhangyou24@mails.tsinghua.edu.cn
{xin.li, chenzhibo}@ustc.edu.cn

Abstract

Text-driven instruction-based video editing in complex scenes remains challenging: purely textual prompts often fail to capture precise spatial relationships and physical constraints, resulting in target ambiguity and physically implausible outcomes. To address this, we propose a plan–guide–edit framework that explicitly bridges semantic intent and spatial execution. In our framework, a Chain-of-Thought (CoT)-enhanced multimodal large language model (MLLM) serves as a planner, performing structured reasoning over the video and instructions to derive a precise sequence of bounding boxes and attribute-enriched editing directives. These spatial priors then guide a box-conditioned mask generator, transforming ambiguous global retrieval into localized, context-aware refinement and producing masks that more accurately capture object scale, contact relationships, and placement. Building on these spatial and semantic signals, a diffusion-based editor integrates the masks, enriched instructions, and frame features to render high-fidelity edits that remain temporally coherent and spatially well aligned. Trained first in a modular manner and then jointly, our framework achieves superior performance with reduced data requirements, delivering precise localization in scenes with multiple similar objects and physically consistent object additions, and extensive experiments demonstrate state-of-the-art performance over multiple strong baseline methods. More details are available at: https://github.com/flying-sky999/CoT-Edit

1. Introduction

Instruction-based video editing [2, 27, 29] has emerged as a transformative capability, enabling users to manipulate video content through natural language without requiring professional post-production skills, with applications spanning content creation and augmented reality. However, in complex real-world scenarios involving multiple coexisting objects, densely distributed similar entities, or dynamic interactions, reliance solely on textual prompts frequently leads to control failures. Purely text-driven editing models [27, 34, 45] must simultaneously comprehend cross-frame semantics, localize targets, and execute edits based on ambiguous linguistic signals. This multifaceted challenge compels models to depend on large-scale datasets and high-capacity models, yet these models still frequently result in localization drift, erroneous edits, and temporal instability due to insufficient spatial grounding and enforcement

Figure 1. Illustration of the challenges in text-driven instruction-based video editing. Given instructions such as "Add a UFO that flies in an elliptical pattern in the sky" (top) and "Change the yellow dog into an orange cat" (bottom), text-only methods struggle to respect trajectory constraints and to reliably select the intended target among multiple similar objects.

*Equal contribution.
†Corresponding authors.

of physical constraints. As illustrated in Fig. 1, even seemingly simple instructions such as "Add a UFO that flies in an elliptical pattern in the sky" or "Change the yellow dog into an orange cat" can lead text-only methods to violate motion trajectories or to alter the wrong dog among several similar ones.

A feasible way to mitigate these issues is to introduce additional conditioning to decouple *what to edit* from *where to edit*. Masks are among the most practical spatial conditioning signals, offering clear cues that transform global retrieval into controllable local editing, reducing data requirements and enhancing controllability. However, if masks are generated solely from the original textual instruction, semantic ambiguity and spatial uncertainty are directly propagated to the mask layer, compromising its quality and robustness. More critically, for object addition tasks, text-derived masks offer no actionable physical priors (e.g., plausible position, scale, or motion logic for the new object), leaving the model without a grounded basis for generation. Therefore, instruction-based video editing calls for a new framework that can build a robust bridge between high-level semantics and low-level pixels: it should preserve the expressiveness of language while offering fine-grained spatial and physical constraints, while avoiding heavy reliance on large amounts of aligned annotations.

To address this need, we propose a *Plan–Guide–Edit* paradigm for instruction-based video editing, which explicitly structures the mapping from semantic intent to spatial execution. First, a structured planning stage precedes editing, decomposing the originally implicit text comprehension process into interpretable, multi-step reasoning. Specifically, we leverage a Chain-of-Thought (CoT)-enhanced multimodal large language model (MLLM) to perform a hierarchical analysis of input video keyframes and text instructions. It first identifies the editing objects and action intents, and then, while modeling spatial localization, temporal consistency, and physical feasibility, generates a sequence of keyframe bounding boxes to precisely anchor the spatiotemporal positions to be edited. In parallel, it outputs an enriched instruction that supplements the original description with target attributes, interaction tendencies, and scene constraints, providing stronger semantic priors to subsequent modules. Guided by this, mask generation is no longer performed solely from ambiguous text but is instead guided by the explicit spatial prior of the bounding box to generate high-quality local masks, significantly boosting localization accuracy and providing essential constraints on scale and contact relationships, particularly for object addition. Finally, the main diffusion-based editing module integrates the precise spatial constraints from the masks, the enriched semantic control from the augmented instructions, and underlying video features, grounding the high-level semantic intent into specific spatio-temporal locations for more controllable instruction-based video editing.

The key to our proposed framework's ability to resolve the deficiencies of text-only methods lies in its establishment of a reliable semantic–spatial link through structured planning. First, the CoT-enhanced MLLM planner acts as a semantic–spatial translator, proactively resolving textual ambiguities before execution. The explicit generation of bounding boxes decouples the burden of spatial localization from the diffusion model, substantially simplifying subsequent mask generation and editing. Second, our design ensures physical consistency: the planner explicitly reasons about scale, spatial relationships, and kinematics, and the resulting bounding boxes and enriched instructions provide actionable physical priors for the edited objects. Finally, this decoupled *Plan–Guide–Edit* paradigm clarifies the learning objectives of each module and, via staged training, reduces the dependence on large-scale aligned data while improving stability and efficiency.

Our contributions can be summarized as follows:

- We introduce a structured *Plan–Guide–Edit* paradigm for instruction-based video editing that parses editing targets and localizes them in space and along the motion dimension before editing, enabling a clear mapping from high-level instructions to concrete video edits.
- We develop a CoT-enhanced MLLM planner that provides structured, physically aware guidance from video and instructions, and introduce a box-conditioned mask prediction branch that converts this guidance into explicit spatial priors to ease the diffusion-based editor's generation.
- Extensive experiments show that our framework achieves accurate spatial localization and consistent physical and temporal behavior, outperforming existing baselines on complex instruction-based video editing tasks.

2. Related Work

2.1. Video Generation

Early video generation relied on GANs/VAEs but faced low resolution, flickering, and temporal incoherence [5, 6, 13, 21, 32, 40]. Diffusion models revolutionized the field by extending image diffusion techniques to video, achieving high-fidelity results through temporal denoising and cascaded architectures [9, 10, 31, 37]. Scaling these models enabled large-scale text-to-video systems with broad scene coverage [11, 23, 38]. To enhance control beyond text, image-guided methods emerged, animating static inputs or adapting image models zero-shot [12, 17, 48]. Multi-condition frameworks [3, 16, 18, 20, 25, 26, 50] now integrate text, images, poses, masks, and videos to improve controllability, though consistency challenges remain.

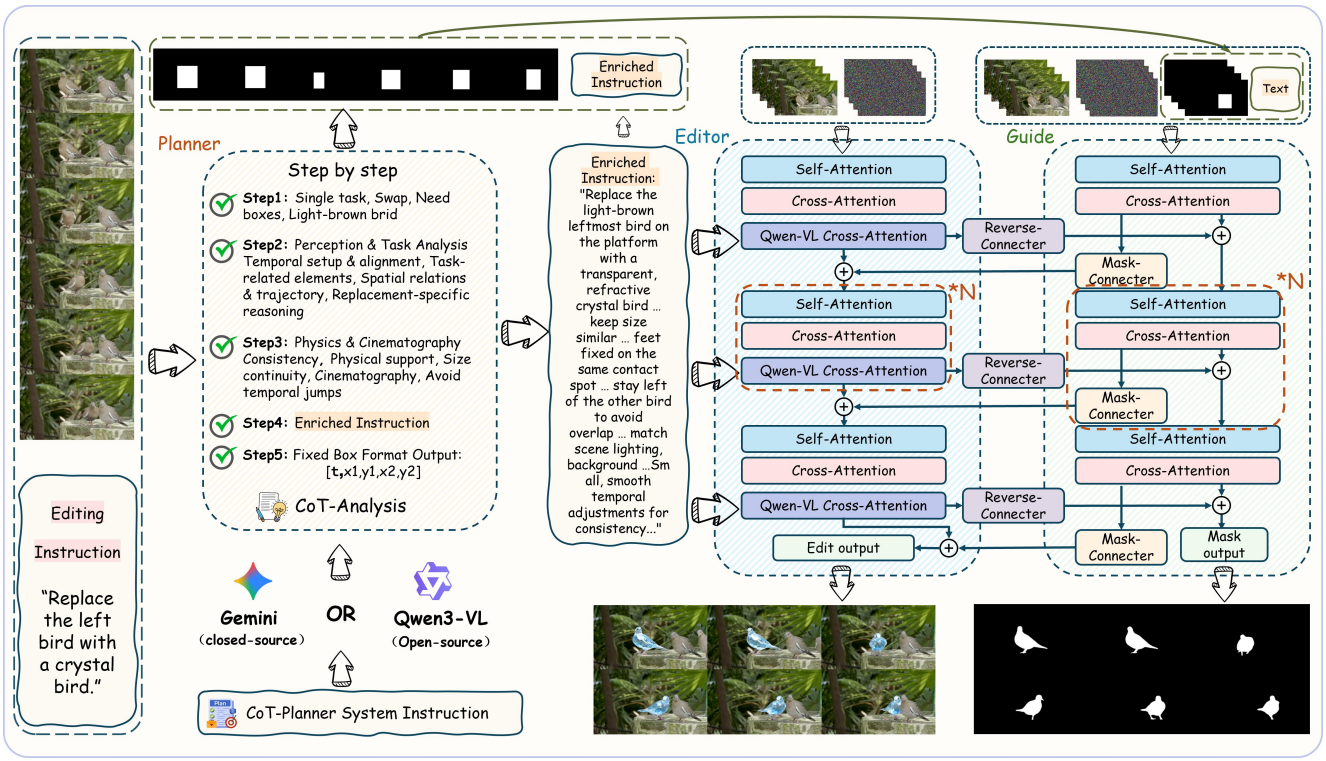

Figure 2. Overview of the proposed "Plan–Guide–Edit" framework for instruction-based video editing. Given a video and an instruction, a CoT-enhanced VLM planner performs step-by-step analysis to produce an enriched instruction and a temporal sequence of bounding boxes. The Guide branch turns these spatial priors into spatio-temporally consistent masks, while the Editor fuses text, video features, and mask guidance through bidirectional connectors to render the final edited video.

2.2. Instruction Video Editing

Non-instructional video editing methods [24, 41, 42, 44] rely on auxiliary inputs or per-video fine-tuning, which compromises flexibility and efficiency. To reduce interaction costs, the instruction-guided paradigm emerged, inspired by InstructPix2Pix [2], which requires only the original video and a natural language instruction. However, it faces dual challenges: the scarcity of large-scale, high-quality instruction-video aligned data, and maintaining spatio-temporal consistency and physical plausibility during editing. To address the data bottleneck, InstructVid2Vid [29] constructs edit pairs via per-video optimization but suffers from poor scalability. EffiVED [49] and InsV2V [4] extend image-editing techniques to synthesize pseudo video data, while recent InsViE [45] and Ditto [1] scale datasets to millions of samples through model distillation or image-to-video generation pipelines, enabling diverse editing tasks. End-to-end models (e.g., Lucy-1.1 [34]) validate the feasibility of video-instruction inputs yet exhibit inadequate spatial localization for small targets or ambiguous instructions. InstructX [27] enhances semantic generalization using multimodal large models but relies on implicit attention for spatial constraints. In contrast, methods [22, 46] emphasizing geometric consistency improve cross-frame coherence via explicit alignment yet require external modules and are not inherently instruction-driven. However, most instruction-driven methods still struggle to turn ambiguous high-level instructions into precise, physically plausible spatiotemporal edits, underscoring the need for a framework that jointly performs instruction parsing, spatial grounding, and fine-grained editing.

3. Method

We propose a *Plan–Guide–Edit* framework for instruction-based video editing tasks, as illustrated in Fig. 2. The framework is implemented through three core modules: the **Planner**, which translates high-level semantic intent into executable spatial constraints; the **Guide**, which generates spatiotemporally consistent masks based on explicit spatial priors; and the **Editor**, which executes appearance modification and content generation. This modular design ensures semantic coherence, spatial precision, and temporal continuity while reducing the complexity of editing tasks.

Base Model and Setup. For the latter two branches of our framework, we select Wan2.2 5B as the base diffusion backbone. Given the original video S, we obtain its low-

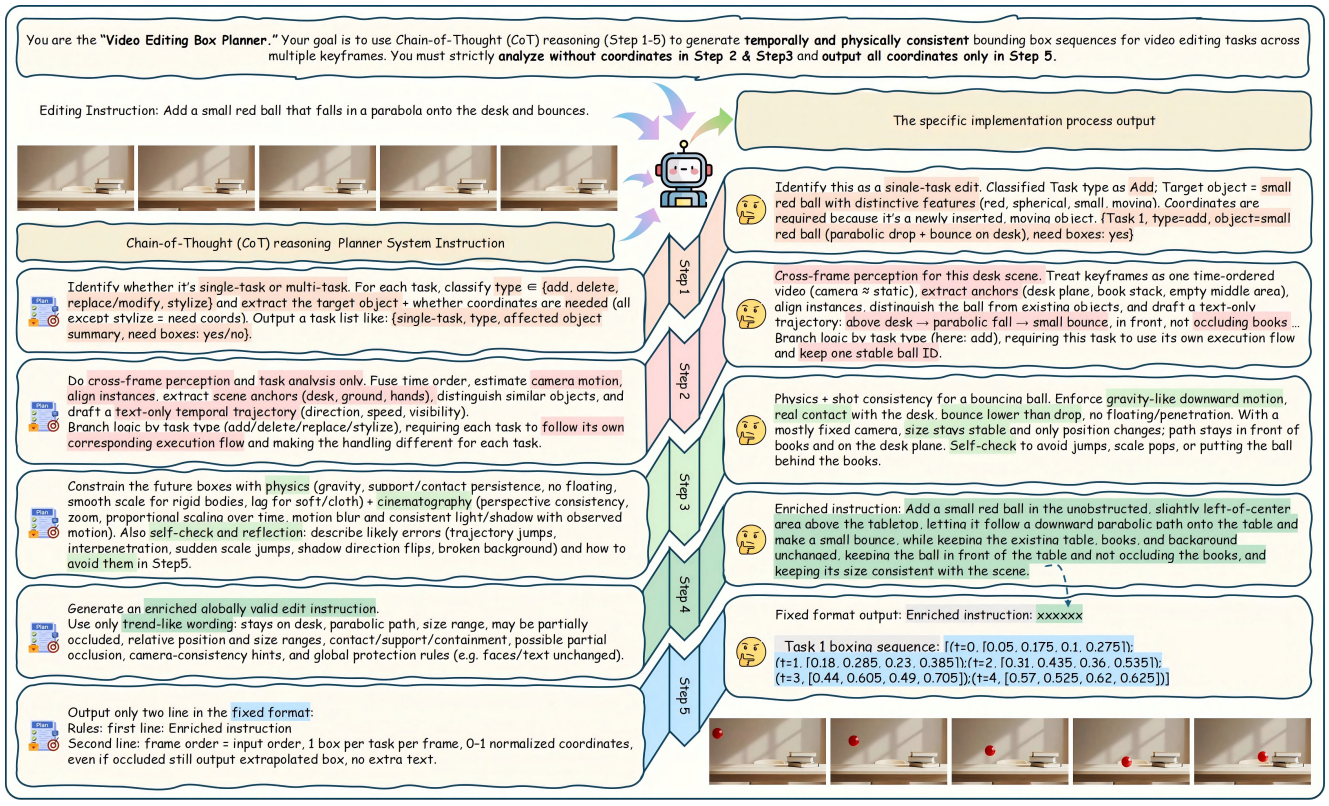

Figure 3. Chain-of-Thought (CoT) reasoning process of the MLLM planner. Given an editing instruction and keyframes, the planner follows a five-step procedure: it parses the task type and target object, performs cross-frame perception and temporal analysis, enforces physics- and cinematography-aware consistency, synthesizes an enriched instruction, and finally outputs the enriched instruction together with a temporally coherent sequence of normalized bounding boxes in a fixed format.

dimensional latent representation y_0 via a 3D VAE. During the diffusion process, noise is injected to obtain y_{noise}, and the instruction text is denoted as I. We concatenate y_0 and y_{noise} along the channel dimension:

$$y_{\text{input}} = \text{ChannelConcat}(y_{\text{noise}}, y_0), \tag{1}$$

and expand the input channel dimension of the tokenizer T to 32 while keeping the output dimension unchanged, thus avoiding disruption to the original Wan2.2 5B distribution.

3.1. Planner: Semantic–Spatial Planning with CoT-Enhanced Reasoning

The planner serves the pivotal role of translating high-level semantic intents into executable spatial constraints. We implement it as an MLLM explicitly embedded with a Chain-of-Thought (CoT) reasoning process. This design bridges ambiguous semantic instructions and precise spatial execution before mask prediction and diffusion-based editing, explicitly determining spatial localization, positional relationships, and physical consistency so that downstream modules focus on localized refinement and appearance generation under well-defined spatial priors.

Inputs and Outputs. The planner takes as input a temporally ordered sequence of keyframes $\{I_t\}_{t=1}^{T}$, together with the user's editing instruction. It produces two parallel, structured outputs: (1) a keyframe-aligned sequence of bounding boxes $\{b_t\}_{t=1}^{T}$ that explicitly specifies where the edit should occur; and (2) an enriched instruction that extends the original text with semantic priors such as target attributes, relative spatial relations, contact patterns, and camera-consistency hints, which are used to condition the mask and generation branches. For non-spatial tasks (e.g., stylization), the planner emits an empty box sequence.

Structured Reasoning Procedure

As illustrated in Fig. 3, we organize the planner's reasoning into an explicit multi-step Chain-of-Thought process instead of a direct single-pass output. This is motivated by three factors: (i) sequentially decomposing complex instructions significantly reduces task difficulty; (ii) premodeling physical and cinematic constraints prevents localization errors from amplifying in later generation stages; and (iii) a dedicated stage creates structured, interpretable guidance for downstream modules. Based on this, we decompose the planning process into three tightly connected

components described below.

(i) **Task Parsing and Cross-Frame Perception.** The objective of this stage is to establish an accurate understanding of both the editing intent and the video content. On the instruction side, the planner parses the text to determine the task type (e.g., addition, swap) and extracts the primary objects. On the visual side, it performs cross-frame perception: estimating camera motion (to distinguish from object motion) and conducting cross-frame task-relevant instance identification and tracking. This stage thus anchors an initially ambiguous linguistic reference to concrete video instances, providing a reliable basis for subsequent precise localization and temporal reasoning.

(ii) **Physics and Temporal Consistency Modeling.** This stage is a critical component of our planner, compelling the MLLM to explicitly model physical constraints and cross-frame continuity. First, inter-frame consistency reasoning requires the MLLM to infer the target's displacement, scale changes, and visibility across frames to generate a smooth localization trajectory by integrating object motion, camera motion, and perspective changes. Second, physical rule embedding leverages the MLLM's world knowledge for constraints (e.g., gravity, contact, occlusion), which, combined with a final self-check and reflection step, mitigates common failure modes in purely text-driven editing, such as inaccurate localization or physical implausibility.

(iii) **Generating Spatial and Semantic Guidance.** Following rigorous parsing and consistency modeling, the planner produces its two structured outputs:

- **Temporal Bounding Box Sequence.** A sequence of normalized bounding boxes $[t, x_1, y_1, x_2, y_2]$, denoted $\{b_t\}_{t=1}^{T}$, is generated, aligned with all keyframes. This sequence specifies *where to edit* and ensures plausible and coherent locations.
- **Structured Enriched Instruction.** The planner further produces an augmented textual instruction that preserves the original user intent while injecting attributes discovered during reasoning, such as key spatial relations, likely interactions, motions, and scene-level constraints.

This dual-channel guidance decouples rigid spatial constraints (boxes) from flexible semantic control (enriched instruction): the former provides verifiable spatial anchors and the latter supplies transferable semantic and physical priors, forming a robust prior for mask generation and diffusion editing.

3.2. Guide: Box-Guided Mask Generation with Spatiotemporal Constraints

Given the planner's outputs, namely the bounding box sequence $B = \{b_t\}_{t=1}^{T}$ and the enriched instruction EI, we transform mask generation from a difficult global problem into a local refinement problem under explicit positional priors. Concretely, this branch treats B as hard constraints and uses video features together with EI as soft constraints to generate a binary mask M. The resulting spatiotemporal features at layer l, denoted C_l^M, are used as implicit spatial guidance for the downstream editing module. Leveraging the accurate spatial anchors provided by the planner, the mask predictor no longer needs to solve a hard localization-and-generation problem, and instead only needs to infer the precise shape required by the editing operation within the specified regions.

Spatiotemporal self-attention and cross-modal constraints. The mask prediction branch (i.e., the Guide) shares the same hierarchical architecture as the Editor. Self-attention is propagated temporally and across neighboring scales. Cross-attention connects to two types of conditions: global video features that compensate for context, and the planner's outputs (EI and B) that inject semantic constraints and accurate spatial position priors. This design greatly reduces the difficulty of mask prediction, allowing the generated masks to correctly capture the true spatial footprint of the target and its motion or contact relationships, thereby yielding semantically meaningful mask outputs.

Reverse-Connector to Editor Branch. Masks generated solely from bounding boxes and local features can be unstable for thin structures or heavily occluded regions. To address this, we introduce a Reverse-Connector from the Editor back to the Guide:

$$C_l^M = C_l^M + \text{Reverse\_Connector}(Q_l^E), \tag{2}$$

where Q_l^E denotes the editor features at layer l. This operation performs a backward correction based on the editor's understanding. Intuitively, the semantic understanding of the editing requirements helps the mask branch recover details that are otherwise easy to miss, and the connection is multi-layer and repeatable.

Mask-Connector and Explicit Mask Output. For the Guide branch, we use a *Mask-Connector* to map the refined spatiotemporal features to mask maps that share the same spatial resolution as the input frames. Concretely, we apply a per-frame decoding head, enforce constraints from the bounding boxes, preserve the temporal order to maintain inter-frame continuity of the masks, and finally output a set of masks $\{M_t\}$ that can be directly supervised. These masks can be used as explicit spatial conditions for diffusion-based editing and can also participate in joint optimization during training, stabilizing the learning of the backbone.

In parallel, we inject the mask information back into the

Editor branch using

$$Q_l^E = Q_l^E + \text{Mask\_Connector}(C_l^M), \quad (3)$$

where Mask_Connector$(\cdot)$ projects the mask features C_l^M at layer l to an additive modulation of the editor features Q_l^E. By explicitly injecting the current-frame mask into several layers, the editor features at different depths remain aware of low-level spatial guidance.

With these design choices, the mask branch maps instructions to temporally consistent masks: the planner resolves semantic ambiguities, bounding boxes anchor spatial uncertainty, and spatiotemporal mechanisms mitigate temporal instability, thereby providing an accurate guiding condition for the editing branch.

3.3. Editor: Diffusion-Based Editing with Multi-Condition Fusion

Given the precise spatiotemporal masks $\{M_t\}$ and the structured enriched instruction EI, the Editor branch performs appearance modification and content generation. Its goal is to inject these conditions into the diffusion generator, preserving unedited regions and producing visually natural, plausible, and smooth results.

Conditioned diffusion backbone. In this branch, cross-attention modules receive the enriched instruction EI. In particular, we take the last-layer visual-language features V from Qwen-VL and inject them into a dedicated Qwen-VL cross-attention module. We keep only the first 1024 tokens of V and map their channel dimension from 3584 down to the 3072 channels used by Wan2.2 5B through a multilayer perceptron (MLP). We then augment each editing block with a specialized cross-attention to these MLLM features:

$$Q_l^E = Q_l^E + \text{QVLcrossattn}(\text{MLP}(V), C_l^M), \quad (4)$$

where QVLcrossattn$(\cdot,\cdot)$ denotes cross-attention from the editor features Q_l^E to the projected MLLM tokens, modulated by the mask features C_l^M. In this way, the Editor not only receives visual priors from the original video but also fully exploits the semantic and world knowledge that the MLLM has already organized in the planning stage.

Bidirectional collaboration with the mask branch. The masks are not fed into the Editor only once; instead, the Editor and Guide cooperate bidirectionally through multi-layer interactions. As described in the previous section, the Reverse-Connector feeds high-level semantics from the Editor back to the mask branch, while the Mask-Connector injects the mask into multiple editor layers, ensuring multi-depth features remain aware of this low-level guidance. This bidirectional coupling improves the stability of the generated results in terms of boundaries, occlusions, and lighting, which is especially beneficial for fine-grained local edits.

Framework Loop. In the *Plan–Guide–Edit* framework, the Planner first parses ambiguous language into ordered boxes and an enriched instruction. Anchored by these boxes, the Guide produces spatiotemporal masks, which it refines using feedback from the Editor. The Editor then injects both masks and instructions into the diffusion backbone to perform conditional generation. Finally, a mask-guided compositing strategy applies edited content inside the mask regions while preserving the original video outside, ensuring a temporally coherent, semantically faithful, and physically compatible output.

4. Experiments

4.1. Experimental Settings

Implementation details. We adopt a two-stage training strategy. In the first stage, the Editor branch and the Mask (Guide) branch are trained separately. The *Mask branch* is trained on a mixture of image segmentation datasets (ADE20K [52], PhraseCut [43]) and video segmentation datasets (YouTube-VOS [30], OVIS [28]). The *Editor branch* is trained on two types of data: (i) open-source image/video editing datasets AnyEdit [15], UltraEdit [51], SEED-Data-Edit [7], EditWorld [47]; (ii) open-source video instruction editing datasets Señorita-2M [53], Ditto [1], together with our own instruction-based video editing dataset.

In the second stage, we jointly train the two branches on our internal dataset, which contains about 100k high-quality video editing pairs (before/after) with precise mask annotations. We train the model at a resolution of 720×1280. The first stage runs for 20k steps and the second stage for 10k steps, with batch size 64 for both stages. The Reverse-Connector and Mask-Connector are initialized with zeros.

Evaluation metrics. We evaluate our method from two perspectives: overall visual quality of the edited videos and instruction-following capability. For visual quality, we use FVD [36] and the VBench [14] metrics Background Consistency (BC), Temporal Consistency (TC), Motion Smoothness (MS), and Aesthetics (AES). For instruction following, we employ CLIPScore [8] and use the Gemini [35] model to rate the edited videos in terms of physical plausibility, spatial relations, instruction following, and overall editing quality.

- **Background Consistency (BC).** We compute cross-frame similarity in the CLIP feature space to measure temporal consistency of the global background.
- **Temporal Consistency (TC).** Following VBench, we use CLIP-B to compute similarities between each frame and its adjacent frames, as well as the first frame, to assess temporal coherence.
- **Motion Smoothness (MS).** Motion smoothness measures the continuity and naturalness of subject or cam-

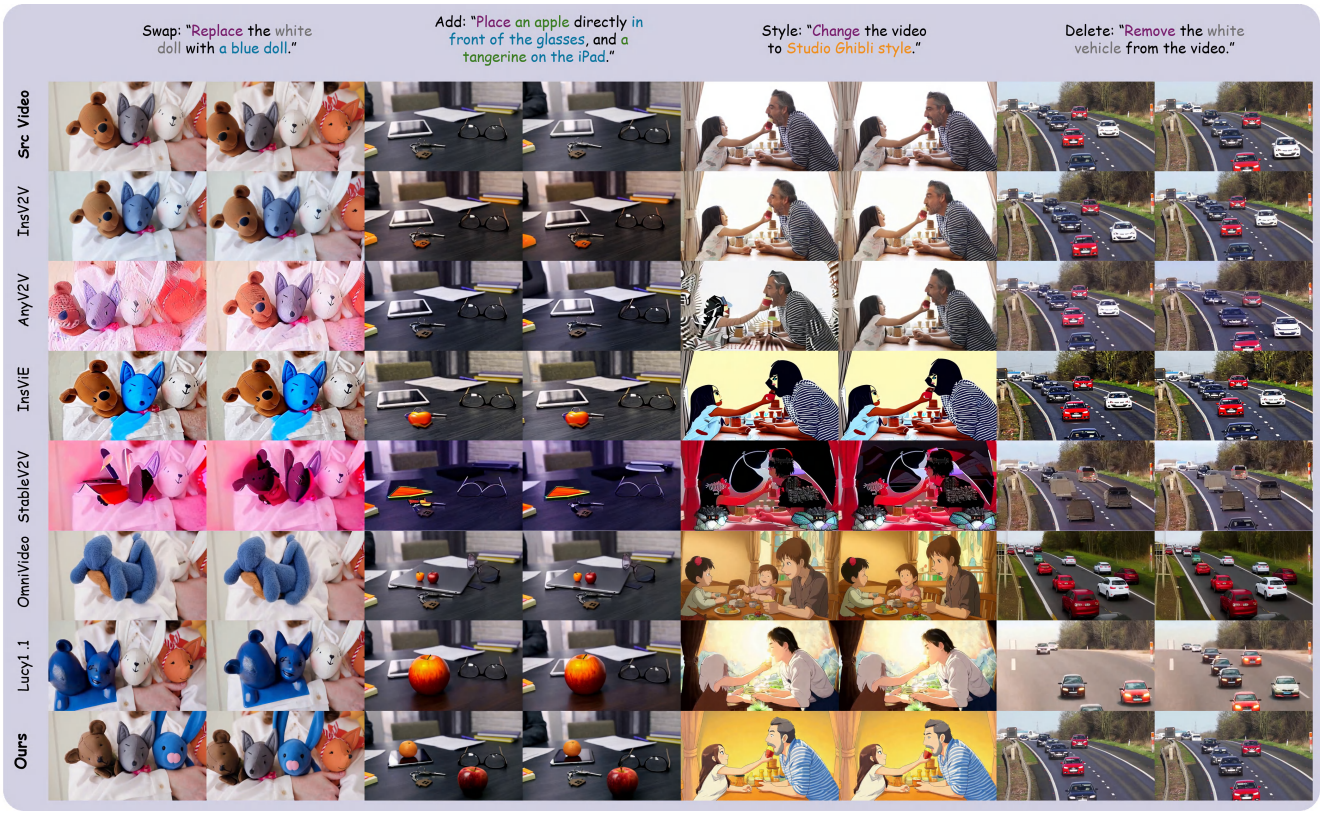

Figure 4. Qualitative comparison of CoT-Edit and open-source baselines, showing more precise target localization, better preservation of non-edited regions, and more faithful handling of complex, physics-aware editing instructions.

era motion. A normal video should avoid jitter and unnatural acceleration changes. We adopt motion priors from a video frame interpolation model to evaluate motion smoothness of edited videos.

- **Aesthetics (AES).** We use the LAION aesthetic predictor to score the aesthetic quality of each frame.
- **CLIPScore.** We use CLIP-B to measure the alignment between the text instruction and the generated video.

Evaluation protocol. We randomly sample 100 videos from the Koala-36M [39] dataset and assign diverse editing instructions, including object addition, deletion, attribute modification, stylization, and physically plausible motions such as parabolic trajectories.

4.2. Comparison with State-of-the-Art Methods

Baselines. We compare our proposed CoT-Edit with several state-of-the-art open-source instruction-based video editing methods, including InsV2V [4], StableV2V [22], InsViE [45], AnyV2V [19], OmniVideo [33], and Lucy-1.1 [34]. OmniVideo [33] combines MLLMs with diffusion models in a unified framework, while InsViE [45] and Lucy-1.1 [34] perform end-to-end video editing purely with diffusion models. However, due to their relatively weak instruction injection mechanisms, these methods are limited in modeling physical laws, spatial relations, instruction following, and overall editing quality.

Qualitative results. We present qualitative comparisons between CoT-Edit and other open-source methods in Fig. 4. CoT-Edit demonstrates: (1) precise localization and editing of the target object specified in the instruction among multiple similar objects; (2) strong preservation of non-edited regions consistent with the original video; (3) robust understanding of complex editing instructions and high-quality multi-task editing. Moreover, CoT-Edit supports instruction-based video edits that explicitly follow physical laws, such as parabolic motion or uniform linear motion. More results are shown in the supplementary material.

Quantitative results. We quantitatively compare CoT-Edit with baseline models using multiple automatic metrics, summarized in Table 1. For visual quality, we report FVD [36] and VBench [14] metrics (BC, TC, MS, AES). For editing quality, we rely on Gemini [35] to score physical plausibility, spatial relations, instruction following, and overall editing quality of the edited videos. CoT-Edit consistently outperforms all baselines across most dimensions, with particularly large gains in spatial reasoning and physical law understanding.

Table 1. Quantitative comparison of CoT-Edit and baseline models on visual and editing quality metrics.

MODEL	BC ↑	CLIPScore ↑	FVD ↓	TC ↑	MS ↑	AES ↑	Physical Rule ↑	Spatial Relation ↑	Instruction Following ↑	Editing Quality ↑
InsV2V	0.921	0.129	4095.42	0.958	0.95	0.57	0.367	0.291	0.401	0.326
StableV2V	0.942	0.263	3222.18	0.925	0.99	0.46	0.305	0.341	0.381	0.319
InsViE	0.936	0.395	2397.65	0.957	1.10	0.48	0.389	0.285	0.355	0.369
AnyV2V	0.927	0.251	2942.77	0.918	0.95	0.41	0.313	0.373	0.265	0.394
OmniVideo	0.972	0.382	1076.44	0.954	1.04	0.44	0.590	0.641	0.526	0.604
Lucy-1.1	0.931	0.376	1488.12	**0.961**	0.98	0.52	0.488	0.769	0.562	0.589
Ours	**0.974**	**0.445**	**1015.67**	0.945	**1.18**	**0.62**	**0.741**	**0.841**	**0.629**	**0.648**

Table 2. Ablation study results on video editing. Higher is better. PR, SR, IF, and OEQ denote physical plausibility, spatial relations, instruction following, and overall editing quality, respectively.

	PR	SR	IF	OEQ
E w/o MLLM	0.501	0.754	0.575	0.598
E w/ MLLM	0.674	0.758	0.598	0.631
E + M w/ Mc	0.643	0.807	0.586	0.638
E + M w/ Mc & Rc	0.681	0.815	0.609	0.647

4.3. Further Analysis and Ablation Studies

CoT ablation. In Section 3.1, we introduce the CoT scheme, which decomposes instruction understanding into intermediate reasoning steps and outputs both the editing region locations and an enhanced instruction text. To verify its effectiveness, we compare four variants: (i) Qwen3-VL-32B; (ii) Qwen3-VL-32B with CoT; (iii) Gemini 2.5 Pro [35]; (iv) Gemini 2.5 Pro with CoT. Qualitative results in Fig. 5 show that, without CoT, the ping-pong ball fails to exhibit physically plausible motion. In contrast, with CoT, both variants generate a clear parabolic motion where the ball first follows a parabolic trajectory, hits the table, and then bounces upwards, closely following physical laws. These results indicate that our CoT scheme not only improves target object localization but also enables physically consistent motion generation.

Mask branch ablation. The proposed Mask branch is responsible for task decomposition and region localization. It provides implicit mask signals to guide the Editor branch on where to edit, thus reducing the learning burden of the Editor. In the Editor branch, we further introduce Qwen3-VL-8B as an instruction enhancement module: the original video and instruction are first processed by Qwen3-VL, and the extracted features are injected into the Editor to improve instruction understanding. We compare four configurations corresponding to those in Table 2: (i) E w/o MLLM, the Editor branch without Qwen3-VL; (ii) E w/ MLLM, the Editor branch with Qwen3-VL features; (iii) E + M w/ Mc, the Editor plus the Mask branch with Mask-Connector; (iv) E + M w/ Mc & Rc, the full model with both Mask-Connector and Reverse-Connector. Experimental results in Table 2 show that Qwen3-VL significantly improves instruction-following performance, while the Mask branch enhances the Editor's understanding of spatial locations.

Figure 5. Qualitative results of the CoT ablation study.

User study. To complement automatic metrics, we conducted a user study assessing perceptual quality and physical consistency. Participants ranked CoT-Edit against baseline methods on visual quality and physical consistency. The results showed a clear preference for CoT-Edit in all aspects. Detailed protocols and statistics are provided in the supplementary material.

5. Conclusion

We presented a *Plan–Guide–Edit* framework where a CoT-enhanced MLLM planner, a box-guided mask branch, and a diffusion editor link semantic intent to spatial execution. By explicitly modeling spatial constraints and priors, our method achieves precise localization, temporal coherence, and preservation of unedited regions. Experiments show consistent gains over prior text-driven methods in spatial reasoning and visual quality.

Acknowledgments

This work was supported in part by NSFC under Grant 62371434, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20252293, the China Postdoctoral Science Foundation-Anhui Joint Support Program under Grant Number 2024T017AH, China Postdoctoral Science Foundation under Grant Number 2025M783529, Anhui Postdoctoral Scientific Research Program Foundation (No.2025A1015), the Fundamental Research Funds for the Central Universities(No. WK2100250064), and ZGCA Project-C20250302.

References

[1] Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025. 3, 6

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 1, 3

[3] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025. 2

[4] Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*, 2023. 3, 7

[5] Xin Ding, Lei Yu, Xin Li, Zhijun Tu, Hanting Chen, Jie Hu, and Zhibo Chen. Rass: Improving denoising diffusion samplers with reinforced active sampling scheduler. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12923–12933, 2025. 2

[6] Yixin Gao, Xiaohan Pan, Xin Li, and Zhibo Chen. Why compress what you can generate? when gpt-4o generation ushers in image compression fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 371–381, 2025. 2

[7] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 6

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021. 6

[9] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[11] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[12] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8153–8163, 2024. 2

[13] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025. 2

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6, 7

[15] Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025. 6

[16] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17191–17202, 2025. 2

[17] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 2

[18] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[19] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 7

[20] Sen Liang, Zhentao Yu, Zhengguang Zhou, Teng Hu, Hongmei Wang, Yi Chen, Qin Lin, Yuan Zhou, Xin Li, Qinglin Lu, et al. Omniv2v: Versatile video generation and editing via dynamic content manipulation. *arXiv preprint arXiv:2506.01801*, 2025. 2

[21] Cuiyu Liu, Wei Zhai, Yuhang Yang, Hongchen Luo, Sen Liang, Yang Cao, and Zheng-Jun Zha. Grounding 3d scene affordance from egocentric interactions. *arXiv preprint arXiv:2409.19650*, 2024. 2

[22] Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stabilizing shape consistency in video-to-video editing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3, 7

[23] Dingyuan Liu, Qiannan Shen, and Jiaci Liu. The health-wealth gradient in labor markets: Integrating health, insurance, and social metrics to predict employment density. *Computation*, 14(1):22, 2026. 2

[24] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 3

[25] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2

[26] Yue Ma, Xiaodong Cun, Sen Liang, Jinbo Xing, Yingqing He, Chenyang Qi, Siran Chen, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9385–9395. IEEE, 2025. 2

[27] Chong Mou, Qichao Sun, Yanze Wu, Pengze Zhang, Xinghui Li, Fulong Ye, Songtao Zhao, and Qian He. Instructx: Towards unified visual editing with mllm guidance. *arXiv preprint arXiv:2510.08485*, 2025. 1, 3

[28] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 6

[29] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 1, 3

[30] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pages 208–223. Springer, 2020. 6

[31] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[32] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022. 2

[33] Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation. *arXiv preprint arXiv:2507.06119*, 2025. 7

[34] DecartAI Team. Lucy edit: Open-weight text-guided video editing, 2025. 1, 3, 7

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6, 7, 8

[36] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6, 7

[37] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[38] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[39] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 7

[40] Xingrui Wang, Xin Li, Yaosi Hu, Hanxin Zhu, Chen Hou, Cuiling Lan, and Zhibo Chen. Tiv-diffusion: Towards object-centric movement for text-driven image to video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7988–7996, 2025. 2

[41] Yuanzhi Wang, Yong Li, Mengyi Liu, Xiaoya Zhang, Xin Liu, Zhen Cui, and Antoni B Chan. Re-attentional controllable video diffusion editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8123–8131, 2025. 3

[42] Yuanzhi Wang, Yong Li, Mengyi Liu, Xiaoya Zhang, Zhen Cui, and Jian Yang. Training-free controllable text-guided video editing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026. 3

[43] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 6

[44] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7623–7633, 2023. 3

[45] Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16692–16701, 2025. 1, 3, 7

[46] Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. Veggie: Instructional editing and reasoning video concepts with grounded generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15147–15158, 2025. 3

[47] Bohan Zeng, Ling Yang, Jiaming Liu, Minghao Xu, Yuanxing Zhang, Pengfei Wan, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-

following image editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12674–12681, 2025. 6

[48] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[49] Zhenghao Zhang, Zuozhuo Dai, Long Qin, and Weizhi Wang. Effived: Efficient video editing via text-instruction diffusion models. *arXiv preprint arXiv:2403.11568*, 2024. 3

[50] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27957–27967, 2025. 2

[51] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024. 6

[52] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International journal of computer vision*, 127(3):302–321, 2019. 6

[53] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Se\~ norita-2m: A high-quality instruction-based dataset for general video editing by video specialists. *arXiv preprint arXiv:2502.06734*, 2025. 6