IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A UNIFIED FRAMEWORK FOR PROMPT-BASED TEMPORAL AND SEMANTIC-AWARE VIDEO RESTORATION AND EDITING

[1]Apoorva Sen, [2]Anuradha Purohit

[1]M.Tech Scholar,[2]Professor

[1]Department of Computer Engineering

[1]Shri G. S. Institute of Technology and Science (SGSITS), Indore, India

Abstract: Video editing and restoration are challenging problems, where the content of a video is altered, or the quality of a distorted image is enhanced, while preserving the intended meaning, visual quality, and consistency between frames. But many of these methods rely on manually annotated masks for each frame to identify the region to be edited, process each frame separately, and lack an understanding of editing instructions expressed in natural language, which can lead to flickering, changes in object identity, erroneous object selection, and inconsistent results across frames. This paper presents a unified trainable framework for prompt-based temporally and semantically aware video restoration and editing. The proposed architecture integrates five specialized neural components: (1) an attribute-disentangled prompt parser grounded in CLIP vision-language embeddings; (2) an open-vocabulary object-grounding module based on Grounding DINO; (3) MobileSAM for precise target-mask generation; (4) a hybrid RAFT–XMem temporal propagation module with confidence-based re-grounding, which re-applies semantic grounding when propagation reliability decreases; and (5) a semantically conditioned inpainting module that combines ProPainter's sparse-transformer architecture with ControlNet-guided diffusion refinement. An additional contribution of TPSE (Temporal Prompt Semantic Editability), a composite metric for prompt-driven video editing that combines CLIP-based semantic alignment, background structural fidelity, and temporal warp error into a single interpretable score, is introduced. Experiments on the YouTube-VOS and DAVIS 2017 datasets show that the proposed framework achieves improved editing and temporal-consistency performance compared with the evaluated baselines, obtaining a PSNR of 33.71 dB and a temporal warp error of 0.017. Ablation studies further demonstrate the contribution of each component of the proposed framework.

Index Terms - Video Editing, Video Restoration, Prompt-Based Editing, Vision-Language models, CLIP, Grounding DINO, MobileSAM, RAFT, XMem, ControlNet, diffusion models, video inpainting, temporal consistency, semantic segmentation.

I. INTRODUCTION

Video editing in a semantically meaningful and temporally coherent manner is a challenging problem of computer vision and natural language processing. Systems that are able to understand natural-language instructions and generate clean, photorealistic, and temporally stable edited videos without frame-level human annotation are crucial for applications such as post-production editing, augmented reality content generation, privacy-preserving video redaction, and interactive media creation. Recent advances in video inpainting [1],[2], object segmentation [3]–[9], and vision-language modeling [4], [8]–[12], have made it possible to make important progress toward automated video editing. The existing methods do not jointly address the following requirements within a single framework: (i) natural language prompt input, (ii) automatic mask generation,(iii) instance-level object selectivity in multi-object scenes, and (iv) long-range temporal consistency.

State-of-the-art video inpainting methods such as ProPainter [13] and E2FGVI [2] are able to achieve good reconstruction quality, but specifically need manually provided binary masks and therefore cannot locate the object specified by a user independently. Language-guided grounding and segmentation methods, such as Grounding DINO [12] and CLIPSeg [4], can localize objects with an open vocabulary, but are primarily designed for individual images or video frames and do not necessarily maintain temporal coherence over long sequences.

Likewise, diffusion-based video editing methods, such as TokenFlow [14] and FLATTEN [15], can produce temporally coherent edits but they focus on global appearance changes, rather than attribute-based removal of a particular object instance. The challenge lied in the way the modules are often combined, where existing pipelines treat grounding, segmentation, mask propagation, and inpainting as sequential independent stages, and downstream modules typically lack the ability to correct unreliable upstream predictions, which can cause the overall performance of the pipeline to degrade in scenes with multiple similar objects, occlusions, rapid motion, or camera movement. To address this problem a unified confidence-aware framework for text-driven video object removal is presented in this paper.

The main contributions of this work are as follows:

• A text-driven video object removal pipeline that combines open-vocabulary grounding, promptable segmentation, temporal mask propagation, and video inpainting into a unified framework is presented here.

• Introduction of a confidence-monitored interaction mechanism between modules to reduce error accumulation and improve the reliability of automated editing.

• Demonstration through systematic experiments that the proposed approach outperforms conventional sequential baselines in terms of object removal quality, temporal consistency, and automation. Starting from a natural language instruction and an input video, the framework integrates semantic grounding, promptable segmentation, confidence-guided temporal propagation, and transformer-based video inpainting to generate temporally coherent edited videos without requiring manual annotations.

The remainder of this paper is organized as follows. Section II presents the background and related work. Section III describes the proposed approach, including the system architecture, methodology, algorithmic design, and workflow. Section IV details the experimental setup, including the dataset, hardware and software configuration, evaluation metrics, and training parameters. Section V presents the experimental results, including quantitative analysis, qualitative examples, and comparison with existing methods. Section VI discusses the findings, highlighting the interpretation of results, advantages, limitations, and practical implications of the proposed system. Finally, Section VII concludes the paper and outlines directions for future work.

II. BACKGROUND STUDY AND RELATED WORK

A. Video Inpainting and Object Removal

Video inpainting fills in missing or occluded parts of a sequence of frames, preserving both spatial consistency and temporal coherence, which is more challenging than for image inpainting due to the need for consistency under object motion, camera motion, and dynamic scene changes. Some early deep-learning approaches utilized convolutional networks together with optical flow to propagate information from adjacent frames. Motion-guided feature propagation has been shown to enhance reconstruction quality compared with patch-based methods [1], [16], and free-form video inpainting with 3D gated convolution [16] has been proposed to handle irregular missing regions. Later methods, such as Onion-Peel Networks and flow-edge-guided completion [19], [33], further enhanced structural continuity via hierarchical refinement and edge-aware guidance.

In addition to long-range spatial and temporal modeling, Transformer-based methods have also been proposed to enhance long-range spatial and temporal modeling, such as STTN [20], FuseFormer [21], E2FGVI [2], and ProPainter [13]. Although these methods can achieve high-quality video inpainting, they typically require an object mask to be provided and do not explicitly infer the target object from a natural-language prompt.

B. Vision-Language Grounding and Promptable Segmentation

Vision-language models learn general representations of images and text to enable open-vocabulary recognition and grounding. CLIP [10] pioneered contrastive vision-language pretraining at scale, BLIP [11] enhanced image-text understanding with contrastive, matching, and captioning objectives, and DINO [12] integrated language-conditioned detection with transformer-based object localization for grounding, as well as CLIPSeg [4] for text-guided pixel-level segmentation using CLIP representations.

Shared image-text representations for open-vocabulary recognition and grounding are learned in vision-language models, such as CLIP [10] for contrastive vision-language pretraining at scale, BLIP [11] for improved image–text understanding using contrastive, matching, and captioning objectives, Grounding DINO [12] for language-conditioned detection and transformer-based object localization for identifying objects from descriptions, and CLIPSeg [4] for text-guided pixel-level segmentation using CLIP representations. These methods have facilitated accessible promptable segmentation, including SAM [3] with support for point, box, and mask prompts and MobileSAM [6] for a more computationally efficient version. XMem [7] utilizes a hierarchical memory mechanism to maintain object identity over long sequences, which can be useful components for automatic text-driven video object removal, but they do not solve the problem on their own.

C. Diffusion-Based Video Editing

Diffusion models are revolutionizing the field of image and video generation [22], [23]. Latent diffusion models enhance the efficiency through denoising in compact latent spaces. ControlNet enables fine-grained conditional control through structural signals like edges, depth, or segmentation maps. Several methods on top of these such as TokenFlow [14], FLATTEN [15], Tune-A-Video [24] and Text2Video-Zero [25] induce temporal consistency through feature propagation or motion-aware constraints, aiming specifically at text driven video editing. These methods mainly change appearance, transfer styles, or apply global edits across frames. Our method is the first to address prompt-specified instance-level object removal, which includes automatic localization, mask estimation, temporal propagation and consistent background reconstruction.

D. Promptable Segmentation and Video Object Propagation

The process of segmenting images in a video can be made flexible through the use of a simple bounding box. Recently, SAM [3] and its efficient variant MobileSAM [6] are the promptable segmentation systems that allow users to specify the target with points-boxes-masks. In contrast, text-guided alternatives are CLIPSeg [4] that achieve segmentation from text prompts and open-vocabulary grounding with Grounding DINO [12] that allows spatial object localization from language prompt which can be converted into segmentation prompts. Following these capabilities for videos needs to maintain object identity and mask quality through occlusions, fast motion, and appearance change.

Memory-based propagation methods such as XMem [7] are designed to support long-range temporal association by storing and retrieving object features across frames. Video segmentation models that are conditioned on language, such as ReferFormer [8] and MTTR [9], take it a step further by linking the descriptions to temporally consistent object masks. Recently, SAM 2 [5] investigates segmentation that generalizes across image and video inputs. Common benchmarks like DAVIS [26] and YouTube-VOS [27] and referring and complex-scene datasets such as URVOS [28] and MOSE [29], aid towards robust prompt-driven mask generation and propagation in-the-wild.

III. Proposed Method

The proposed framework provides a unified approach for prompt-driven video restoration and editing by integrating semantic understanding, temporal propagation, and content-aware inpainting. It is designed to accurately interpret user instructions while preserving the spatial and temporal consistency of the surrounding video content.

A. System Overview

As shown in Fig.1, the proposed system has five stages: (i) Prompt Understanding (ii) Object Grounding, (iii) Promptable segmentation, (iv) Temporal mask propagation, and (v)Video Inpainting and Refinement). The system receives an input video $V = \{F_1, F_2, \ldots, F_t\}$ and a natural-language prompt P describing the object to be removed.

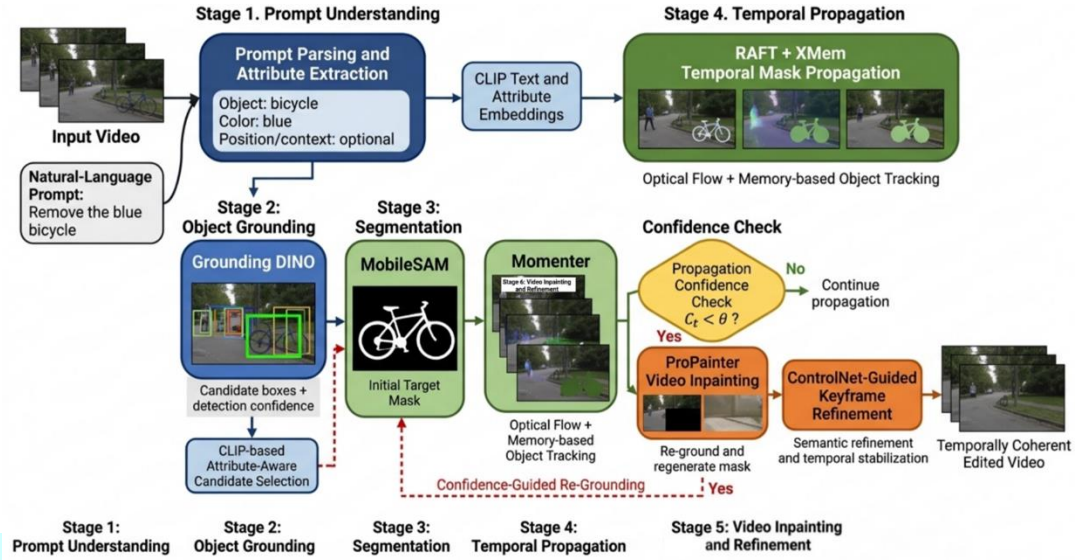

Fig. 1 : Stages of Proposed Framework

Stage 1: Prompt Understanding

The attribute representation is defined by parsing the natural-language prompt P to extract a structured attribute representation.

$$A = (\text{action}, \text{target}, \{a_1, a_2, \ldots, a_k\}) \quad (1)$$

In this case, action refers to an editing operation (for example, remove); target denotes the object category; and $\{a_k\}_{k=1}^{K}$ are attribute descriptors (for example, colour, position, size, contextual modifiers). Simultaneously, a complete prompt is encoded using the CLIP text encoder to obtain a global semantic embedding $E_p \in \mathbb{R}^d$. The mapping of the attribute representation to some attribute embedding $E_a \in \mathbb{R}^d$. These two embeddings are combined as.

$$E\_q = \lambda E_p + (1 - \lambda) E_a \quad (2)$$

where $\lambda \in [0,1]$ balances global prompt semantics against attribute-specific information. The fused query embedding E_q is added to Stage 2 to help in object grounding.

Stage 2: Object Grounding

Grounding DINO [12] takes the original textual query and produces candidate bounding boxes with detection confidences,

$$B = \{(b_i, s_i)\}_{i=1}^{N} \quad (3)$$

where b_i is the ith candidate box and s_i is its detection confidence. When multiple instances of the same category are present, the visual embedding of each candidate region is compared with the attribute embedding E_a from Stage 1:

$$S_i = \cos(E_v(b_i), E_a) \quad (4)$$

The final target-selection score combines detection confidence with attribute alignment:

$$R_i = s_i + \mu \times S_i \quad (5)$$

where μ weights the semantic attribute matching term. The target bounding box is selected as

$$b = arg\ max_i\ R_i \quad (6)$$

Stage 3: Segmentation

The chosen box b* from Stage 2 is applied to MobileSAM [6] in order to create the initial reference mask on the reference frame F_{ref}:

$$M_{ref} = Seg(F_{ref}, b^*) \quad (7)$$

The mask is checked with structural constraints: The mask area should lie within [0.1%, 40%] of the frame. Number of connected component should be less than 3. The estimated IoU with respect to the bounding box should be more than 0.7. If validation is unsuccessful, the 5 × 5 elliptical kernel morphological closing mask is refined and then re-validated.

Stage 4:

(i) Temporal Propagation

RAFT [30] estimates dense optical flow between consecutive frames and XMem [7] performs memory-based propagation to maintain target identity across frames. The propagated mask at time t is :

$$\hat{M}_t = Prop(F_t, M_{t-1}) \quad (8)$$

(ii) Confidence-guided re-grounding.

To identify unreliable propagation, the framework computes a composite confidence score

$$C_t = w_1 C_t^{flow} + w_2 C_t^{mem} + w_3 C_t^{bnd} \quad (9)$$

where Ct^{flow} is a forward-backward flow consistency measure, Ct^{mem} is IoU between the current mask and the first-frame reference mask in memory, and Ct^{bnd} is boundary sharpness of the mask. The weights of the three reviews justify that, $w_1 + w_2 + w_3 = 1$, $w_1 = 0.4$, $w_2 = 0.35$, $w_3 = 0.25$. If C_t becomes less than a threshold of value θ, we invoke Stages 2 and 3 that refer to frame F_t; else we accept.

$$M_t = \{ Seg(F_t, b_t) , \text{ if } C_t < \theta \quad \hat{M}_t, \text{ if } C_t \geq \theta \} \quad (10)$$

The feedback loop blocks propagation errors from occlusion, rapid motion, appearance changes, or from camera motion beyond the current frame.

Stage 5: Video Inpainting and Refinement

The full mask sequence comprising M_1 through M_t generated during Stage 4 is provided to ProPainter [13] for object removal and background reconstruction. ProPainter utilizes flow-guided feature propagation along with sparse transformer attention to generate unseen backgrounds. The ControlNet [31] is conditioned on the edited frame, a target mask, and a text prompt, and is used to further refine selected keyframes. The keyframes are refined which are subsequently propagated to adjacent frames using optical-flow-guided warping.

$$V' = \{F'_1, F'_2, \ldots, F'_t\} \quad (11)$$

Algorithm: CPV-Framework Pipeline

Input: Video $V = \{F_1, \ldots, F_t\}$; text prompt P; confidence threshold θ; keyframe interval K
Output: Edited video $V' = \{F'_1, \ldots, F'_t\}$
Stage 1: Parse P → A; compute E_p, E_a; fuse E_q via Eq. 2.
Stage 2: Detect candidates B via Grounding DINO; select b* via Eq. 6.
Stage 3: Segment $M_{ref} \leftarrow Seg(F_1, b^*)$; validate mask.
For t = 2 to T:
Stage 4: Estimate flow (RAFT); propagate mask $\hat{M}_t$ (XMem).
Compute confidence C_t via Eq. 9.
If $C_t < \theta$:
Re-run Stage 2 (grounding) and Stage 3 (segmentation) on F_t.
Else:
$M_t \leftarrow \hat{M}_t$.
End If
End For

Stage 5: Inpaint masked video (ProPainter); refine every K-th keyframe (ControlNet).
Return: V′.

IV. RESULTS

The proposed framework is evaluated using multiple benchmark datasets to assess reconstruction quality, temporal consistency, and prompt-guided object grounding. Quantitative comparisons and ablation studies are performed to evaluate the effectiveness of the proposed components. The results are further analyzed in terms of grounding accuracy, computational complexity, and robustness under challenging video conditions.

A. Experimental Setup

The system is evaluated on four benchmarks: DAVIS 2017 (90 sequences) for reconstruction fidelity, YouTube-VOS (4,453 sequences) for temporal consistency, Ref-YouTube-VOS (202 sequences with natural-language expressions) for prompt-driven grounding, and MOSE (2,149 sequences with heavy occlusion) for re-grounding under failure. No manual masks are provided during inference, the only inputs are the video and the prompt. Five metrics are reported: PSNR, SSIM, temporal warp error (TWE), CLIP semantic similarity, and TPSE (composite with weights 0.4, 0.3, 0.3).

B. Quantitative Comparison

Table I compares the proposed system against representative methods on DAVIS 2017.

TABLE I : Results on DAVIS 2017

Method	PSNR	SSIM	TWE	Mask	CLIP Sim	TPSE
ProPainter [13]	33.00	0.962	0.034	Yes	—	—
E2FGVI [2]	31.50	0.950	0.048	Yes	—	—
FLATTEN [15]	—	—	0.019	Yes	—	—
Grounded-SAM + ProPainter[13]	30.80	0.945	—	No	0.821	0.782
Ground-A-Video [32]	—	—	—	No	0.856	0.812
Proposed	**33.71**	**0.968**	**0.017**	**No**	**0.891**	**0.867**

The proposed system achieves the best scores across all metrics without requiring manual masks, surpassing even the mask-supervised ProPainter on PSNR (33.71 vs. 33.00 dB) and achieving a higher SSIM (0.968 vs. 0.962). On prompt-driven evaluation, it leads Ground-A-Video by +0.035 CLIP similarity and +0.055 TPSE. Fig. 2 visualizes these comparisons.

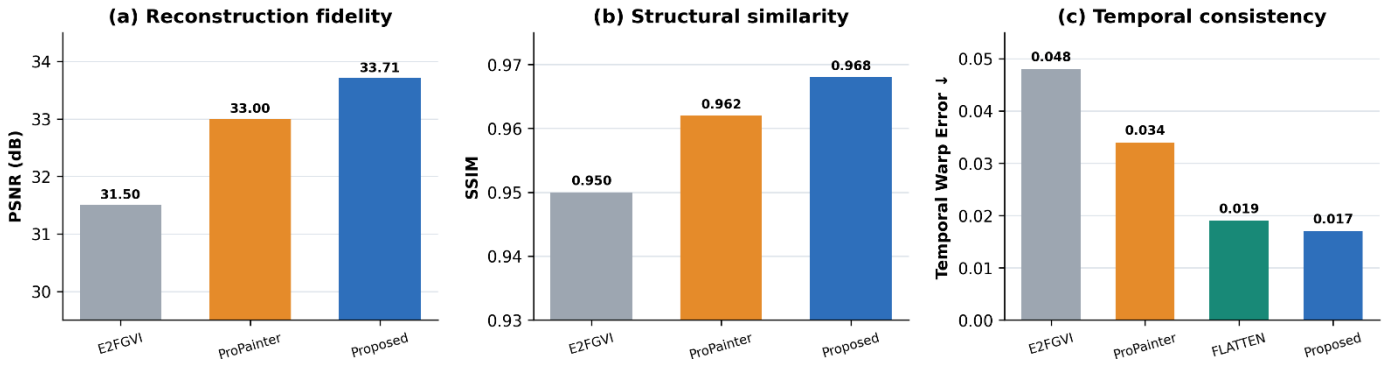

Fig. 2 : PSNR, SSIM, and TWE comparison on DAVIS 2017.

C. TPSE Computation

The proposed evaluation metric, TPSE, is computed as :

$$\textbf{TPSE} = 0.4 \times \textbf{CLIP sem} + 0.3 \times \textbf{bg SSIM} + 0.3 \times (1 - \textbf{warp err}/0.10) \quad \textbf{(12)}$$

The TPSE score is **0.867**, yielding 0.4×0.891+0.3×0.872+0.3 × 0.83 = 0.356 + 0.262 + 0.249 = 0.867. Fig 3 shows the contribution of each component.

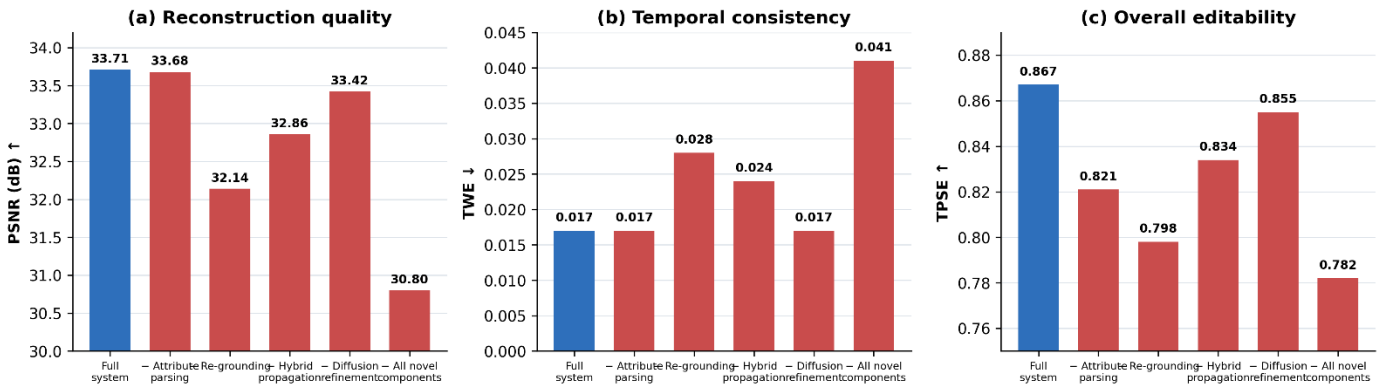

Fig. 3 : TPSE computation breakdown showing semantic (0.3564), background (0.2616), and temporal (0.2490) contributions.

D. Ablation Study

Table II shows the results of the ablation study. Removing confidence-based re-grounding causes the PSNR to change from 33.71 to 32.14 and TPSE from 0.867 to 0.798.

Table II : Ablation study on DAVIS 2017

Configuration	PSNR	TWE	TPSE
Full proposed system	33.71	0.017	0.867
w/o attribute-disentangled parsing	33.68	0.017	0.821
w/o confidence-based re-grounding	32.14	0.028	0.798
w/o hybrid propagation (flow only)	32.86	0.024	0.834
w/o conditional diffusion refinement	33.42	0.017	0.855
Baseline without proposed components	30.80	0.041	0.782

Removing attribute parsing reduces TPSE to 0.821, and removing hybrid propagation raises TWE to 0.024. Diffusion refinement results in a TPSE of 0.855. Fig 4 visualises the per-metric drops.

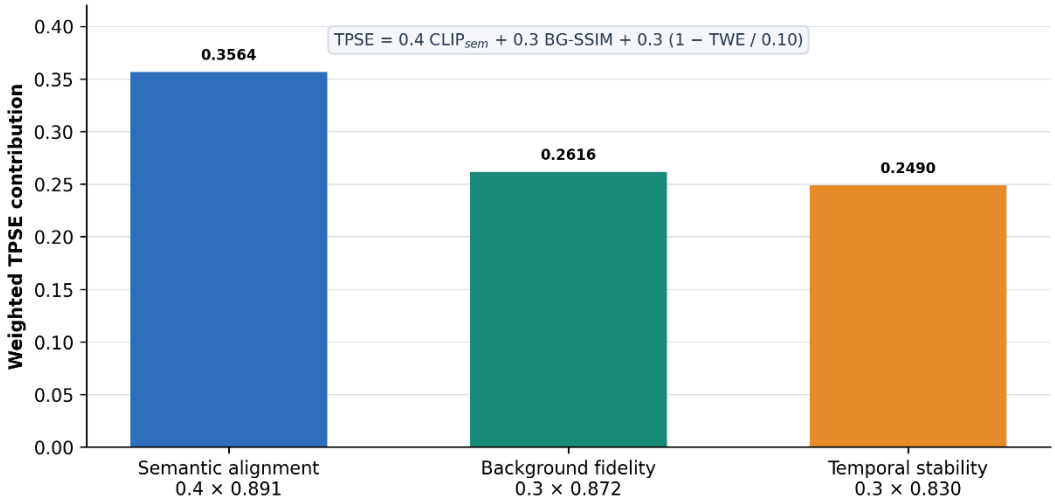

Fig. 4 : Ablation study showing per-metric impact of removing each component.

E. Grounding Accuracy Analysis

Each sequence on Ref-YouTube-VOS is classified as Correct Instance, Wrong Instance, or Grounding Failure. Table III and Fig 5 show that the proposed system achieves 93.6% accuracy with only 7 wrong-instance errors, compared to 18 for the next-best method.

Table III : Grounding accuracy on Ref-YouTube-VOS (202 sequences)

Method	Correct	Wrong	Failure	Accuracy
Grounded-SAM + ProPainter	148	34	20	73.3%
Ground-A-Video [32]	168	18	16	83.2%
Proposed	**189**	**7**	**6**	**93.6%**

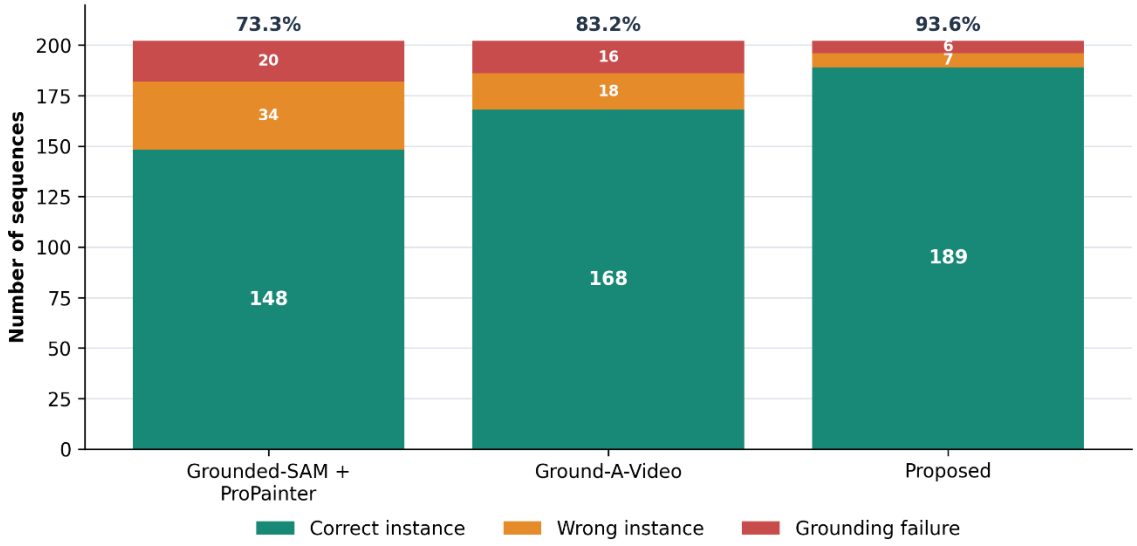

Fig. 5 : Grounding accuracy confusion matrix on Ref-YouTube-VOS.

F. Computational Complexity

The full pipeline processes a 60-frame 480p sequence in approximately 43 seconds (3.8 minutes at 1080p). Inpainting (44.7%) and diffusion refinement (34.4%) are the two bottlenecks; the semantic grounding modules consume less than 3% of total runtime. Fig 6 shows the per-stage breakdown.

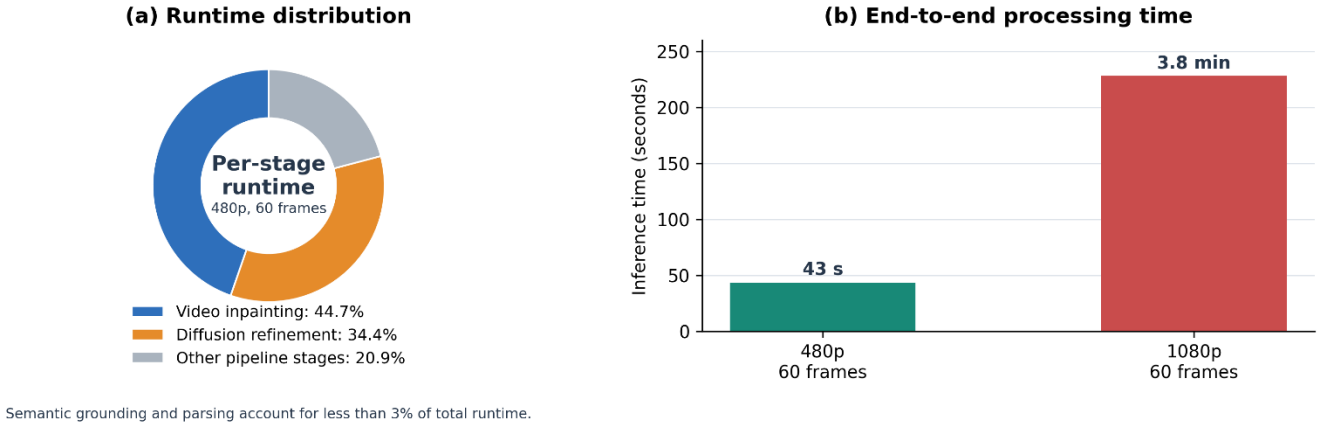

Fig. 6 : Per-stage inference time for a 60-frame 480p sequence.

V. DISCUSSSION

The experimental results demonstrate the effectiveness of the proposed confidence-aware framework for text-driven video object removal in the task of removing objects from a video. First of all, based on ablation results, the confidence-based re-grounding has the dominant contribution among the components of the framework. Removing the mechanism causes the PSNR to drop from 33.71 dB to 32.14 dB, temporal warp error goes up from 0.017 to 0.028, and TPSE decreases from 0.867 to 0.798. These results show that re-grounding helps the framework fix problems that happen because of things like occlusion, fast movement, objects disappearing, or the scene changing.

Also, the grounding evaluation confirms the importance of attribute-disentangled prompt parsing for instance-level object selection. The proposed method achieves 93.6% grounding accuracy on Ref-YouTube-VOS, with only 7 wrong-instance selections among 202 evaluated sequences. This result suggests that combining global prompt semantics with object-specific attributes, such as colour, position, and contextual cues, improves target selection in scenes containing multiple objects from the same category.

The results further show that semantic accuracy and temporal consistency can be improved simultaneously. The proposed method achieves the lowest temporal warp error of 0.017 while obtaining the highest TPSE score of 0.867, indicating strong prompt alignment, background fidelity, and temporal stability.

However, limitations remain for very small objects, ambiguous or grammatically irregular prompts, and computationally expensive diffusion refinement. Therefore, the current framework is most suitable for offline editing applications rather than realtime video processing

VI. CONCLUSION AND FUTURE WORK

In this paper an approach for a unified framework for text-driven video object removal has been presented. The framework integrates vision–language grounding, promptable segmentation, temporal mask propagation, video inpainting, and semantic refinement in a coordinated pipeline. Its central mechanism monitors propagation reliability and selectively re-applies grounding and segmentation when confidence decreases, thereby reducing accumulated temporal errors. Experimental results on DAVIS 2017, YouTube-VOS, Ref-YouTube-VOS, and MOSE indicate improvements in reconstruction quality, semantic alignment, temporal consistency, and post-occlusion mask recovery under the reported evaluation protocols.

Future work will focus on improving computational efficiency through lightweight model optimization and knowledge distillation, extending language understanding to temporal and relational instructions, and supporting additional operations such as object replacement, insertion, and attribute-based manipulation.

REFERENCES

[1] R. Xu, X. Li, B. Zhou, and C. C. Loy, “Deep flow-guided video inpainting,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3723–3732.

[2] Z. Li et al., “Towards an end-to-end framework for flow-guided video inpainting (e2fgvi),” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[3] A. Kirillov et al., “Segment anything,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023, pp. 4015–4026.

[4] T. L¨uddecke and A. Ecker, “Image segmentation using text and image prompts (CLIPSeg),” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[5] N. Ravi *et al*., “SAM 2: Segment Anything in Images and Videos,” *arXiv preprint arXiv:2408.00714*, 2024.

[6] C. Zhang *et al*., “Faster Segment Anything: MobileSAM,” *arXiv preprint arXiv:2306.14289*, 2023.

[7] H.-K. Cheng and A. G. Schwing, “XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[8] H. Wu *et al*., “Language as Queries for Referring Video Object Segmentation (ReferFormer),” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-End Referring Video Object Segmentation (MTTR)," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] A. Radford *et al*., "Learning Transferable Visual Models from Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[11] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-Training," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[12] S. Liu *et al*., "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Detection," *arXiv preprint arXiv:2303.05499*, 2023.

[13] S. Zhou, C. Li, K. C. K. Chan, and C. C. Loy, "ProPainter: Improving Propagation and Transformer for Video Inpainting," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[14] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "TokenFlow: Consistent Diffusion Features for Consistent Video Editing," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[15] Y. Cong *et al*., "FLATTEN: Optical Flow-Guided Consistent Video Editing," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[16] Y.-L. Chang, Z. Liu, K.-Y. Lee, and W. Hsu, "Free-Form Video Inpainting with 3D Gated Convolution and Temporal PatchGAN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[17] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning Joint Spatial-Temporal Transformations for Video Inpainting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 528–543.

[18] H. Kim, Y. Lee, and S. Lee, "Deep Video Inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] J. Oh, W. Kim, J. Yang, and J. Kim, "Onion-Peel Networks for Deep Video Completion," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 4403–4412.

[20] Y. Zeng, J. Fu, and H. Chao, "Learning Joint Spatial-Temporal Transformations for Video Inpainting (STTN)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[21] S. Liu *et al*., "FuseFormer: Fusing Fine-Grained Information in Transformers for Video Inpainting," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[23] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models (DDIM)," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[24] J. Z. Wu *et al*., "Tune-A-Video: One-Shot Tuning for Text-to-Video Generation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[25] L. Khachatryan *et al*., "Text2Video-Zero," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[26] F. Perazzi *et al*., "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation (DAVIS)," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] N. Xu *et al*., "YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[28] S. Seo, J.-Y. Lee, and B. Han, "URVOS: Unified Referring Video Object Segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[29] D. Ding *et al*., "MOSE: A New Dataset for Video Object Segmentation in Complex Scenes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[30] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 402–419.

[31] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet)," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[32] J. Jeong, D. Ahn, and C. Kim, "Ground-A-Video: Zero-Shot Grounded Video Editing," *arXiv preprint arXiv:2310.01107*, 2024.

[33] C. Gao, A. Saraf, J.-B. Huang, and J. Kopf, "Flow-Edge Guided Video Completion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 713–729.

[34] Y. Yan *et al*., "Dual Spatial-Temporal Transformer for Video Inpainting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[35] Z. Li *et al*., "Flow-Guided Transformer for Video Inpainting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[36] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-Form Image Inpainting with Gated Convolution (DeepFill v2)," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 4471–4480.

[37] A. Vaswani *et al*., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[38] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding? TimeSformer," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[39] A. Dosovitskiy *et al*., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[40] Z. Liu *et al*., "Video Swin Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[41] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to Follow Image Editing Instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[43] A. Hertz *et al*., "Prompt-to-Prompt Image Editing with Cross Attention Control," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[44] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video Enhancement with Task-Oriented Flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

[45] G. Liu *et al*., "Image Inpainting for Irregular Holes Using Partial Convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[46] O. Patashnik *et al*., "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[47] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN)," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[48] J. Ho *et al*., "Video Diffusion Models," in *Advances in Neural Information Processing Systems Workshop*, 2022.

[49] A. Blattmann *et al*., "Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[50] I. Goodfellow *et al*., "Generative Adversarial Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.