RACCooN: Remove, Add, and Change Video Content with Auto-Generated Narratives

Anonymous Author(s)
Affiliation
Address
email

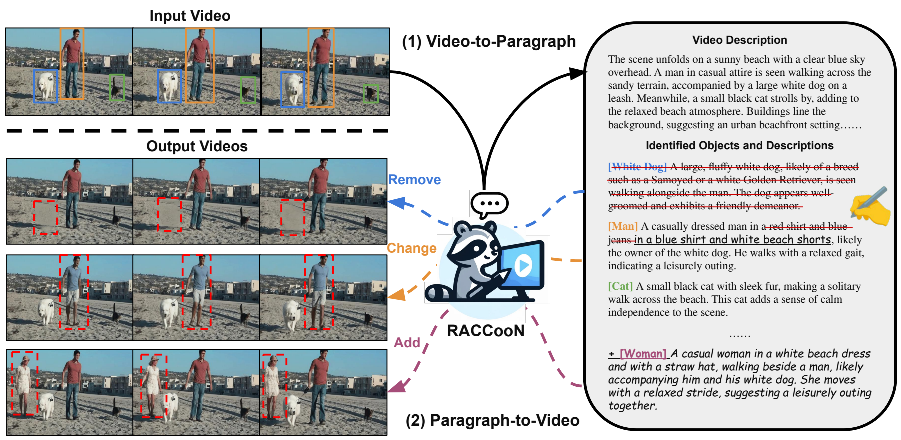

Figure 1: **Overview of RACCooN**, a versatile and user-friendly video-to-paragraph-to-video framework, enables users to remove, add, or change video content via updating auto-generated narratives.

Abstract

This paper proposes **RACCooN**, a versatile and user-friendly **video-to-paragraph-to-video** generative framework that supports multiple video editing capabilities such as removal, addition, and modification, through a unified pipeline. The proposed approach stands out from other methods through several significant contributions: (1) suggests a multi-granular spatiotemporal pooling strategy to generate well-structured video descriptions, capturing both the broad context and object details without requiring complex human annotations, simplifying precise video content editing based on text for users. (2) RACCooN incorporates auto-generated narratives or instructions to enhance the quality and accuracy of the generated content. It supports the addition of video objects, inpainting, and attribute modification within a unified framework, surpassing existing video editing/inpainting benchmarks by demonstrating impressive versatile capabilities in video-to-paragraph generation (up to $9.4\%p \uparrow$ absolute improvement in human evaluations against the baseline), video content editing (relative $49.7\% \downarrow$ in FVD).

1 Introduction

Despite advancements in recent video editing modelssignificant challenges remain in developing a versatile and user-friendly framework that facilitates easy video modification for personal use. The primary challenges include: 1) the complexity of training a unified framework encompassing multiple video editing skills (e.g., remove, add, or change an object). Training a single model to perform various editing skills is highly challenging, and recent video editing methods often focus on specific tasks, such as background inpaintingor attribute editing. 2) the necessity for well-structured textual

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

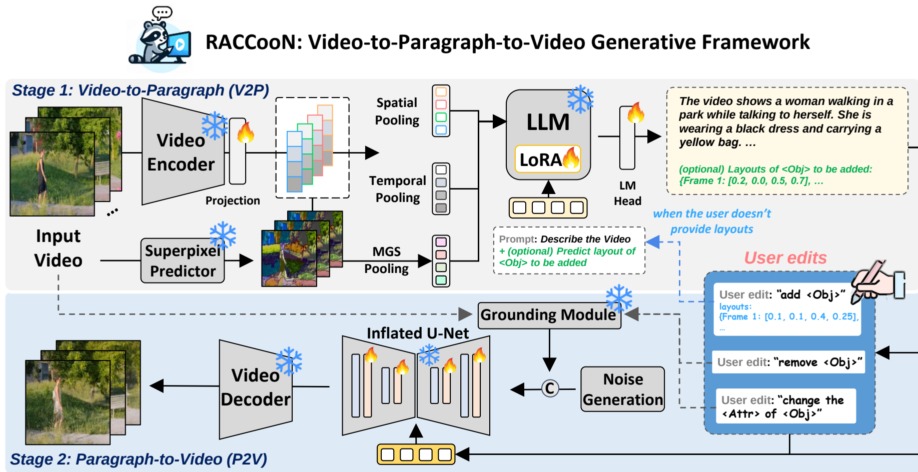

Figure 2: **Illustration of RACCooN framework.** RACCooN generates video descriptions with the three distinct pooled visual tokens, including Multi-Granular Spatiotemporal (MGS) Pooling. Next, users can edit the generated descriptions by adding, removing, or modifying words to create new videos. Note that for adding object tasks, if users do not provide layout information for the objects they want to add, RACCooN can predict the target layout in each frame.

prompts that accurately describe videos and can be edited to support diverse video editing skills. The quality of prompts critically influences the models' capabilities and the quality of their outputs. Generating detailed prompts is time-consuming and costly, and the quality varies depending on the expertise of the annotators. Although Multimodal Large Language Models (MLLMs)have been explored for automatically describing videos, they often overlook critical details in complex scenes. This oversight compromises the development of a seamless pipeline, hindering both user convenience & the effectiveness of video generative models.

To tackle these limitations, we introduce **RACCooN**, a novel **video-to-paragraph-to-video (V2P2V)** generative framework that facilitates diverse video editing capabilities based on auto-generated narratives, as illustrated in Fig. 1. RACCooN allows for the seamless removal and modification of subject attributes, as well as the addition of new objects to videos **without requiring densely annotated video prompts or extensive user planning**. Our framework operates in two main stages: *video-to-paragraph* (V2P) and *paragraph-to-video* (P2V) (Please see Fig. 2). In the V2P stage, we introduce a new video descriptive framework built on a pre-trained Video-LLM backboneWe find that existing Video-LLMs effectively capture holistic video features, yet often overlook detailed cues that are critical for accurate video editing, as users may be interested in altering these missing contexts. To address this, we propose a novel multi-granular video perception strategy that leverages superpixelsto capture diverse and informative localized contexts throughout a video. Next, in the P2V stage, to integrate multiple editing capabilities into a single model, we fine-tuned a video inpainting model that can paint video objects accurately with detailed text, object masks, and condition video. Then, by utilizing user-modified prompts from generated descriptions in the V2P stage, our video diffusion model can accurately *paint* corresponding video regions, ensuring that textual updates from prompts are reflected in various editing tasks. Moreover, to better support our model training, we have collected the **V**ideo **P**aragraph with **L**ocalized **M**ask (**VPLM**) dataset—a collection of over 7.2K high-quality video-paragraph descriptions and 5.5k detailed object descriptions with corresponding masks, annotated from the publicly available ROVIdataset using GPT-4V.

We emphasize that RACCooN enhances the quality and versatility of video editing by leveraging detailed, automatically generated textual prompts that minimize ambiguity and refine the scope of generation. We validate the extensive capabilities of the RACCooN framework in both V2P generation, text-based video content editing, and video generation on ActivityNetYouCook2UCF101DAVISand our proposed VPLM datasets. On the V2P side, RACCooN outperforms several strong video captioning baselines, particularly improving by average **+9.1%p** on VPLM and up to **+9.4%p** on YouCook2 compared to PG-VL, based on both automatic metrics and human evaluation, as shown in Tabs. 1 and 2. On the P2V side, RACCooN surpasses previous strong video editing/inpainting baselines over three subtasks of video content editing (remove, add, and change video objects) over 9 metrics. Detailed results and visualizations are in Tab. 3 and Fig. 3.

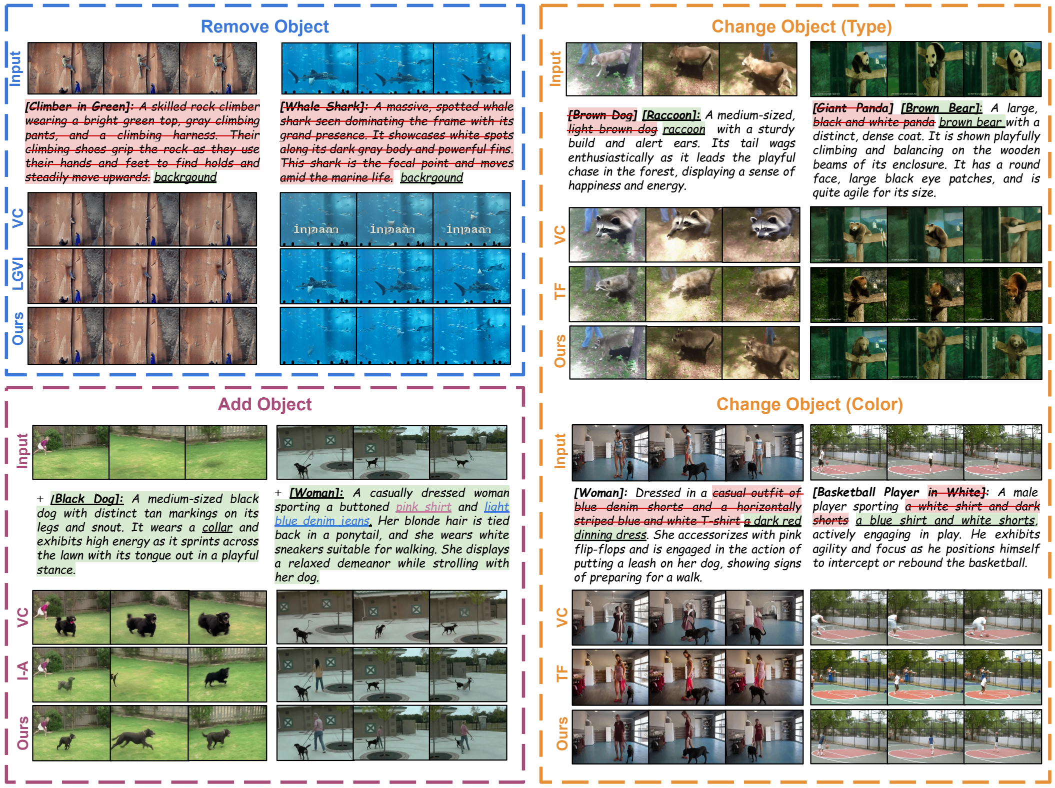

Figure 3: **Qualitative Comparison between RACCooN and other baselines.** Baseline names are abbreviated: **VC**: VideoComposer**I-A**: Inpainting Anything**TF**: TokenFlowWe underlined visual details in our caption. Best viewed in color.

Table 1: **Results of Single Object Prediction** on VPLM test set. Metrics are abbreviated:**S**: *SPICE*, **B**: *BLEU-4*, **C**: *CIDEr*.

Methods	S	B	C	IoU	FVD	CLIP
open-source MLLMs						
LLaVA	17.4	27.5	18.5	-	-	-
Video-Chat	18.2	25.3	19.1	-	-	-
PG-VL	18.2	27.4	14.6	-	-	-
proprietary MLLMs						
Gemini 1.5 Pro	19.2	23.5	11.0	0.115	**371.63**	0.978
GPT-4o	20.6	28.0	**37.4**	0.179	447.67	0.977
RACCooN	**23.1**	**31.0**	33.5	**0.218**	432.42	**0.983**

Table 2: **Results of Human Evaluation** on YouCook2. We measure the quality of the description through four metrics: Logic Fluency (Logic), Language Fluency (Lang.), Video Summary (Summ.), and Video Details (Details). We report the normalized score $s \in [0, 100]$.

Methods	Logic	Lang.	Summ.	Details	Avg.
Ground Truth	66.7	42.2	41.7	**72.2**	55.7
PG-VL	77.2	81.1	69.4	62.8	72.6
RACCooN	**80.6**	**85.0**	**72.2**	**72.2**	**77.5**

Table 3: **Results of Video Content Editing on three sub-tasks** on VPLM test. We gray out models that conduct the DDIM inversion process and have a different focus on our inpainting-based model.

Model	Change Object			Remove Object			Add Object		
	CLIP-T↑	CLIP-F↑	Qedit↑	FVD↓	SSIM↑	PSNR↑	FVD↓	SSIM↑	PSNR↑
Inversion-based Models									
FateZero	25.18	94.47	1.01	1037.05	47.35	15.16	1474.80	47.65	15.45
TokenFlow	29.25	96.23	1.31	1317.29	47.06	15.83	1373.20	49.95	15.95
Inpainting-Based Models									
Inpaint Anything	24.86	92.01	1.01	383.81	82.33	27.69	712.59	77.75	22.41
LGVI	23.82	**95.33**	1.04	915.24	56.16	19.14	1445.43	47.93	16.09
VideoComposer	27.61	94.18	**1.25**	827.04	47.34	17.55	1151.90	48.01	15.76
RACCooN	**27.85**	94.78	1.15	**162.03**	**84.38**	**30.34**	**415.82**	**77.81**	**23.38**