# RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models

## Supplementary Material

## S.1. Dataset

The availability of datasets featuring a wide range of editing tasks and motions for the evaluation of zero-shot text-guided video editing tasks is limited (see Section S.3 for a detailed discussion). To foster standardized evaluations in future video editing research, we have curated a dataset comprising 186 text-video pairs from diverse sources, including Pexels [3], Pixabay [4], DAVIS [33], and Internet videos used by previous approaches. The prompts are obtained from ChatGPT, drawn from previous approaches, or contributed by the authors.

**Length and resolution** The dataset categorizes video lengths into three segments—8, 36, and 90 frames—with 10, 15, and 6 videos respectively, to analyze the impact of duration on video editing methods. Note that the existing methods can handle only up to a certain number of frames; for instance FLATTEN [11] can handle up to 27 frames on RTX4090* and FateZero [34] is up to 45 frames on A40 with their official repositories. Hence, assessing video editing methods for longer video lengths is crucial. We also use resolutions of 512 × 320 or 512 × 256 for rectangular, and 512 × 512 for square videos.

**Types of Edits** We broadly classify editing types into **style editing** and **shape editing**. Style editing is divided into:

- **Local editing** for localized changes (*e.g*. jacket color)
- **Visual style editing** for artistic style changes (*e.g*. 'watercolor style')
- **Background editing** for background or setting changes (*e.g*. beach background)

Furthermore, shape editing is divided into two types:

- **Shape/attribute editing** focusing on altering an object's shape or attributes (e.g., wolf to cat), and
- **Extreme shape editing** for major transformations of an object's shape or nature (e.g., car to tractor)

See Fig. S.8 for examples of each edit type.

**Motion** We assess the robustness of baseline approaches across different types of motions.

- **Exo-motion**: In this type of motion, only the object within the video is in motion, while background stays the same.
- **Ego-motion**: It involves scenarios where the camera itself is in motion.
- **Ego-exo motion**: It combines both camera and object movement in the videos.

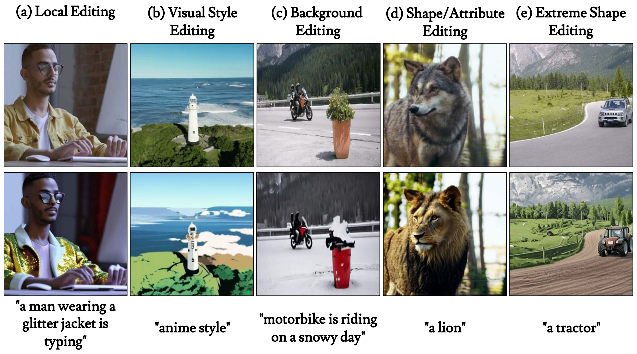

(a) Local Editing
(b) Visual Style Editing
(c) Background Editing
(d) Shape/Attribute Editing
(e) Extreme Shape Editing
"a man wearing a glitter jacket is typing"
"anime style"
"motorbike is riding on a snowy day"
"a lion"
"a tractor"

Figure S.8. ***Types of edits in our dataset.***

- **Occlusion**: It incorporates scenarios where objects are partially or fully hidden from view.
- **Multiple objects with appearances/disappearances**: It features videos with multiple objects, some appearing and disappearing throughout the video.

By considering these diverse types of motions, we can thoroughly evaluate the robustness of the approaches in addressing the complexities of real-world scenarios. This assessment provides valuable insights into the methods' adaptability and effectiveness across a wide range of dynamic situations in video editing.

**Diversity of text prompts** Our dataset involves employing varied levels of detail in our text prompts. This includes examples such as '*a zombie*' and more elaborate descriptions like '*Soft, blended colors and visible brushstrokes make the scene appear as if painted with watercolors*'.

## S.2. Limitations

**Extreme shape editing in long videos** Shape editing is a challenging task in the field of video editing, with most existing methods struggling to maintain consistent shape transformations. Often, approaches that focus on shape editing rely on complex procedures like atlas editing [26], and even these can lead to unsatisfactory outcomes. Our approach, in contrast to many current text-guided video editing models, is capable of handling shape edits from simple to extreme examples. For instance, our method can transform a wolf into a cat, bear, or dinosaur (as shown in Fig. 1), or can convert a boat into a jeep or a monkey into a bear (illustrated in Fig. 6 of the main paper). Moreover, our method can handle **extreme** shape edits, such as transforming a car into a fire-truck, train, tractor, and so on. However, while our method can handle these edits successfully, it encounters limitations when performing **extreme** shape edits

*FLATTEN Last accessed: 2023-11-26.

as the video length increases. In particular, the ability of our method to maintain the distinct shape of these extreme objects weakens, resulting in some flickering. It is noteworthy that in cases of extreme editing, such as with the car-turn example, our method effectively manages shape transformations for up to 27 frames, beyond which the quality of the edit starts to degrade. This 27-frame threshold is significant as it represents the upper limit of the editing capabilities of many competing methods, such as FLATTEN [11] (on RTX4090), for similar tasks.

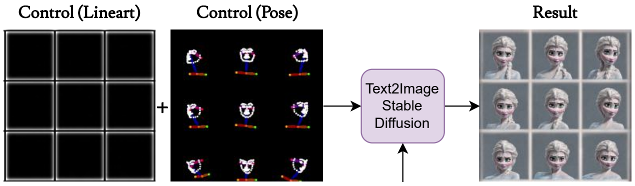
**Fine details flickering** Certain extreme shape editings (e.g., transforming the wolf into 'a unicorn') require high-frequency edits in the video (such as long and rich hair details of the unicorn). In such cases, flickering may occur as our model does not explicitly utilize pixel-level methods to address video deflickering. Furthermore, the unavoidable losses incurred during the compression in the encoding/decoding steps of latent diffusion models and the selection of inversion methods (DDIM inversion in our case) impact the quality of reconstructing fine details. Note that this is a common challenge present in existing approaches as well.

## S.3. Existing Datasets in Video Editing Literature

TokenFlow [14] utilized 61 text-video pairs sourced from DAVIS [33] and Internet Videos. Rerender [51] employed test videos from Pexels [3] and Pixabay [4], while Text2Video-Zero [25] randomly selected 25 videos generated by CogVideo [20]. FateZero [34] utilized videos from DAVIS and other in-the-wild videos, with text prompts created by the authors. Pix2Video [10] employed a subset of videos from DAVIS, along with prompts acquired from previous works, some of which were generated by users. FLATTEN [11] used 16 videos from DAVIS and 37 videos from Videvo [†], each with 4 prompts and 32 frames per video. Tune-A-Video [50] employed 42 videos from DAVIS with 140 manually crafted text prompts. While some of these datasets are publicly available and others are not, there remains a notable lack of a standardized video dataset in the literature that includes a wide range of motions and longer-duration videos.

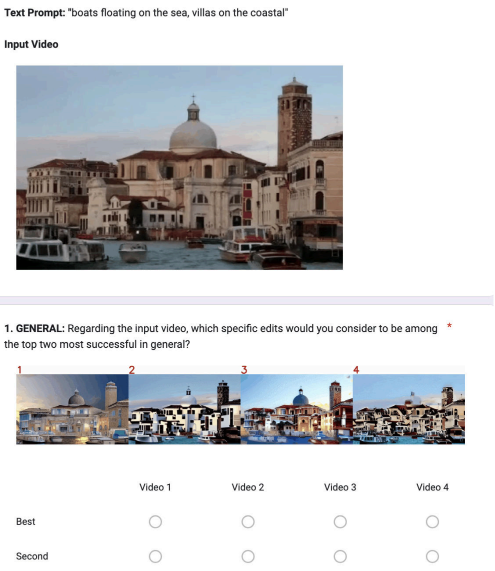
## S.4. Grid Trick

Often referred as the *character sheet*, specifying desired characteristics in the text prompt and utilizing a grid (on the left) as a condition when employing ControlNet. The resulting output maintains the grid format during the editing process, ensuring consistent styles.



"(a character sheet of Elsa from different angles with a gray background:1.4), white hair,blue eyes open, cinematic lighting, Hyperrealism, depth of field, photography..."

Figure S.9. ***Grid trick.***



**Text Prompt:** "boats floating on the sea, villas on the coastal"

**Input Video**

**1. GENERAL:** Regarding the input video, which specific edits would you consider to be among the top two most successful in general? *

|  | Video 1 | Video 2 | Video 3 | Video 4 |
|---|---|---|---|---|
| Best | ○ | ○ | ○ | ○ |
| Second | ○ | ○ | ○ | ○ |

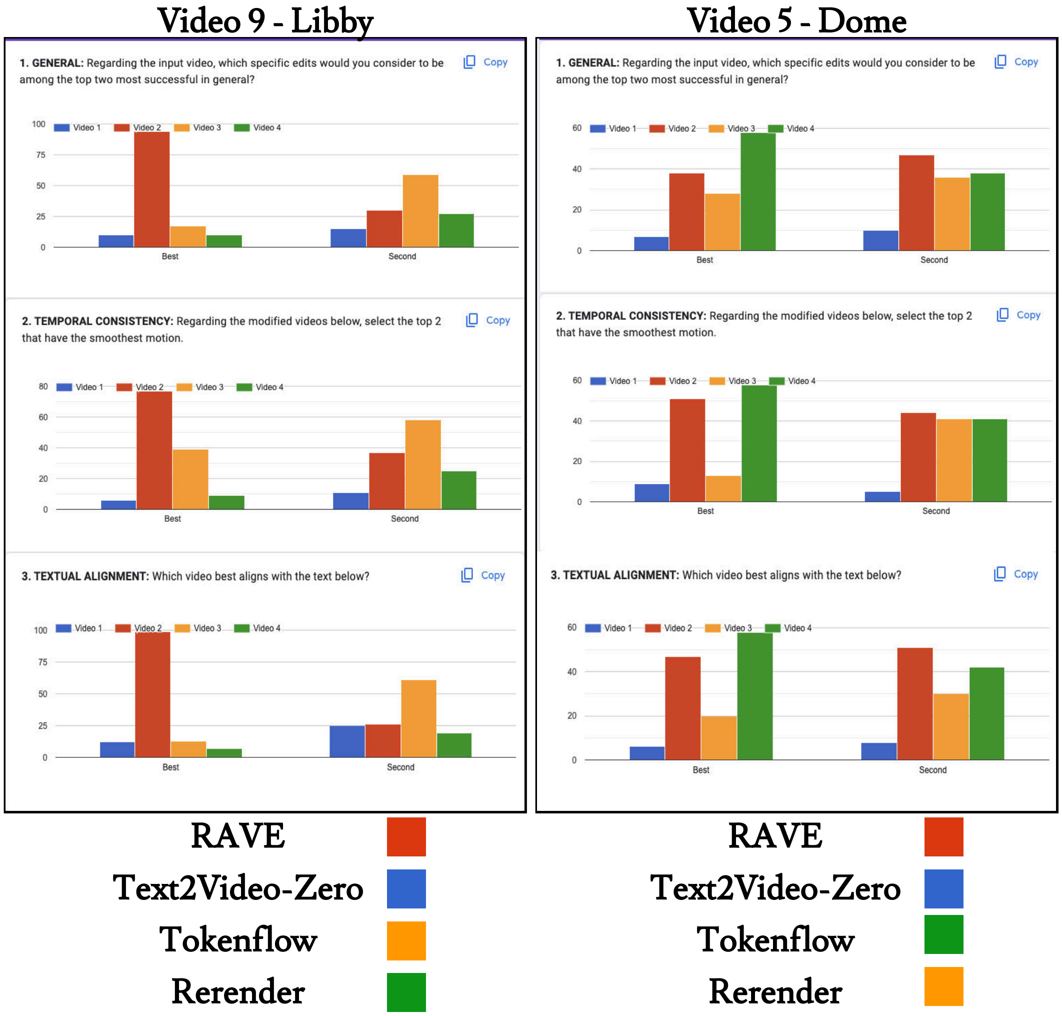
Figure S.10. ***A screenshot of the user study.***

## S.5. User Study Details

We conducted a user study involving 130 anonymous participants recruited from Prolific, a crowd-sourcing platform commonly utilized in research studies. The study focused on 23 randomly selected video-text pairs from our dataset. The comparison in the user study was made among Text2Video-Zero [25], Rerender [51], Tokenflow [14], and our approach. Note that we employed Stable Diffusion v1.5 in all comparative analyses, including user study.

Note that Tokenflow [14] did not conduct a user study, whereas Rerender [51] surveyed 30 people to assess general editing capabilities. FateZero [34] conducted a study with 20 participants, focusing on questions related to textual alignment, temporal consistency, and general editing capabilities. Additionally, Pix2Video [10] conducted a survey with 37 participants, asking questions about general editing capabilities. FLATTEN [11] conducted a study with 16 participants, examining semantical, motional, temporal, and structural consistencies. In contrast, Tune-A-Video [50] did not specify the number of users in their survey, concentrating on questions concerning temporal consistency and textual alignment. As opposed to previous works, our survey includes 130 anonymous participants.

[†] Videvo Last accessed: 2023-11-16.



1. GENERAL: Regarding the input video, which specific edits would you consider to be among the top two most successful in general?

2. TEMPORAL CONSISTENCY: Regarding the modified videos below, select the top 2 that have the smoothest motion.

3. TEXTUAL ALIGNMENT: Which video best aligns with the text below?

Figure S.11. ***Example results from user study.*** The results from two scenarios in our user study with 130 participants are illustrated. On the left, RAVE achieves the top-1 ranking on 'libby' video, while on the right, Tokenflow secures the top-1 position. Each participant responded to three questions for each video.

### S.5.1. Questions

We presented three questions to the participants, requesting them to rank the top two video edits among the four provided videos based on the following questions:

- **Question 1 - General Editing (GE)**: "Regarding the input video, which specific edits would you consider to be among the top two most successful in general?"
- **Question 2 - Temporal Consistency (TC)**: "Regarding the modified videos below, select the top 2 that have the smoothest motion."
- **Question 3 - Textual Alignment (TA)**: "Which video best aligns with the text below?"

The screenshot of the survey form for a single question and video is depicted in Fig. S.10. It's important to mention that for each user and video, the order of videos produced by each method is randomly shuffled to ensure an unbiased comparison.

### S.5.2. Further analysis

Note that we formulate a metric as the frequency of each method chosen among the top two edits, as provided in Table 1. We provide the results of two examples (complete videos are available in the Supplementary Website) from our user study, one selected as the best and the other not selected, in response to Question 1 with that metric. (Fig. S.11)

We notice that videos with relatively stable backgrounds yield nearly evenly distributed selections among the different approaches. In contrast, videos featuring ego-exo motion and occlusions within dynamic scenes consistently demonstrate superior performance for our method compared to previous approaches. Please refer to the Supple-

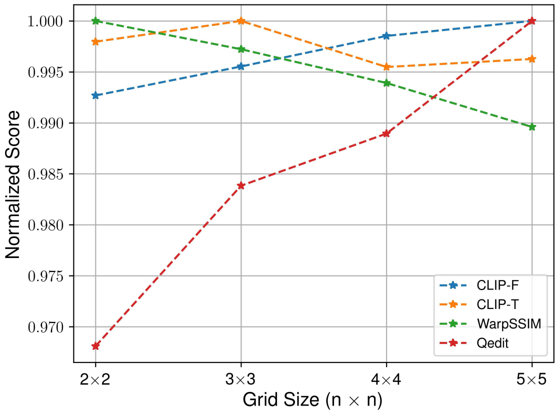


Figure S.12. ***Ablation study on grid size***

mentary Website for the complete videos.

## S.6. Further Results

### S.6.1. Stable diffusion vs. realistic vision

In our qualitative results, we utilize Realistic Vision V5.1 to harness its diverse editing capabilities (note that for comparison with other methods, we used Stable Diffusion v1.5 in order to have a fair comparison). As an ablation of using Realistic Vision model vs. Stable Diffusion, we also perform a comparison with the results obtained with each method. As observed, in both cases our method is able to apply the edits successfully, and temporal consistency is maintained. Please see Supplementary Website for examples.

### S.6.2. Grid size vs metrics

We conduct an additional ablation to explore the impact of grid size on the quality of editing. In Fig. S.12, the normalized metrics averaged over 90-frame videos are presented for grid sizes of $2 \times 2$, $3 \times 3$, $4 \times 4$, and $5 \times 5$. As anticipated, there is a noticeable improvement in CLIP-F and $Q_{\text{edit}}$ as the number of frames within a grid increases, attributed to the increased level of interaction that plays a more significant role in longer videos. Note that we chose $3 \times 3$ in our experiments since the GPU requirement increases in proportion to the grid size.

## S.7. Comparison with Keyframe Propogation

We also evaluate our approach against a straightforward method. In this method, we produce keyframes using the grid trick and employ off-the-shelf tools such as Ebsynth [21] to fill in between these keyframes.

Fig. S.13 illustrates an instance where Ebsynth is combined with the 'Grid without shuffling' approach. It is evident that significant alterations occur in the structure of the car and the environment across consecutive frames of the keyframes in the Ebsynth method. In contrast, our approach demonstrates superior handling of temporal structural consistency.

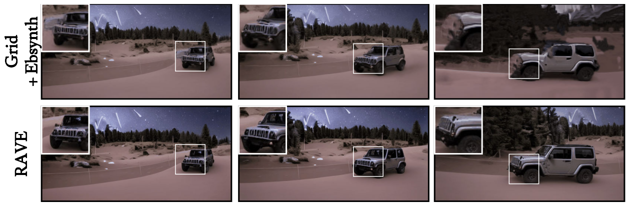


Figure S.13. ***Comparison with Ebsynth.*** We generated keyframes using the grid trick and employed Ebsynth to fill out between the keyframes.