SUPPLEMENTARY OF COHERENT VIDEO-TO-VIDEO TRANSFER USING SYNTHETIC DATASETS

A INSTRUCTION FOR ZERO-SHOT MPT DATA GENERATION:

The instruction for MPT comprises three parts. The first is the task description which delineates the objectives for the bot.

```
You are a bot to generate synthetic text data for generating video
↪  clip. You will creatively generate paired prompt triplet for
↪  synthetic data generation for video clip. You will be given an
↪  input prompt and you should return the edit prompt and output
↪  prompt. The output prompt reflects the sentence after applying
↪  edit prompt on the input prompt. Ensure that the prompt are
↪  proper for generating video clip (i.e. prompt should describe
↪  a scene).

Successful editing do changing the main subject, modifying the
↪  context or setting or altering the artistic style.

Here are some examples of editing that are likely to success (do
↪  not limit to the verb used in the follow. You must be creative
↪  and the editing should be diverse):

Edit of Landscape:
Edit: Convert the cityscape to a seascape.
Edit: Turn the desert scene into a lush forest.

Replacement of Characters:
Edit: Replace the cowboys with astronauts.
Edit: Turn the group of children into a group of elderly people.

Edit of Time:
Edit: Switch the night scene to a day scene.
Edit: Transform the contemporary setting into a medieval setting.

Addition of Significant Elements:
Edit: Add a full moon to the clear sky.
Edit: Include a rainbow in the cloudy scene.

Edit of Weather or Season:
Edit: Make the sunny day into a snowfall.
Edit: Transform the summer scene into autumn.

Edit the Action or Activity:
Edit: Change the soccer game to a ballet performance.
Edit: Replace the cooking scene with a gardening scene.

Edit of Artistic Style:
Edit: Make it look like a watercolor painting.
Edit: It is now in the style of Van Gogh. (do not only use Van
↪  Gogh)
```

The subsequent phase of the instruction involves presenting MPT with five randomly selected examples from the LAION-IPTP dataset. These samples have been previously successful in generating paired prompts that meet the CLIP filter criteria. The depiction below illustrates this process:

```
Here are some success examples (please be creative and not limited
↪  to examples)

Input: Graham Wands - George Square, Glasgow, watercolour
Edit: Turn the watercolour into a pencil sketch
Output: Graham Wands - George Square, Glasgow, pencil sketch

Input: Pierre de Clausade, (French, 1910-1976), Winter at the Lake
Edit: make it a sunset
Output: Pierre de Clausade, (French, 1910-1976), Sunset at the
↪  Lake

Input: Rex Beanland, Charing Cross, watercolour, 9 12
Edit: make it an oil painting
Output: Rex Beanland, Charing Cross, oil painting, 9 12

Input: Mark Van Crombrugge, Old Milk Bottle and Grapes, oil, 31 x
↪  59.
Edit: Make the bottle transparent.
Output: Mark Van Crombrugge, Transparent Old Milk Bottle, oil, 31
↪  x 59.

Input: """Large Original Oil painting on canvas. Beautiful
↪  portrait of a woman 24x24""""""""
Edit: make the woman a cat
Output: """Large Original Oil painting on canvas. Beautiful
↪  portrait of a cat 24x24"""""
```

Finally, MPT is provided with video captions from the WebVid dataset that are designated for processing. This marks the initiation of the generation process for novel, paired prompts.

```
Generate triplet for following inputs:

Merida, mexico - may 23, 2017: tourists are walking on a roadside
↪  near catholic church in the street of mexico at sunny summer
↪  day.

Fun clown - 3d animation

Happy family using laptop on bed at home

11th march 2017. nakhon pathom, thailand. devotees goes into a
↪  trance at the wai khru ceremony at wat bang phra temple. what
↪  bang phra is famous for its magically charged tattoos and
↪  amulets.

Decorate with pineapple sweet cake roll.

Beautiful lake aerial view

Frankfurt, germany-circa 2013:traffic with skyscrapers in
↪  background at night along the river main in frankfurt, time
↪  lapse

Young positive couple laughing in the backyard in front of the
↪  large house under falling snow. bearded man and attractive
↪  woman in warm clothes have winter fun. the guy showing thumb
↪  up
```

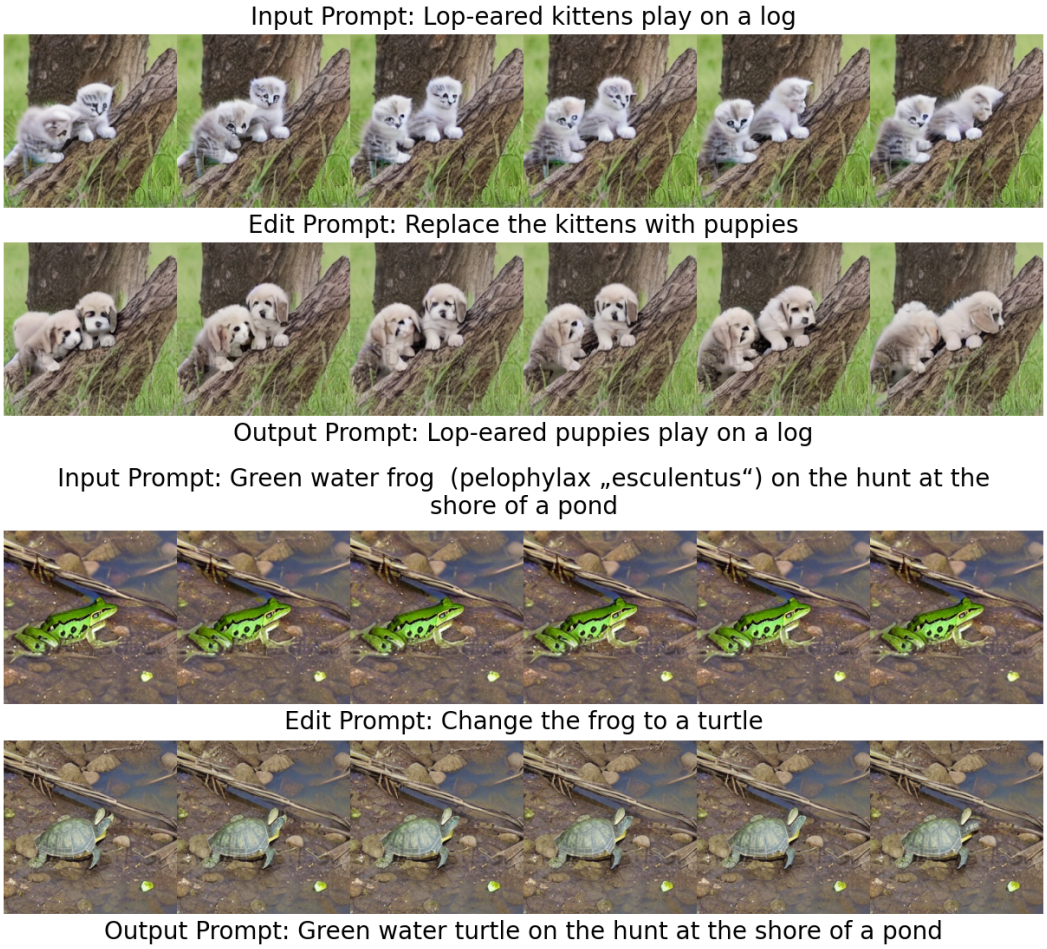

Figure 5: Visualization of synthetic samples produced by the data generation pipeline in Section 3. Each sample contains two generated videos: the upper one represents the input prompt, while the lower one depicts the output prompt. The actual videos comprise 16 frames while 6 subsampled frames are displayed here for brevity.

```
Broadcast twinkling squared diamonds, multi color, abstract,
↪  loopable, 4k

Wheat harvesting. combine harvester gathers the wheat crop on the
↪  field.
```

B Visualization of Synthetic Video Dataset

In Section 3, we introduce a pipeline designed to produce synthetic paired videos. Further visual demonstrations of these results can be found in Figures 5 and 6. The generated paired videos maintain a similar structure, and the editing can showcase the edit prompt.

C Long Video Sampling Detail Illustration and Examples

In Section 4.4, we addressed the challenge of ensuring consistency across different batches when sampling long videos, particularly when each batch is processed separately. This section introduces an illustrative explanation of the Long Video Sampling Correction (LVSC) method in Figure 7. In

Input Prompt: Chicago, illinois - 2010: this is a sculpture of a lion located in front of the art institute of chicago on michigan avenue. this lion can be seen in chicago movies like ferris bueller's day off.

Edit Prompt: Replace the lion with a dragon

Output Prompt: Chicago, illinois - 2010: this is a sculpture of a dragon located in front of the art institute of chicago on michigan avenue.

Input Prompt: Honey bees on the pink flowers of a tree peony collect honey and pollen.

Edit Prompt: The pink flowers are replaced with sunflowers.

Output Prompt: Honey bees on the sunflowers collect honey and pollen.

Input Prompt: Woman is eating pizza at home

Edit Prompt: Replace pizza with sushi

Output Prompt: Woman is eating sushi at home

Figure 6: Visualization of synthetic samples produced by the data generation pipeline in Section 3. Each sample contains two generated videos: the upper one represents the input prompt, while the lower one depicts the output prompt. The actual videos comprise 16 frames while 6 subsampled frames are displayed here for brevity.

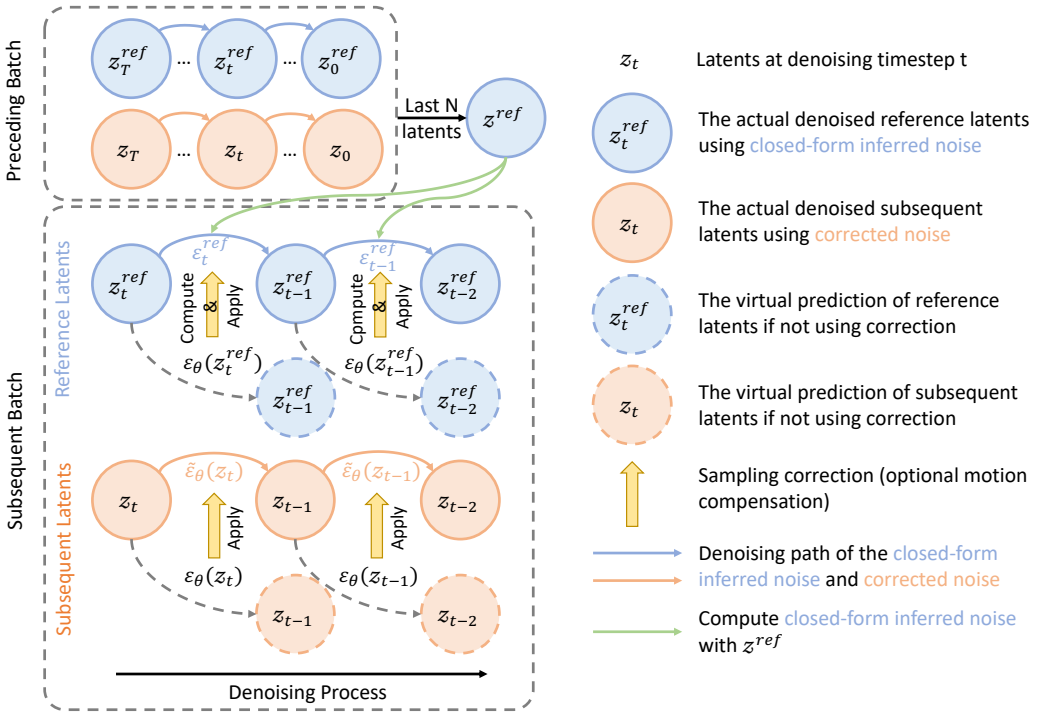

Figure 7: Schematic of Long Video Sampling Correction (LVSC) mechanism. The diagram illustrates the interaction between reference latents from a preceding batch and subsequent latents during the denoising process. The closed-form inferred noise (Equation (4)) is computed for the N reference latents (shown in blue), which then guides the correction of the actual denoised subsequent latents (shown in orange).

LVSC, there are N overlapping reference frames between two consecutive batches, but this method transcends a basic sliding window technique. Instead of allowing the editing model unrestricted freedom in transferring edits, the reference frames guide the appearance of subsequent frames during the editing process.

Additionally, we present a visual comparison to demonstrate the impact of LVSC. Figure 8 contrasts the results with and without LVSC implementation. This comparison clearly shows the lack of continuity in the sampled frames between batches when LVSC is not applied. In contrast, employing LVSC achieves noticeable consistency, aligning the first frame of a subsequent batch with the last frame of the previous batch.

D Effect of Motion Compensation in Long Video Sampling Correction

In our experiments, we observed that the presence of holistic camera motion could degrade the transfer results of subsequent batches processed by the LVSC model (see red boxes in Figure 9). This degradation arises because the same regions across different frames require consistent score corrections. In other words, score corrections should "travel" with the regions affected by camera movement, ensuring that identical corrections are applied to the same areas regardless of motion. To address this issue, we introduce a motion compensation strategy. Specifically, we first employ the RAFT flow estimator Teed & Deng (2020) to compute the optical flow between each reference frame and the remaining frames in subsequent batches. The original equation for LVSC (Equation (5)) is then modified as follows:

$$\tilde{\varepsilon_\theta}(z_t)[:,m] = \varepsilon_\theta(z_t)[:,m] + (\frac{1}{N}\sum_{i=1}^{N} o(\varepsilon_t^{ref}[:,i]), i \to m) - \varepsilon_\theta(z_t)[:,m] \tag{6}$$

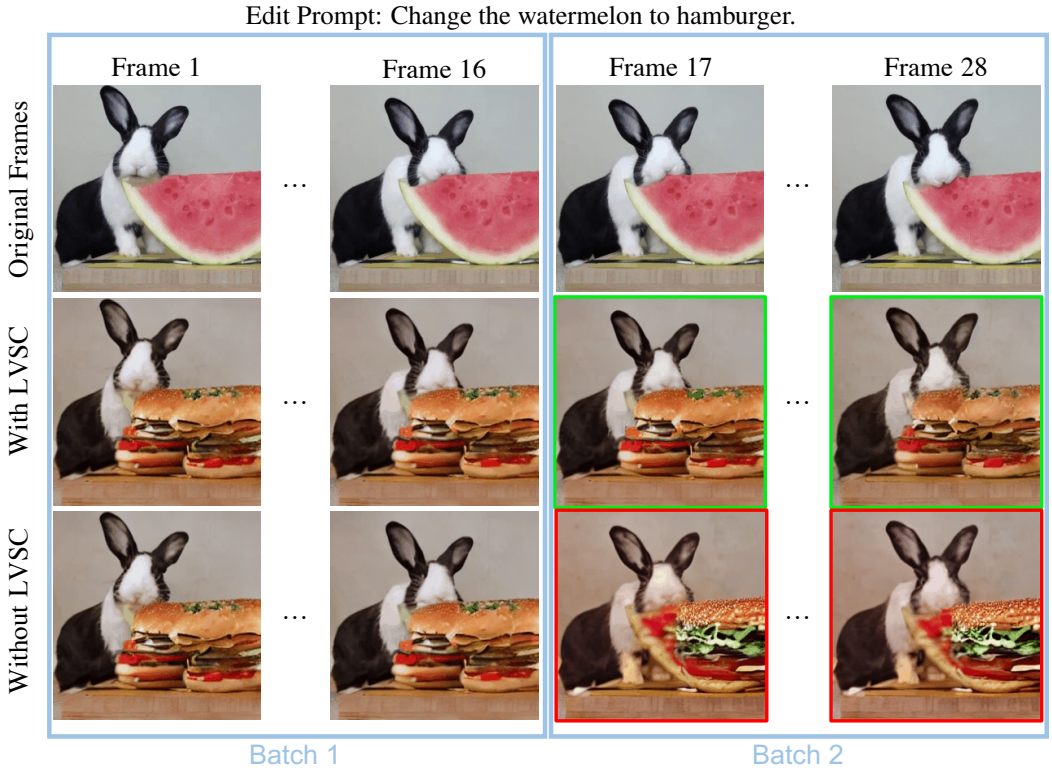

Figure 8: In the process of video sampling, we utilize a batch size of 16. The 17th to 28th frames in the video are processed in the second batch, and they reference the last four sampled frames from the preceding batch. By employing the Long Video Sampling Correction (LVSC), we can ensure the content consistency between sampled frames across different batches (green boxes in the figure). In contrast, sampling two batches separately may lead to inconsistencies at the boundaries where the batches change (rex boxes in the figure).

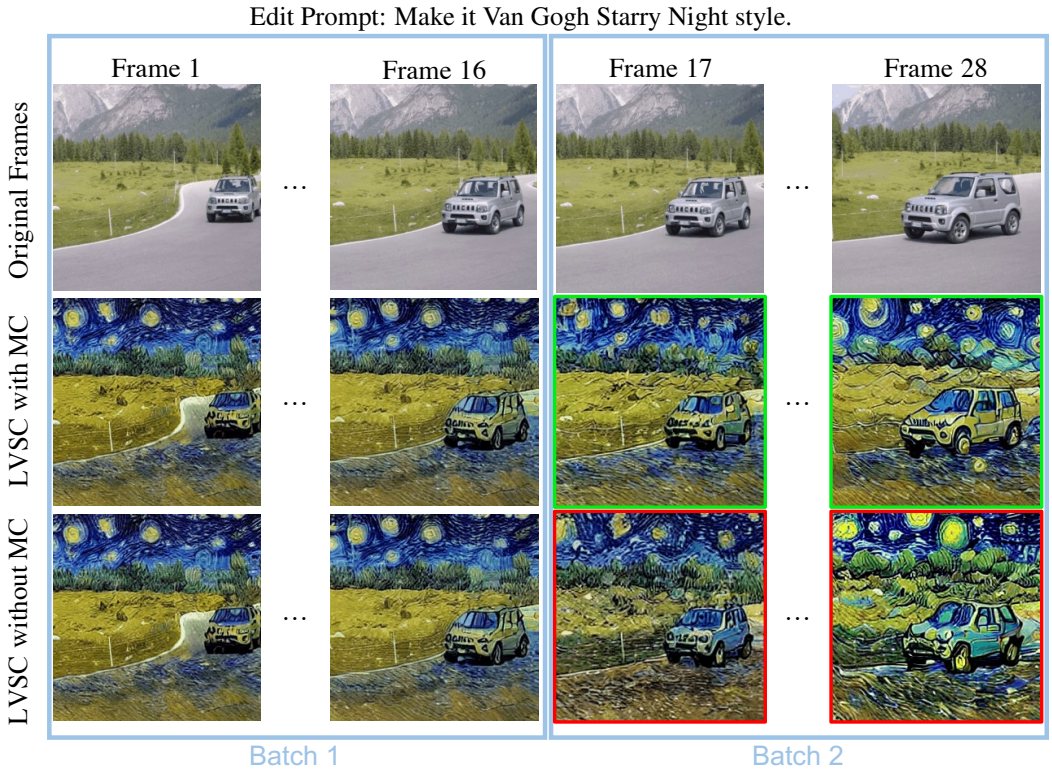

Figure 9: When the video exhibits extensive camera motion, the LVSC model without motion compensation (LVSC without MC) suffers significant quality degradation, particularly in frames from subsequent batches (red boxes in the figure). Incorporating motion compensation (LVSC with MC) substantially improves frame consistency with the reference and maintains high transfer quality, even when the camera view changes dramatically in later batches (green boxes in the figure).

Here, $m = [1, 2, ..., M]$ represents the indices of frames in the subsequent batches, which contain a total of M frames, $o(\cdot, i \rightarrow m)$ is a warping function that uses optical flow to align the i-th reference frame with the m-th frame in the subsequent batch. This modification ensures better transfer quality by making the score corrections to be aware of camera movement. The green boxes in Figure 9 reveal that applying motion compensation significantly enhances content consistency and overall quality. This improvement is particularly noticeable in the last few frames of the subsequent batches.

E Effect of Sampling Hyperparameters and Picking Criteria

In our experiments, we observed that the video CFG and resolution have a greater impact on the generated video than the text CFG. To streamline the parameter search process, we thus focus solely on picking the video CFG and resolution, effectively reducing the search space. Detailed visual effects of the hyperparameter choices are provided in Figure 10. We opt for PickScore Kirstain et al. (2023) as our automated selection criterion. Our choice is motivated by the fact that PickScore aligns more closely with human perception compared to the CLIP score, as indicated in Kirstain et al. (2023).

F Qualitative Comparisons

We provide qualitative comparisons with baselines in Figures 11 to 14.

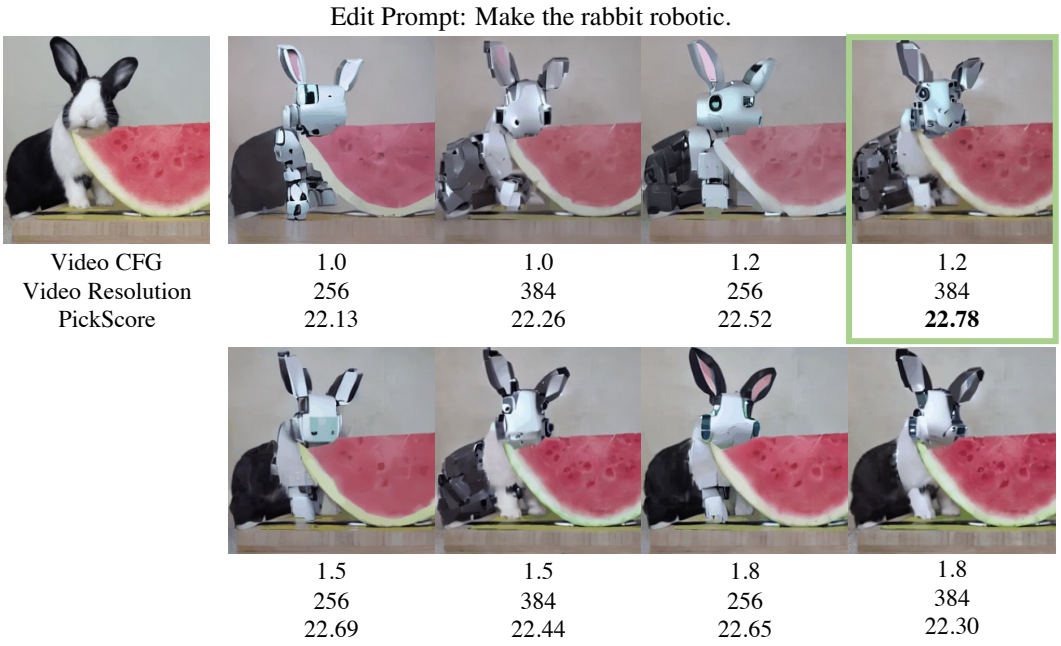

Figure 10: During sampling, the interplay between the scale of image classifier-free guidance (CFG) and image size can significantly influence the transfer result. In our evaluation, we sample videos by employing various combinations of CFG scales and resolutions, utilizing the average PickScore across all frames as the selection metric. The video with the highest PickScore is designated as the final transfer result. As indicated by the green box in the figure, the chosen video and its corresponding hyperparameters bear the highest PickScore, thereby making it the final selection.

G Failure Cases

Our model, InsV2V, occasionally encounters challenges in video transfer, particularly when the object of interest is difficult to detect. Such difficulties arise when the object is positioned close to the video's edge, is unusually small or large, or is partially obscured. These scenarios can lead to the object either vanishing in the edited video or exhibiting inconsistent appearances.

For instance, as depicted in Figure 17, when a person is situated near the frame's border and occupies a small area, InsV2V struggles to maintain the person's structural integrity, resulting in their disappearance from the transferred video. However, when the person moves away from the border, InsV2V can recognize them again, but with a notably altered appearance.

Figure 18 presents another scenario where the object, in this case, the front part of a truck, is only partially visible in certain frames. While InsV2V can accurately transfer the image when the truck is fully visible, it misinterprets the partially visible truck as a car's front, leading to inconsistent results in the transferred video.

In summary, when object detection is hindered due to size, positioning, or occlusion, InsV2V may not deliver satisfactory transfer results.

Figure 11: Swans gliding over a lake. → Pink flamengos gliding over a lake.

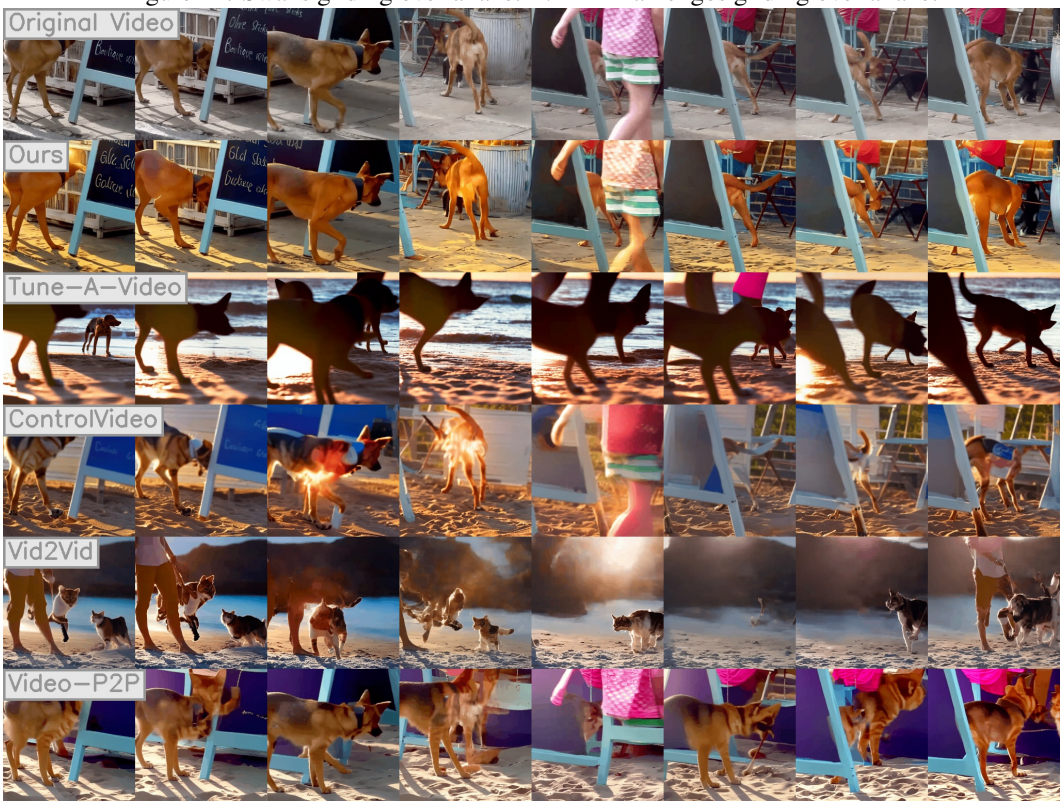

Figure 12: A cat and a dog playing on the street while a girl walks around them.
→ A cat and a dog playing on the beach while a girl walks around them, golden hour lighting.

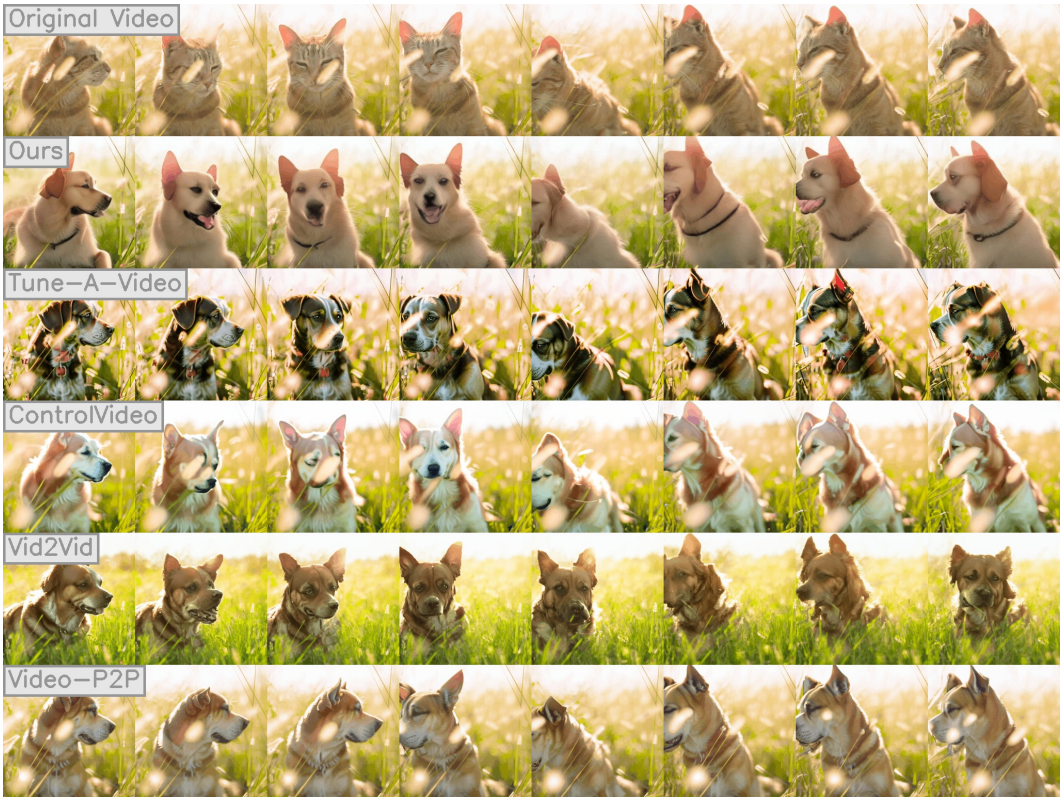

Figure 13: A cat in the grass in the sun. → A dog in the grass in the sun.

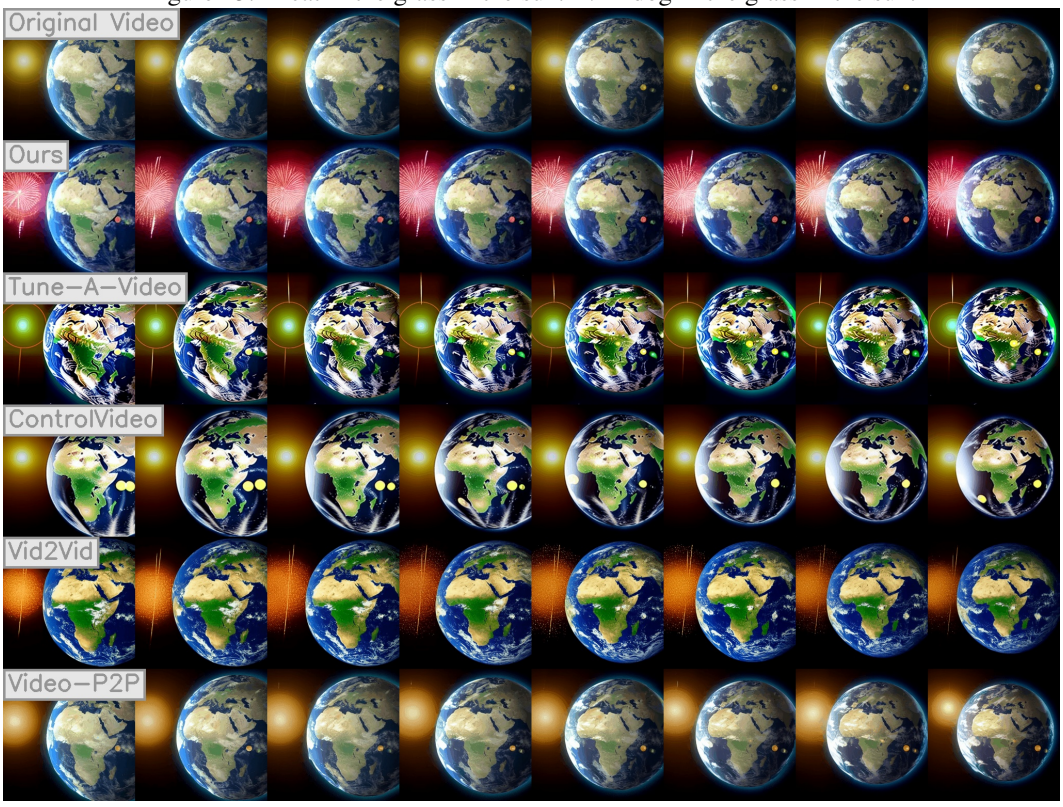

Figure 14: Full view of the Earth as it moves slowly toward the sun.
→ Full view of the Earth as it moves slowly through a fireworks display.

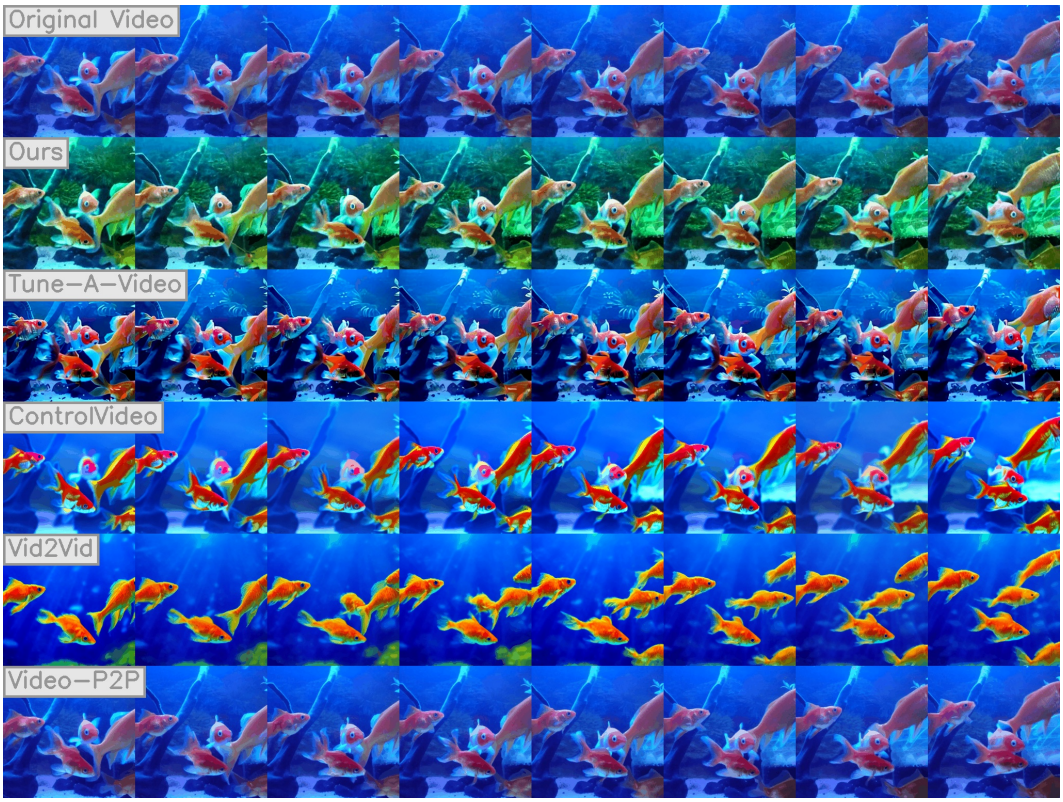

Figure 15: Several goldfish swim in a tank.→ Several goldfish swim in a pond.

Figure 16: A static shot of red roses in sunlight, gently swaying in the breeze.
→ A static shot of red roses in sunlight, gently swaying in the breeze, origami style.

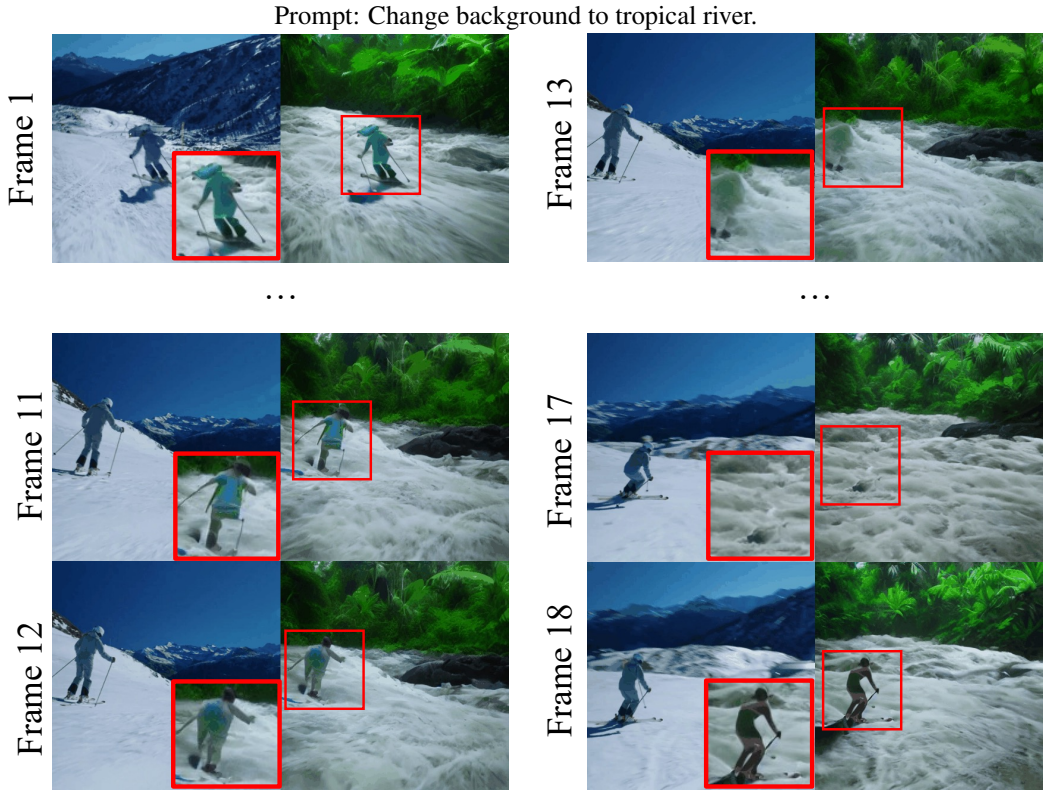

Figure 17: Illustration of a failure case where InsV2V struggles to maintain the structure of a person located near the video's edge and occupying a small area, resulting in the disappearance of the person from the transferred video.

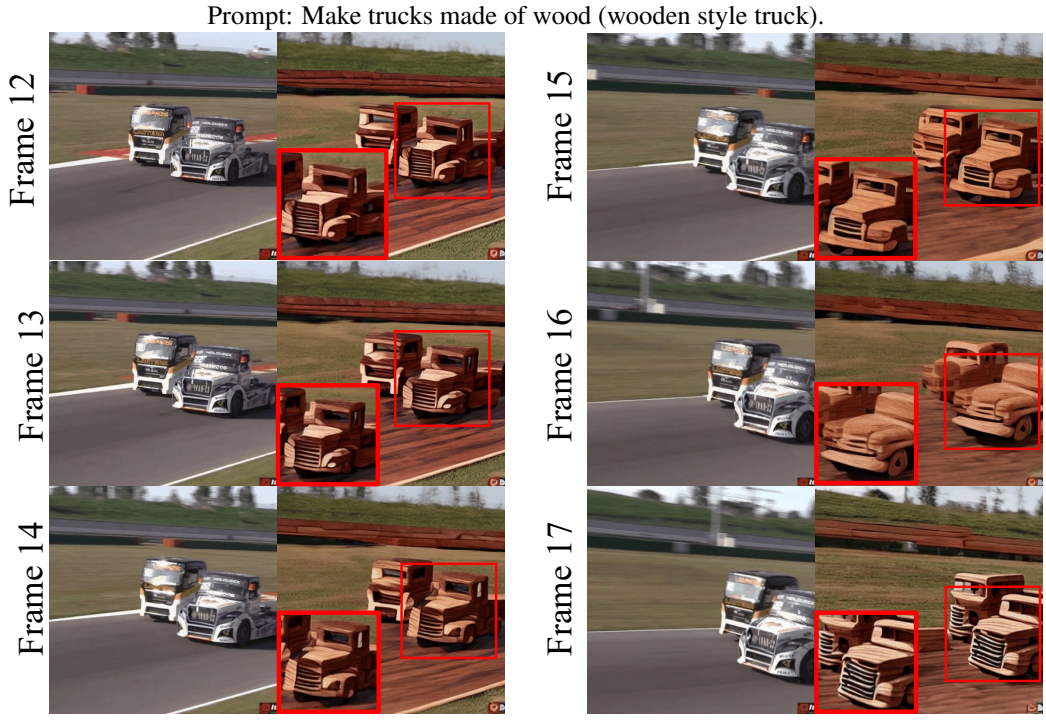

Figure 18: Demonstration of a failure scenario in InsV2V where a partially visible truck is misidentified as a car's front, leading to inconsistent transfer results in the edited video.