

MeDM: Mediating Image Diffusion Models for Video-to-Video Translation with Temporal Correspondence Guidance

Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, Jun-Cheng Chen

Research Center for Information Technology Innovation, Academia Sinica

128 Academia Road, Section 2

Nankang, Taipei 115, Taiwan

{shchu, jason890425, joseph, pullpull}@citi.sinica.edu.tw

Abstract

This study introduces an efficient and effective method, MeDM, that utilizes pre-trained image Diffusion Models for video-to-video translation with consistent temporal flow. The proposed framework can render videos from scene position information, such as a normal G-buffer, or perform text-guided editing on videos captured in real-world scenarios. We employ explicit optical flows to construct a practical coding that enforces physical constraints on generated frames and mediates independent frame-wise scores. By leveraging this coding, maintaining temporal consistency in the generated videos can be framed as an optimization problem with a closed-form solution. To ensure compatibility with Stable Diffusion, we also suggest a workaround for modifying observed-space scores in latent-space Diffusion Models. Notably, MeDM does not require fine-tuning or test-time optimization of the Diffusion Models. Through extensive qualitative, quantitative, and subjective experiments on various benchmarks, the study demonstrates the effectiveness and superiority of the proposed approach.

1 Introduction

Recently, diffusion-based deep generative models (DMs) (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021; Song et al. 2021) have set new benchmarks in image generation, particularly for text-to-image tasks. Major tech companies are investing heavily in large-scale text-to-image models, like DALLE-2 (Ramesh et al. 2022), Imagen (Saharia et al. 2022), and Stable Diffusion (Rombach et al. 2022). These models deliver high-quality images within seconds based on text descriptions. Additionally, models like InstructPix2Pix (Brooks, Holynski, and Efros 2023) and ControlNet (Zhang and Agrawala 2023) enable instruction-based and structure-preserving text-guided image-to-image translations. These advances have spurred novel applications in content creation. However, generating videos or performing video-to-video translation while maintaining temporal consistency remains a challenging and ongoing problem.

To tackle this challenge, many text-to-video models have been proposed, including Imagen Video (Ho et al. 2022), Make-a-Video (Singer et al. 2022), etc. However, these approaches demand substantial training videos and computational resources. To mitigate these requirements, researchers have extended text-to-image models to encom-

pass video generation, incorporating a temporal aspect into the latent-space-based DM (Blattmann et al. 2023). In addition, several zero-shot methods are also proposed to directly leverage the pre-trained image-based or text-to-image-based DMs for text-guided video-to-video translation, including Pix2Video (Ceylan, Huang, and Mitra 2023), Rerender-a-Video (Yang et al. 2023), ControlVideo (Chen et al. 2023), etc. While these methods achieve video transformation, some still experience persistent flickering issues, while others involve intricate and meticulous processes.

To perform efficient and high-quality temporally consistent video-to-video translation, we propose a novel temporal correspondence guidance based on optical flows to Mediate image Diffusion Models (MeDM, pronounced “medium”) for video-to-video translation. Explicit optical flows are utilized to establish a pragmatic encoding of pixel correspondences across video frames. By leveraging this encoding, maintaining temporal consistency in the generated videos can be framed as an optimization problem with a closed-form solution. Then, the proposed approach enables high-quality and temporally consistent video-to-video translation by iteratively aligning corresponding noisy pixels across video frames using the provided temporal correspondence guidance derived from optical flows. Notably, MeDM does not necessitate additional video data for fine-tuning or test-time optimization of the DMs. With extensive quantitative, qualitative, and subjective evaluation on different benchmarks and tasks, our approach outperforms other state-of-the-art alternatives in terms of video quality and temporal consistency. The effectiveness of MeDM is also demonstrated in Figure 1.

Our main contributions are summarized as follows.

- We develop a practical coding algorithm based on optical flows to mediate independently estimated image scores. This mediation can seamlessly integrate into current state-of-the-art image-based denoising cycles, enhancing their ability for high-quality, consistent video-to-video translation.
- We suggest a workaround for modifying observed-space scores in latent-space Diffusion Models.
- We show that the proposed approach is effective in other real-world applications, yielding promising subjective evaluations. These applications include text-guided video

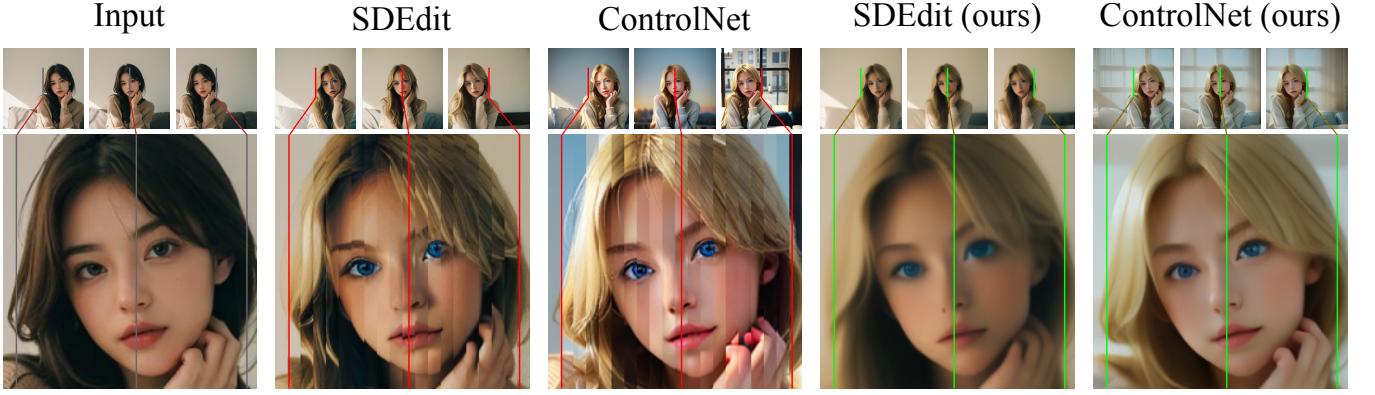


Figure 1: Text-guided video editing. From a series of video frames (top row), we extract a 20-pixel-wide vertical segment of pixels from each generated frame and stack them horizontally (bottom). A fluent video should reconstruct a stripe-free image. From left to right: input video (synthetic motion), results from SDEdit (Meng et al. 2022) and ControlNet (Zhang and Agrawala 2023), before and after applying the proposed harmonization framework. For SDEdit, we start the denoising process from $0.6T$ step. For ControlNet, we start from the pure noise and use the lineart (Chan, Durand, and Isola 2022) derived from the input as conditions. Prompt: *I girl, soft light, blonde hair, blue eyes.*

editing and video anonymization.

2 Related Work

In this section, we briefly review the recent relevant works of video-to-video generation and the mediation of multiple diffusion generations.

2.1 Diffusion-Based Video-to-Video Translation

To address the resource-intensive nature of training video DMs, various zero-shot techniques have arisen for adapting existing image DMs to video-to-video translation. For instance, Pix2Video (Ceylan, Huang, and Mitra 2023) employs a pre-trained structure-guided DM with self-attention features for modifying each denoising step. Yang et al. ensure consistency in global style and texture using a hierarchical inter-frame restricted DM, with the generated key frames propagating to in-between frames through temporal-sensitive patch matching and blending. ControlVideo (Zhang et al. 2023), an extension of ControlNet (Zhang and Agrawala 2023), incorporates cross-frame interaction for visual consistency and video quality preservation. Additionally, Chen et al. introduce motion-enhanced noise initialization and conditioning techniques for consistent and text-aligned videos, even with small datasets. In addition, there are other similar works (Xing et al. 2023; Wu et al. 2022; Shin et al. 2023; Liu et al. 2023; Qi et al. 2023; Khachatryan et al. 2023). Although the aforementioned techniques succeed in transforming videos to different styles, artifacts and flickering in the video remains clearly visible. In comparison, our approach enables efficient and effective harmonious video editing.

2.2 Mediating Multiple Diffusion Samples

MultiDiffusion (Bar-Tal et al. 2023) samples separate views from a DM. A view is a subset of pixels in the final image,

which have intersection. The DM gives diverse score estimates for each view's intersection during denoising. Bar-Tal et al. present view harmonization as a least-square optimization problem. In practice, they show a closed form solution as averaging views over the intersection during iterative denoising process, while testing mainly on simple views like rectangular crops and quantized masks from low-dimensional latent vectors of LDMs (Rombach et al. 2022). In contrast, our approach deals with more general views spanning temporal and spatial dimensions. These views can be subsets of consecutive and nonconsecutive pixels in video frames on the original dimension.

3 Optical Flow-Based Temporal Correspondence

In this section, we introduce Flow Coding, a physics-inspired coding at the core of the proposed method. Next, we propose an objective function with a closed-form solution for harmonization based on Flow Coding.

3.1 Flow Coding

Beginning with the notion that the color¹ of a point on a real-world material remains constant over time when the material possesses a time-invariant electronic structure, we apply this concept to rasterized digital video. Within such videos, points are collectively perceived within a pixel via the sampling theorem. We extend this idea by assuming that pixels through which the same set of underlying analog points are viewed should exhibit identical colors. To achieve this, we use a pixel repository to represent all rasterized points of the observed object. Each video frame serves as a view of this pixel repository, with views being defined by pixel-level optical flows, as depicted in Figure 2.

¹The shade, though subject to changes due to surrounding and environmental lighting, is intentionally disregarded in this study for the sake of brevity, with plans to address it in future research.

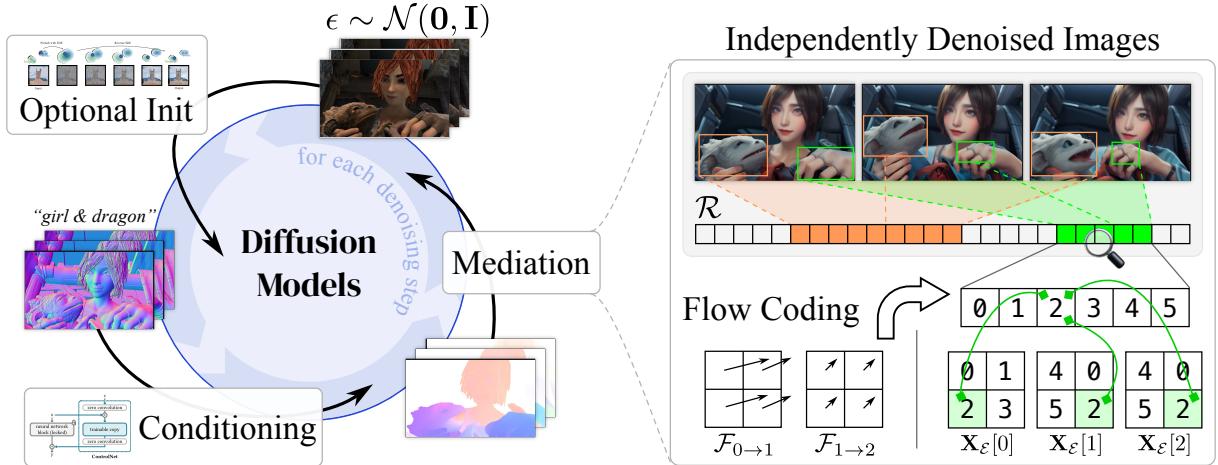


Figure 2: **Left: the proposed framework for video-to-video translation.** We extend the previous works in the image domain (Meng et al. 2022; Zhang and Agrawala 2023) to the video domain without any fine-tuning or iterative optimization. Our method mediates independent image score estimations after every denoising step, making them fluent motion pictures when viewed sequentially. Image credits to aforementioned authors. **Right: an illustration of a global pixel repository for a video and the proposed Flow Coding.** Video pixels are essentially views to the underlying objects. We construct an explicit pixel repository \mathcal{R} to represent the underlying world. \mathcal{R} is derived from the optical flows \mathcal{F} through the proposed Flow Coding and stores all unique pixels of the video. The encoded frames $\mathbf{X}_{\mathcal{E}}$ and the repository \mathcal{R} enable efficient harmonization of the divergent frame-wise score estimations during the generation process of Diffusion Models.

Algorithm 1: Flow Coding

Input: Flows \mathcal{F} and occlusions \mathcal{O} **Output:** Encoded frames $\mathbf{X}_{\mathcal{E}}$ and number of unique pixels n

```

1: if  $\mathcal{F}$  is forward flow then
2:   Temporally flip  $\mathcal{F}$  and  $\mathcal{O}$ 
3: end if
4: Let  $n = \text{height}(\mathcal{F}) \times \text{width}(\mathcal{F})$ 
5: Let encoded frame  $\mathcal{E}$  be a matrix with shape of  $\mathcal{F}$ 
6: Fill  $\mathcal{E}$  with codes from 0 to  $n - 1$ 
7: Let  $\mathbf{X}_{\mathcal{E}} = \{\mathcal{E}\}$ 
8: for each flow  $\mathcal{F}_i$  and occlusion  $\mathcal{O}_i$  in  $\mathcal{F}$  and  $\mathcal{O}$  do
9:   Let  $\mathcal{E}_{\text{prev}} = \mathcal{E}$ 
10:  Let  $\mathcal{V}$  be coordinate vectors that cover every point in
    the spatial dimension of  $\mathcal{F}_i$ 
11:  Let warped destination  $\mathcal{D} = \text{round}(\mathcal{V} + \mathcal{F}_i)$ 
12:  Propagate codes:  $\mathcal{E}[\mathcal{V}] = \mathcal{E}_{\text{prev}}[\mathcal{D}]$ 
13:  Let coordinate vectors of novel pixel  $\mathcal{N} = \mathcal{V}[\mathcal{O}]$ 
14:  Fill  $\mathcal{E}[\mathcal{V}]$  with codes from  $n$  to  $n + |\mathcal{N}| - 1$ 
15:   $n = n + |\mathcal{N}|$ 
16:   $\mathbf{X}_{\mathcal{E}} = \mathbf{X}_{\mathcal{E}} \cup \{\mathcal{E}\}$ 
17: end for
18: if  $\mathcal{F}$  is forward flow then
19:   Temporally flip  $\mathbf{X}_{\mathcal{E}}$ 
20: end if
21: return  $\mathbf{X}_{\mathcal{E}}, n$ 

```

In practice, we introduce Flow Coding, which uses per-frame optical flows and occlusions between adjacent frames. Flow Coding encodes video frame pixels as indices pointing

to a pixel repository. Initially, the repository contains all first-frame pixels. The process iteratively propagates codes from the first encoded frame to pixels covering the same underlying analog points in the next frame based on flows. Finally, novel pixels, specified by the occlusions, are added to the repository in each iteration. Refer to Algorithm 1 for the full procedure of Flow Coding.

3.2 Global Harmonization

After applying the Flow Coding algorithm, let $\mathbf{X}_{\mathcal{E}}$ represent the encoded frames (tensors of indices), and n be the count of unique pixels. We use \mathcal{R} to denote the shared pixel repository with a size of n . To decode frames, we adopt the indexing notation $\mathcal{R}[\mathbf{X}_{\mathcal{E}}]$, which retrieves pixel values from the repository. Placing this notation on the right side of an equation fetches pixel values, while on the left side, it denotes references to store pixels.

During the denoising process of a DM, we let \mathbf{X}^t be the noisy samples from the model evaluation of step t . The DM we consider here has no temporal information, and each video frame in \mathbf{X}^t is generated independently. Therefore, to enforce temporal constraint, we construct a temporal consistency loss using Frobenius norm:

$$\mathcal{L}^t = \|\mathcal{R}^t[\mathbf{X}_{\mathcal{E}}] - \mathbf{X}^t\|^2, \quad (1)$$

$$\mathcal{R}^t[\mathbf{X}_{\mathcal{E}}] \leftarrow G(\mathbf{X}^t), \quad (2)$$

where G is a function that mixes the pixels in \mathbf{X}^t into the ones in $\mathcal{R}^t[\mathbf{X}_{\mathcal{E}}]$, and \mathcal{R}^t is the pixel repository at time t . Notably, $\mathcal{R}^t[\mathbf{X}_{\mathcal{E}}]$ contains significantly fewer unique pixels than \mathbf{X}^t . For a concrete example in the right half of Figure 2, according to the optical flows, the bottom-left pixel of $\mathbf{X}_{\mathcal{E}}[0]$,

Listing 1: Unweighted global harmonization in Pytorch

```

1 # assume accumulate=True
2 cnt.index_put_((X_E,), 1)
3 repo.index_put_((X_E,), X_t)
4 avg = torch.where(cnt>0, repo/cnt, repo)

```

the bottom right pixels of $\mathbf{X}_{\mathcal{E}}[1]$ and $\mathbf{X}_{\mathcal{E}}[2]$ are all associated and synchronized with pixel 2 in \mathcal{R} . G must be applied and harmonized the corresponding three pixels in \mathbf{X}^t into a common values before they can be assigned to $\mathcal{R}^t[\mathbf{X}_{\mathcal{E}}]$.

The intuition behind Equation 1 is that under the physical constraint posed by the coding, we would like to find a frame composition that matches the unconstrained \mathbf{X}^t as close as possible. To solve the system analytically, we consider an DM-agnostic² solver following (Bar-Tal et al. 2023)

$$G_{avg} = \arg \min_G \mathcal{L}^t, \quad (3)$$

where G_{avg} performs unweighted averages on pixels in \mathbf{X}^t that are associated with the same pixel in \mathcal{R}^t according to $\mathbf{X}_{\mathcal{E}}$. In practice, G_{avg} can be implemented in Pytorch in three lines of code (Listing 1).

4 Mediating Image Diffusion Samples with Temporal Correspondence Guidance

In this section, we briefly review the essential formulation in DMs and introduce the proposed temporal correspondence guidance along with techniques to make it compatible with Latent Diffusion Models.

4.1 Diffusion Models

Given a predefined noise scheduling $\{\alpha^t | t \in [0, T]\}$ and a data point \mathbf{X}^0 , the noisy data sample \mathbf{X}^t for step t is defined as

$$\mathbf{X}^t = \sqrt{\alpha^t} \mathbf{X}^0 + \sqrt{1 - \alpha^t} \boldsymbol{\epsilon}^t, \text{ where } \boldsymbol{\epsilon}^t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (4)$$

Equivalently,

$$\mathbf{X}^0 = \frac{1}{\sqrt{\alpha^t}} (\mathbf{X}^t - \sqrt{1 - \alpha^t} \boldsymbol{\epsilon}^t), \quad (5)$$

$$\boldsymbol{\epsilon}^t = \frac{1}{\sqrt{1 - \alpha^t}} (\mathbf{X}^t - \sqrt{\alpha^t} \mathbf{X}^0). \quad (6)$$

A neural network parameterized by θ is often used to represent either the estimated data point $\mathbf{X}_\theta^0(\mathbf{X}^t)$ or the estimated noise residual $\boldsymbol{\epsilon}_\theta^t(\mathbf{X}^t)$ for the construction of a less noisy data distribution.

4.2 Temporal Correspondence Guidance

In this section, we introduce how to leverage the pixel repository \mathcal{R} to help an image DM generate temporally consistent

²A DM-aware solver $\min_{\mathbf{X}^t} \mathcal{L}^t$ is also possible and can potentially remove the need of post-hoc G . However, it would require less efficient gradient-based optimization through the DM parameters.

video frames. For each model evaluation of the DM parameterized by θ , we modify the noisy samples as follows:

$$\mathbf{X}_\theta^t \leftarrow (1 - w) \mathbf{X}_\theta^t + w \mathcal{R}^t [\mathbf{X}_{\mathcal{E}}], \quad (7)$$

or using the score estimation formulation following Ho and Salimans:

$$\boldsymbol{\epsilon}_\theta^t \leftarrow (1 - w) \boldsymbol{\epsilon}_\theta^t + w \mathcal{R}_\epsilon^t [\mathbf{X}_{\mathcal{E}}], \quad (8)$$

$$\mathcal{R}_\epsilon^t [\mathbf{X}_{\mathcal{E}}] \leftarrow G(\boldsymbol{\epsilon}_\theta^t), \quad (9)$$

with weight w on the harmonized samples that controls the strength of temporal correspondence guidance³ and \mathcal{R}_ϵ^t for scores is analogous to \mathcal{R}^t for noisy pixels. However, Equation 9 does not support LDMs (Rombach et al. 2022) because they perform the denoising process in a spatially smaller dimension, and G only operates on the original dimension. Moreover, their perceptual Autoencoders work with neither noisy samples \mathbf{X}^t nor noise predictions $\boldsymbol{\epsilon}^t$, so it is infeasible to cast the samples to a proper space when needed without further fine-tuning the Autoencoders.

To address this issue, we leverage Equation 5 to derive an inaccurate prediction of the original samples $\hat{\mathbf{X}}^{0,t}$ from the predicted noise $\boldsymbol{\epsilon}_\theta^t$. Empirically, we find that $\hat{\mathbf{X}}^{0,t}$ can be faithfully cast between latent space and observed space by pre-trained Autoencoders even in the early stage of denoising process. Therefore, incorporating Equation 6, we can replace Equation 9 with

$$\mathcal{R}_\epsilon^t [\mathbf{X}_{\mathcal{E}}] \leftarrow \frac{1}{\sqrt{1 - \alpha_t}} \left(\mathbf{X}_t - \sqrt{\alpha_t} \Phi_e \left(G \left(\Phi_d \left(\hat{\mathbf{X}}^{0,t} \right) \right) \right) \right) \quad (10)$$

to provide compatibility of LDMs, where Φ_e and Φ_d are the encoder and the decoder of the Autoencoder, respectively.

4.3 Enhancing Video Sharpness

In our early experiments, we found that harmonizing independent score estimations from standard DMs directly results in suboptimal blurry video due to significant visual structure disparities between optical flows and DM estimates. To address this, we incorporate SDEdit (Meng et al. 2022) and ControlNet (Zhang and Agrawala 2023) into our approach. These methods integrate structural and semantic priors, with ControlNet excelling in fusing input structural information and pre-trained DM semantic priors.

5 Experiments

In this section we show that our method outperform other baseline methods in various benchmarks and applications, supported by extensive quantitative metrics, qualitative comparison and human evaluation.

Setup For all video generation, we use DDIM scheduler (Song, Meng, and Ermon 2021) with 20 denoising steps and set the temporal correspondence guidance scale w to 0.8. We use Stable Diffusion (Rombach et al. 2022) with the checkpoint from Civitai⁴ as our pre-trained DM to match up with

³It is possible to use the extrapolation formulation, but we found this lead to serious early saturation in our experiments.

⁴<https://civitai.com/models/43331/majicmix-realistic>

the baseline methods. For all estimated optical flows, we use UniMatch (Xu et al. 2023), and we use Informative Drawings (Chan, Durand, and Isola 2022) for all lineart translations which serve as the input conditions to ControlNet unless further specified. We do not provide a dedicated prompt to each video generation (except for text-guided video editing). Instead, reasonable texture and colors are automatically inferred through our cross-frame harmonization together with ControlNet’s guess mode. Finally, We conduct online user studies for all tasks in this section. The feedback is collected from a total of 71 people with diverse background. For most of the questions, they are asked to give a score ranging from 1 (the worst) to 5 (the best).

5.1 Video Rendering

Our method is capable of efficiently rendering high quality videos solely from 3D assets. We run the experiments on MPI Sintel (Butler et al. 2012) and Virtual KITTI 2 (Cabon, Murray, and Humenberger 2020), synthetic video datasets that provides ground-truth (GT) 3D assets, including optical flows, occlusions, position information (depth, normal) and rendered animations. We consider two variants to this task.

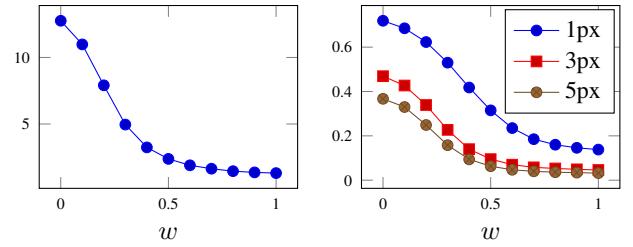
(Full) Rendering. All methods generate videos solely from optical flows and position information (we consider normal maps here). The result video should differ greatly from the original rendered animation in terms of surface texture and colors. We compare our framework with 5 baseline methods: ControlNet (Zhang and Agrawala 2023), ControlVideo (Zhang et al. 2023), Control-A-Video (Chen et al. 2023), Video ControlNet (Chu, Lin, and Chen 2023) and Rerender A Video (Yang et al. 2023).

Assistive Rendering. The methods capable of this task begin with adding noise to the pre-rendered videos and perform the denoising process from $0.5T$ step following SDEdit (Meng et al. 2022). The generated video should be similar to the original animation and incorporating the realistic prior from the pre-trained DM.

For our method, we use the lineart derived from the normal maps as the input conditions to ControlNet⁵. The qualitative comparisons in Figure 4 show that our generated videos are high-quality, fluent and faithfully follow the input conditioning for full video rendering. For assistive rendering, our method further enhances the realism of objects with complex texture, such as hair.

In addition to the scanning, we provide an objective and automatic quantitative measure for the video fluency. We estimate the optical flows from the generated videos and compare the flows with the GT flows. We use the average endpoint error (EPE) between the flows to assess the temporal consistency in videos. Figure 3 validates that a fluent video leads to a lower EPE, and vice versa. As shown in Table 1, our method has the lowest EPEs among all methods in each

⁵Although the pre-trained ControlNet for the depth and normal maps are available, the depth maps fail to encode subtle details like facial features, and the model for normal maps is trained on the estimated normal maps which differ greatly from the GT ones



(a) End-point-error (EPE) (b) Portion of errors larger than

Figure 3: **EPE curves over the temporal correspondence guidance strengths.** The videos are generated from pure noise and conditioned on the assets from MPI Sintel. ($w = 0$ refers to the vanilla ControlNet).

| | Method | Rendering | Assist. Rendering |
|------------|------------------|--------------|-------------------|
| MPI Sintel | ControlNet | 12.757 | 5.924 |
| | ControlVideo | 12.757 | N/A |
| | Control-A-Video | 7.016 | N/A |
| | Video ControlNet | 12.878 | N/A |
| | Rerender A Video | 8.927 | 6.711 |
| | Ours (Est. flow) | 1.501 | 1.570 |
| VKITTI 2 | Ours (GT flow) | 1.456 | 1.202 |
| | Animation | | 0.403 |
| VKITTI 2 | ControlNet | 12.575 | 5.070 |
| | Video ControlNet | 15.285 | N/A |
| | Ours (Est. flow) | 2.857 | 2.695 |
| | Ours (GT flow) | 2.483 | 2.217 |
| Animation | | | 1.737 |

Table 1: **EPE for video rendering on MPI Sintel and Virtual KITTI 2.** The best methods are in **bold**.

task, and it is the closest to the one of the animation on which the flow estimator is trained.

Finally, we close this task with the subjective test. We ask the participants how the presented videos are in terms of quality, sharpness and fluency. We further ask them to rate the realism of the videos, and the results in Table 2 shows that in addition to having the highest quality, our videos are even more realistic than the animation from the dataset.

5.2 Video-to-Video Translation

To show that our method performs well even without high precision optical flows, we conduct evaluation on real videos. Considering two applications of video-to-video translation, we perform text-guided video editing on the videos in DAVIS 2016 (Perazzi et al. 2016) and video anonymization on the videos from CelebV-HQ (Zhu et al. 2022).

Text-Guided Video Editing. We take videos from DAVIS 2016 and caption them with seemingly absurd descriptions to test the models’ ability in combining conflicting ideas into the same canvas. In addition to using the lineart for ControlNet, we also test the InstructPix2Pix version of ControlNet,

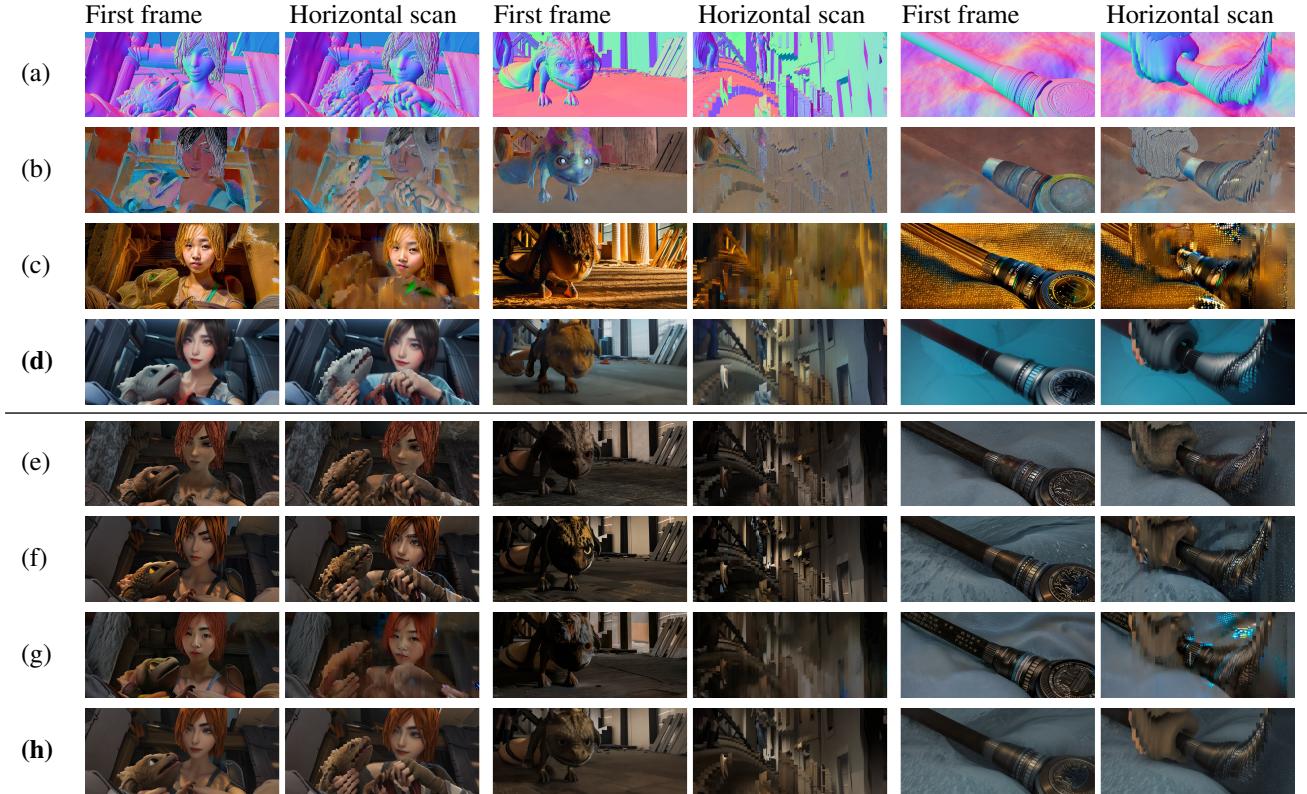


Figure 4: **Visual comparison for video rendering (a-d) and assistive rendering (e-h).** For each method, we provide the first frame and the horizontal scan for visual quality and temporal coherence comparison, respectively. The horizontal scan is composed of the left-shifting vertical segments of pixels from each video frame. A fluent video should reconstruct a stripe-free scan in static area (e.g. background, see the GT video (a,e)). From top to bottom, the presented videos are from (a) input GT normals, (b) Control-A-Video, (c) Rerender A Video, (d) **ours (estimated flow)**, (e) input GT animations, (f) ControlNet, (g) Rerender A Video and (h) **ours (estimated flow)**.

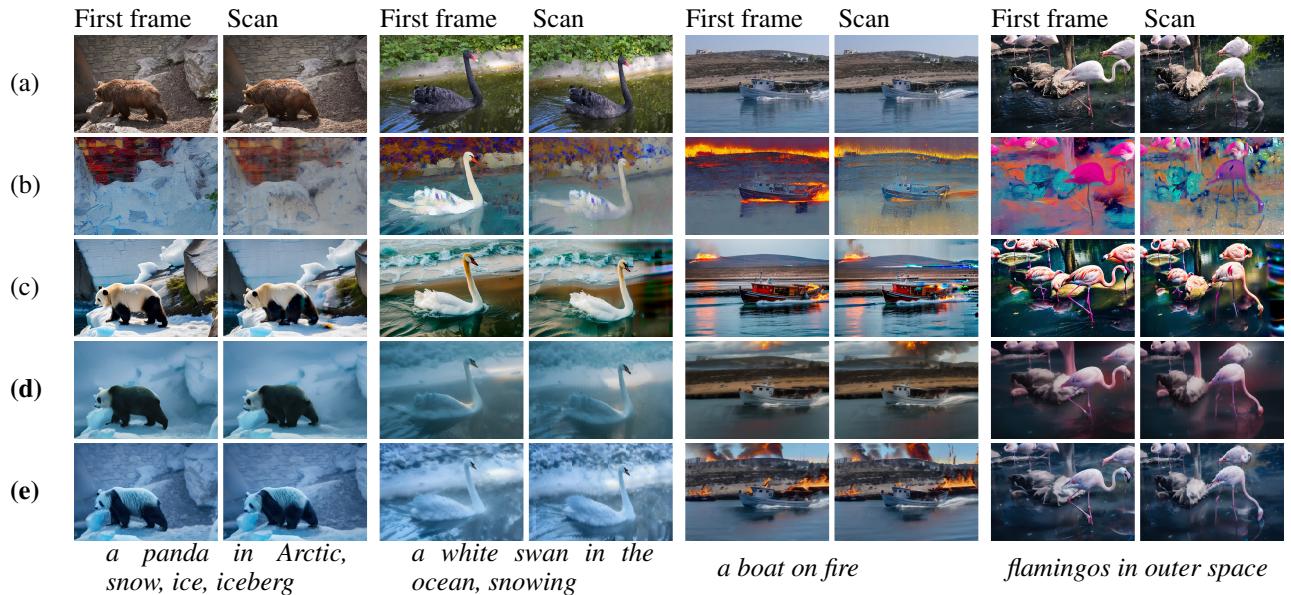


Figure 5: **Visual comparison for text-guided video editing.** For each method, we provide the first frame and the horizontal scan following Figure 4. From top to bottom, the presented videos are from (a) input video, (b) Control-A-Video, (c) Rerender A Video, (d) **ours (Lineart)** and (d) **ours (InstructPix2Pix)**.

| Method | Video quality | Realism |
|---------------|------------------|--------------|
| Non-Assistive | ControlNet | 1.662 |
| | ControlVideo | 2.437 |
| | Control-A-Video | 3.662 |
| | Video ControlNet | 2.197 |
| | Rerender A Video | 2.183 |
| | Ours | 4.423 |
| Animation | 4.423 | 3.113 |
| Assistive | ControlNet | 2.465 |
| | Rerender A Video | 1.887 |
| | Ours | 4.380 |
| | | 3.958 |

Table 2: Human evaluation for video rendering on Sintel.

| Method | Video quality | Text alignment |
|---------------------|---------------|----------------|
| ControlNet | 2.377 | 3.289 |
| Pix2Video | 1.451 | 1.592 |
| ControlVideo | 2.289 | 2.430 |
| Control-A-Video | 2.634 | 1.859 |
| Rerender A Video | 3.042 | 3.099 |
| Ours (Lineart) | 4.042 | 3.810 |
| Ours (Instruct P2P) | 3.901 | 4.338 |

Table 3: Human evaluation for text-guided video editing on DAVIS 2016. The best methods are in **bold**.

which directly consumes real videos. We further consider an inversion-based approach, Pix2Video (Ceylan, Huang, and Mitra 2023), for this task and provide visual comparisons in Figure 5. Together with the subjective scores in Table 3, our methods, especially for the InstructPix2Pix variant, produce the best video quality and image-text alignment in class.

Video Anonymization. Anonymizing videos allows ones to analyze behaviors in videos without exposing individuals’ identities. In order to make the anonymized video useful, we need to preserve anything in the video except the facial identity. We leverage the fact that human visual perception exhibits a remarkable sensitivity to human faces while our ability to detect and recognize other objects is not as specialized. We add noise to a video with a strength of $0.5T$, which is strong enough to erase the identity while preserving other objects and the background scene, and perform denoising using the proposed method to obtain the anonymized video. We conduct an user study to validate the effectiveness of our method. We compare with DeepPrivacy (Hukkelås, Mester, and Lindseth 2019) and report the results in Table 4. We show that a larger portion of the participants⁶ cannot recognize the celebrities from our anonymized videos compared to the ones from DeepPrivacy. The participants also consider our method produce samples that are more realistic and preserves more details in the original videos.

⁶We only consider the participants who know of the celebrities in advance.

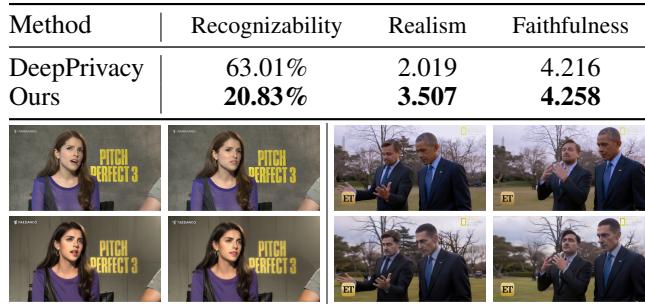


Table 4: User feedback on video anonymization. **Recognizability:** the proportion of participants correctly identifying the celebrity. **Realism:** how realistic is the anonymized video. **Faithfulness:** how much content except the identity is preserved. **Upper row images:** the original clips. **Lower row images:** the anonymized clips (ours only).

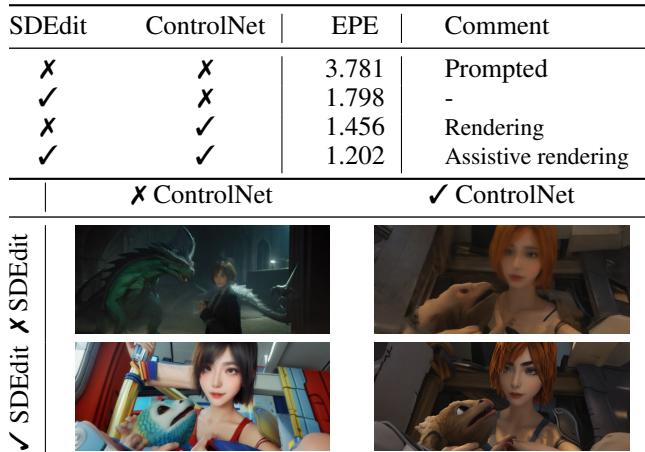


Table 5: Ablation study for ControlNet and SDEdit.

5.3 Without ControlNet or SDEdit

Since we heavily utilize ControlNet and SDEdit in our experiments, we provide an ablation study in Table 5 to justify this usage. From both qualitative and quantitative results, we show that removing either one of the two increases the EPE. When using neither ControlNet nor SDEdit, our method fails to faithfully follow the visual structure hinted by the motion in optical flows.

6 Conclusion and Future Work

In this work, we propose a novel approach, MeDM, which employs optical flows to guide temporal correspondence. This methodology allows for efficient and high-quality video-to-video translation while preserving temporal consistency. By aligning pixel correspondences across video frames using optical flow-derived guidance, MeDM avoids the need for additional video data and achieves superior results on multiple benchmarks compared to existing methods.

While we leverage discretized optical flows for adjacent frames in this work, we are also interested in the potential of point tracking methods (Harley, Fang, and Fragki-

adaki 2022; Doersch et al. 2023; Wang et al. 2023), which track pixels across occlusions and have continuous coordinate system. This helps the proposed Flow Coding build more accurate pixel correspondences. Despite their advantages, these approaches are currently costlier and less robust. Another potential enhancement for our method involves enabling structural translation. While optical flows incorporate structural details, they might not align well with objects that undergo significant contour changes.

References

- Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *International Conference on Machine Learning*.
- Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instruct-Pix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Butler, D. J.; Wulff, J.; Stanley, G. B.; and Black, M. J. 2012. A naturalistic open source movie for optical flow evaluation. In *Proceeding of the European Conference on Computer Vision*, 611–625.
- Cabon, Y.; Murray, N.; and Humenberger, M. 2020. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*.
- Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2Video: Video Editing using Image Diffusion. *arXiv preprint arXiv:2303.12688*.
- Chan, C.; Durand, F.; and Isola, P. 2022. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Chen, W.; Wu, J.; Xie, P.; Wu, H.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2305.13840*.
- Chu, E.; Lin, S.-Y.; and Chen, J.-C. 2023. Video ControlNet: Towards Temporally Consistent Synthetic-to-Real Video Translation Using Conditional Image Diffusion Models. *arXiv preprint arXiv:2305.19193*.
- Doersch, C.; Yang, Y.; Vecerik, M.; Gokay, D.; Gupta, A.; Aytar, Y.; Carreira, J.; and Zisserman, A. 2023. TAPIR: Tracking Any Point with per-frame Initialization and temporal Refinement. *arXiv preprint arXiv:2306.08637*.
- Harley, A. W.; Fang, Z.; and Fragkiadaki, K. 2022. Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. In *Proceeding of the European Conference on Computer Vision*, 59–75.
- Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. *arXiv preprint arXiv:2210.02303*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 6840–6851.
- Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- Hukkelås, H.; Mester, R.; and Lindseth, F. 2019. Deep-Privacy: A Generative Adversarial Network for Face Anonymization. In *Advances in Visual Computing*, 565–578.
- Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*.
- Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; and Jia, J. 2023. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*.
- Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.
- Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, 36479–36494.
- Shin, C.; Kim, H.; Lee, C. H.; Lee, S.-g.; and Yoon, S. 2023. Edit-A-Video: Single Video Editing with Object-Aware Consistency. *arXiv preprint arXiv:2303.07945*.
- Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; Parikh, D.; Gupta, S.; and Taigman, Y. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv preprint arXiv:2209.14792*.
- Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Wang, Q.; Chang, Y.-Y.; Cai, R.; Li, Z.; Hariharan, B.; Holynski, A.; and Snavely, N. 2023. Tracking Everything Everywhere All at Once. *arXiv preprint arXiv:2306.05422*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. *arXiv preprint arXiv:2212.11565*.

Xing, J.; Xia, M.; Liu, Y.; Zhang, Y.; Zhang, Y.; He, Y.; Liu, H.; Chen, H.; Cun, X.; Wang, X.; et al. 2023. Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance. *arXiv preprint arXiv:2306.00943*.

Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; Yu, F.; Tao, D.; and Geiger, A. 2023. Unifying Flow, Stereo and Depth Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. *arXiv preprint arXiv:2306.07954*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077*.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In *Proceeding of the European Conference on Computer Vision*, 650–667.

A System

We run all experiments on the system with the specs listed below.

| | |
|-----|--|
| CPU | Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz |
| GPU | NVIDIA RTX A5000 24GB |
| OS | Ubuntu 20.04.6 LTS |

We heavily utilize the codebase of Hugging Face Diffusers⁷ in this work. The ControlNet weights for lineart⁸ and InstructPix2Pix⁹ are also hosted on Hugging Face. The videos samples from Rerender A Video (Yang et al. 2023) are generated without their patch-based propagation due to the absence of official implementation.

B Reducing Discretization Errors

Per-pixel optical flows are utilized in our study. These flows require rounding during Flow Coding to obtain the integer coordinates of the destination, leading to errors in code propagation due to discretization. To address this problem, we introduce a technique to mitigate these errors.¹⁰

Optical flow for distant frames. When optical flows from distant frames are available alongside those from adjacent frames, these distant flows can be employed to counteract accumulated errors stemming from discretization. Specifically, when d denotes the number of frames preceding the current one, the encoded frame \mathcal{E}_{i+1} prioritizes codes from the distant frame \mathcal{E}_{i-d} over the previous frame \mathcal{E}_i , such as

$$\mathcal{E}_{i+1}[\mathcal{V}] \leftarrow \mathcal{E}_{i-d}[\mathcal{D}_{i-d} \cap \mathcal{O}'_{i-d} \cap \mathcal{O}'_i] \cup \mathcal{E}_i[\mathcal{D}_i \cap \mathcal{O}_{i-d} \cap \mathcal{O}'_i], \quad (11)$$

where \mathcal{V}, \mathcal{D} and \mathcal{O} are the coordinate vectors, destinations and occlusion, respectively, in line with Algorithm 1. The index $i - d$ is clamped to be non-negative. Nevertheless, our empirical observations suggest that distant optical flows tend to exhibit greater estimation errors, implying that their inclusion might not necessarily lead to improved performance

C Local Harmonization

Global harmonization represents a rigorous process, enforcing consistent pixel values over time to visualize a consistent physical point. This approach, however, disregards alterations in shading due to varying lighting conditions. In this section, we introduce a preliminary solution that offers a more relaxed approach to this procedure. Drawing inspiration from the notion that distant pixels are likely to exhibit greater discrepancies than adjacent ones, we organize the pixels observing the same object point into a 1D tensor according to their temporal sequence. To enhance coherence, we apply a smoothing operation to this tensor by convolving it with a predefined Gaussian kernel denoted as K . In practice, we have to construct an inverse pixel repository as

$$\tilde{\mathcal{R}}_c = \{(i, y, x) \mid \mathbf{X}_{\mathcal{E}}[i, y, x] = c\} \quad (12)$$

⁷<https://huggingface.co/docs/diffusers/index>

⁸https://huggingface.co/llyasviel/control_v11p_sd15_lineart

⁹https://huggingface.co/llyasviel/control_v11e_sd15_ip2p

¹⁰It is also possible to leverage point tracking methods (Harley, Fang, and Fragkiadaki 2022), but we leave it for future work.

that stores the coordinates of pixels associated with each code c where i indicates the frame index, and y and x represent the 2D spatial coordinates. With the inverse repository, we can formally define the local harmonization as¹¹

$$\mathcal{R}_c^t = G_{conv}(\mathbf{X}^t) = \mathcal{P}\left(\mathbf{X}^t [\tilde{\mathcal{R}}_c^t]\right) * K, \forall c \in [0, n], \quad (13)$$

where \mathcal{P} is a reflection padding function to maintain the size, $|\mathbf{X}^t[\tilde{\mathcal{R}}_c^t]|$, after convolution and permit batch implementations. A drawback to this approach is that Equation 12 cannot be executed in parallel and has a time complexity of $\mathcal{O}(TWH)$, i.e. the number of pixels in \mathbf{X} . The local harmonization approach is not physically correct but it provides a fluent transition of color in time that is visually appealing.

Execution time. We provide an analysis of execution time required for both global and local harmonization in Figure 6. The results are collected by generating the entire MPI Sintel dataset with GT optical flows. The insights from Figure 6a demonstrate that the total time taken for video synthesis using global harmonization is merely three times that of employing the original Stable Diffusion in combination with ControlNet. Moreover, Figure 6b underscores that in

¹¹ \mathcal{R}_c^t no longer represents a single shared pixel but a tensor of length $|\tilde{\mathcal{R}}_c|$ here. We reuse the notation for convenience.

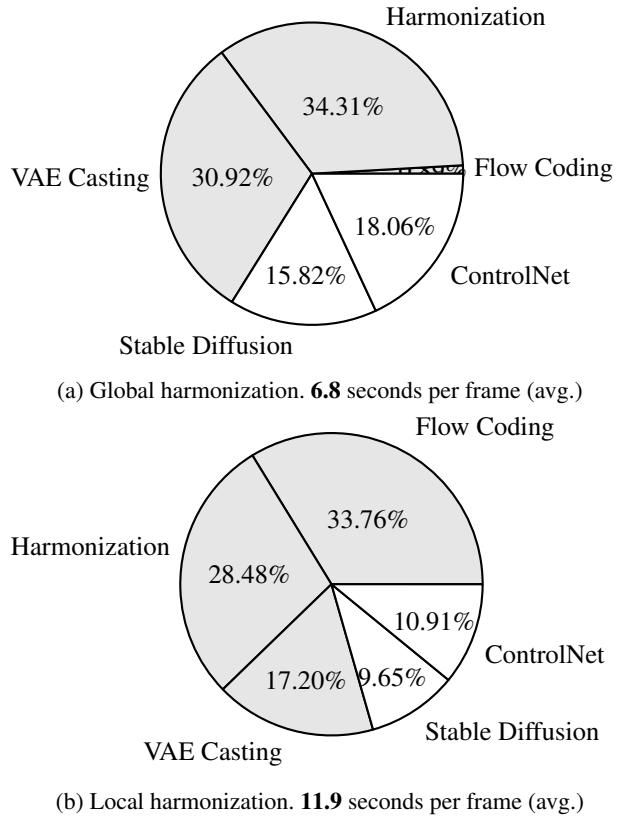
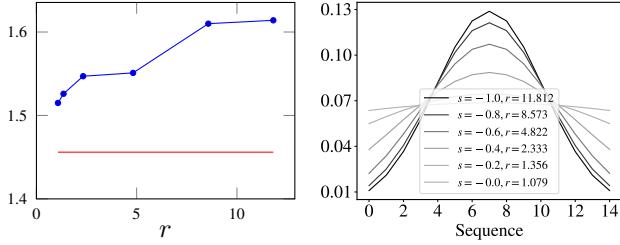


Figure 6: **Execution time distribution for generating videos in the entire MPI Sintel.** The components proposed in this work are shaded.



(a) EPE for local harmonization

(b) Convolution kernels

Figure 7: Local harmonization results on MPI Sintel. The red line in 7a is the EPE of global harmonization.

the case of local harmonization, both the time taken for initial coding (building reverse repository) and the subsequent harmonization (convolution) in each diffusion step are notably extended. As a result, the execution time for the local harmonization approach is almost five times that of utilizing the original Stable Diffusion combined with ControlNet.

C.1 Experiments

In this section, we conduct a quantitative assessment of the efficacy of local harmonization using various Gaussian kernels. The discrete 1D Gaussian kernel K , characterized by a width ℓ , exhibits an impulse response defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\ell/2}{\sigma}\right)^2\right), \quad x \in \mathbb{Z} \cap [0, \ell], \quad (14)$$

where σ denotes the standard deviation regulating the kernel's sharpness. Prior to application, we normalize the kernel to ensure its weights sum to one. We control the harmonization strength by varying σ . To obtain a proper scale of σ , we define a set of seeds s ranging from -1.0 to 0.0 at intervals of 0.2 , and set $\sigma = 100^s + 0.2$. Additionally, r signifies the ratio of the maximum weight to the minimum weight in K , serving as an intuitive indicator of the convolutional harmonization's smoothing level, as illustrated in Figure 7b. We graph the End Point Error (EPE) of generated videos against r in Figure 7a. The outcomes affirm that videos exhibit enhanced fluidity as r diminishes, albeit at the cost of a more rigorous global consistency constraint.

D User Study

We include a representative screenshot as an illustration of our user study in Figure 8. Our survey on anonymization is conducted interactively and possesses a unique aspect: each participant encounters it only once. This survey pertains to guessing the identity concealed within the anonymized rendition of each video. Notably, participants are provided no foreknowledge or contextual information about the celebrity under consideration. This characteristic mandates that the source videos utilized for assessing various methods cannot be reused.

E Additional Samples

We provide additional visual samples of our generated videos in Figure 9 and Figure 10.

Question 1 / 6

Anonymization

▶ 003 / 0:16

Do you know of the celebrity in the video after watching the entire video? *

Yes, I know of at least one of them
 No, I do not know of any of them

Had you guessed the identities of them correctly before watching the answer in the latter part of the video

Yes, I guessed correctly. The anonymization is not effective.
 No, I guessed wrongly. The anonymization is effective.

Does the anonymized video look natural, realistic, fluent and without flickering? *(Compared to the original video)

| | | | | |
|-----------------------------------|---|-------------------------|---|---|
| 1 | 2 | 3 | 4 | 5 |
| Very unnatural. Severe flickering | | Very natural and fluent | | |

Except for the identity, does the anonymized video preserve all other information? *(e.g. background, clothes, objects, etc.)

| | | | | |
|-----------------|---|----------------------|---|---|
| 1 | 2 | 3 | 4 | 5 |
| Completely lost | | Completely preserved | | |

Figure 8: Screenshot of the survey for anonymization comparison. In the video, we first show the anonymized version to the participants, then we show the original version. We have made sure that the participants know of the celebrity before taking their responses into account. More videos can be found in the supplementary materials.



Figure 9: **Samples of video rendering on MPI Sintel.** The videos are generated using the lineart ControlNet from pure noise.



Figure 10: **Samples of assistive video rendering on MPI Sintel.** The videos are generated using the lineart ControlNet with a noise level of $0.5T$.