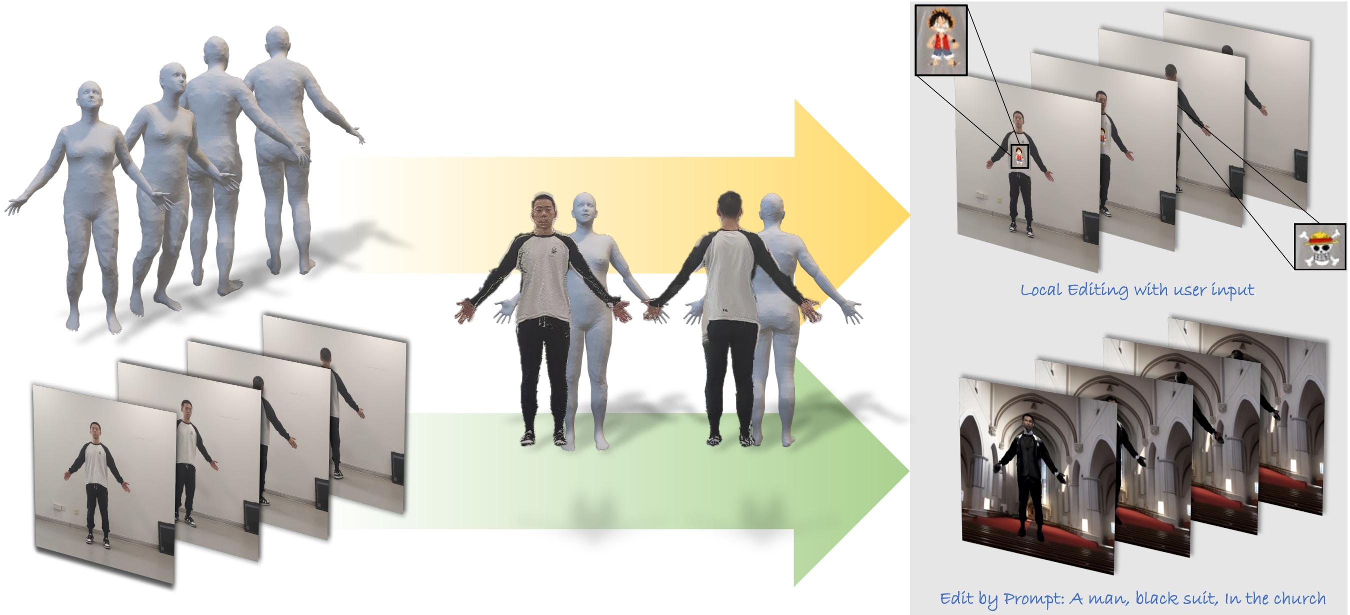
# HeroMaker: Human-centric Video Editing with Motion Priors

Anonymous Authors


Local Editing with user input
Edit by Prompt: A man, black suit, In the church

(a) Human Mesh acquisition (b) Dual Canonical Fields Learning (c) Diverse Video Editing

**Figure 1: We present HeroMaker, a new video representation with motion priors for human-centric video editing, which contains human motion warping, margin refinements, and dual canonical fields. As illustrated in (a), our model employs the body mesh to portray the structure information of people in the video. From (a) to (b), our model reconstructs the video with explicit human motion warping and neural margin refinements between dual canonical fields and each human-centric video frame. (c) shows the two editing results from HeroMaker, which are temporally consistent and plausible.**

## ABSTRACT

Video generation and editing, particularly human-centric video editing, has seen a surge of interest in its potential to create immersive and dynamic content. A fundamental challenge is ensuring temporal coherence and visual harmony across frames, especially in handling large-scale human motion and maintaining consistency over long sequences. The previous methods, such as diffusion-based video editing, struggle with flickering and length limitations. In contrast, methods employing Video-2D representations grapple with accurately capturing complex structural relationships in large-scale human motion. Simultaneously, some patterns on the human body appear intermittently throughout the video, posing a knotty problem in identifying visual correspondence. To address the above problems, we present HeroMaker. This human-centric video editing framework manipulates the person's appearance within the input video and achieves inter-frame consistent results. Specifically, we propose to learn the motion priors, transformations from dual canonical fields to each video frame, by leveraging the body mesh-based human motion warping and neural deformation-based margin refinement in the video reconstruction framework to ensure the semantic correctness of canonical fields. HeroMaker performs human-centric video editing by manipulating the dual canonical fields and combining them with motion priors to synthesize temporally coherent and visually plausible results. Comprehensive experiments demonstrate that our approach surpasses existing methods regarding temporal consistency, visual quality, and semantic coherence.

## CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

## KEYWORDS

Human-centric Video Editing, Diffusion Model, Motion Priors

*ACM MM, 2024, Melbourne, Australia*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

# 1 INTRODUCTION

Human-centric video editing focuses on modifying the individual within a given video and generating temporally coherent results. This technique has numerous potential applications, such as media content production, virtual reality, and video games. A pivotal challenge in human-centric video editing is maintaining coherence and harmonious results across frames when people can move freely in the video.

Recent diffusion-based video editing explores extracting and incorporating various structural correspondences using infer-frame attention maps [6, 8, 32, 55], optical flows [11, 58] and nn-fields [19]. Although the temporal consistency has improved, it still struggles with flickering and length limitations. Alternatively, some researchers have explored video-2D representations, storing the information of the video in atlases [30] or canonical images [46] to propagate changes over time. However, it grapples with accurately capturing complex structural relationships in large-scale human motion. Moreover, many studies have focused on reconstructing a human body in 3D and attempting to edit it. While promising, these methods often present challenges regarding cost, size, and unfriendly user-controlled environments due to the semantic-less texture maps and the requirement for long-term optimization processes.

Since each human part is unique, patterns on the human body appear intermittently throughout the video due to self-occlusion, posing a knotty problem in identifying correspondence to ensure consistent video editing. As a method of video-2D representation, CoDeF[46] builds correspondence via learning a neural deformation field from a canonical image to each frame and does the video editing on the canonical field, which improves the correspondence between frames. Although it achieves high-fidelity reconstruction, the canonical image differs from natural images, leading to difficulties when editing with image editing tools, including ControlNet [62], and resulting in editing challenges to generate semantically plausible results.

In light of CoDeF's success, we leverage the human body mesh to obtain semantic human canonical images, which provide structural and texture correspondence in 3D space. Our framework defines motion priors, incorporating human motion warping, neural margin refinements, and dual canonical fields to achieve this goal. To obtain the motion priors from a given video, our model employs an off-the-shelf human mesh estimator to set up an initial body mesh. Then, we refine the body mesh in a two-step optimization to close its shape to the person's in videos to ensure more accurate human motion warping. Resorting the motion priors, our model defines the dual canonical fields with a frontal and back body mesh under the A-pose to obtain the vast majority of human body information. Subsequently, it reconstructs the video with explicit human motion warping and neural margin refinements between dual canonical fields and each human-centric video frame. Additionally, our model supports diverse user interactions for modifying the videos. HeroMaker performs human-centric video editing by manipulating the semantic-aware dual canonical fields. Together with the motion priors, it synthesizes temporally coherent and visually plausible results.

We summarize our contributions as follows:

- We propose a new human-centric video representation combining motion priors and deformation fields to reconstruct and edit the video.
- We leverage the motion priors with human motion warping based on body mesh, neural margin refinements, and dual canonical fields to identify accurate structural correspondence and produce coherent results.
- Extensive experiments demonstrate that our model could produce temporal coherent and plausible results, especially during large-scale human motion.

# 2 RELATED WORK

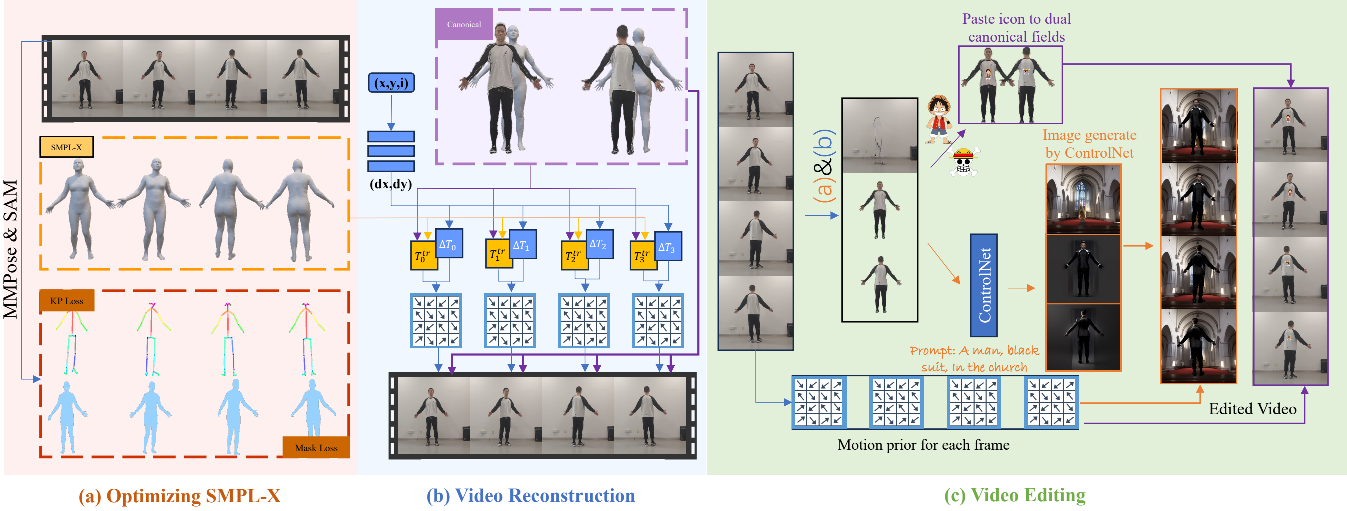
## 2.1 Text-to-Video Generation and Editing.

Recent works attempt to extend a latent diffusion model into a T2V editing model [2, 5, 6, 15–17, 19, 21, 22, 24, 25, 27, 32, 34, 35, 38, 41, 44, 48, 49, 52, 55, 56, 59–61, 65]. Tune-A-Video [55] and Control-A-Video [8] extend a latent diffusion model to the spatial-temporal domain and finetune it with source videos. However, they still have difficulties in modeling complex motions and long sequences. Text2Video-Zero [32] and ControlVideo [64] use ControlNet [62] to preserve the per-frame structure but struggle to temporal consistency. FateZero [48] and vid2vid-zero [53] use attention maps to enhance shape-aware editing based prompt-to-prompt [23], but they still have temporal issues. Rerender-A-Video [58], TokenFlow [19], and VideoControlNet [27] utilize optical flow to control inter-frame relationships to improve consistency. However, they still face challenges when addressing large-scale human motion and rotation issues. TokenFlow [19] enforces linear combinations between diffusion features based on source correspondences. However, the pre-defined combination weights are not adapted to all videos, resulting in high-frequency flickering. Because TokenFlow [19] needs to cache information for each frame when processing long-sequence videos, resulting in insufficient memory, we will not compare it with this method.

The above methods explore the augmentation of inter-frame attention modeling on a diffusion model. They ensure the correct spatial structure but still challenge temporal consistency. Recently, AnimateDiff [21] presented a motion module trained on extensive video data without a fineturing diffusion model, improving temporal consistency. Furthermore, human-centric videos have further explored some workss [26, 57] and achieved visually plausible results. However, generating the same effect in videos in a small amount of video data poses challenges. Unlike these works, our method leverages human motion priors to achieve text-guided video editing effectively.

## 2.2 Temporal Propagation in video editing.

Another significant line of video editing work relies on a powerful video representation. VideoSnap [66] compresses videos using spatio-temporal feature maps into one or several images and then trains an expansion network to transform these images back into videos. The layered neural atlas [30] factorizes the input video using a layered presentation. It maps the subject/background of all frames using 2D UV maps as an intermediate editing representation. Once the layered neural atlas is learned, editing can occur either on keyframes or on the atlas itself, and the editing results consistently



(a) Optimizing SMPL-X

(b) Video Reconstruction

(c) Video Editing

**Figure 2: We propose a multi-stage framework for human-centric video editing. We first acquire the motion priors for each frame based on SMPL-X [47] (a). Building upon the motion priors, we devise an editing-friendly video representation to reconstruct the input video (b). Then, our optimized video representation enables superior editing performance as in (c).**

propagate to other frames [4, 7, 13, 28, 33]. CoDeF [46] incorporates the 3D deformation field with the 2D hash-based canonical image to improve the video representative capability further. However, Video Snapshot [66], atlas [30], and canonical image [46] all utilize optical flow to help to predict the relationships between each frame. They encounter difficulties in reconstructing and editing videos with large-scale human motion. This leads to incorrect correspondence and texture information, resulting in unnatural results.

## 2.3 3D Human Reconstruction and Editing

3D human reconstruction and editing are closely related to human-centric video editing tasks. Many papers aim to reconstruct an accurate human body and texture using the SMPL+D model [1, 18] or implicit functions [20, 29, 54] through monocular videos. However, most focus on reconstructing and driving more accurate human models without considering editing effects and friendliness. Some works attempt to edit in 3D, SINE [3] and SKED [42] support editing a local region of the base NeRF [43]. Dyn-E [63] and Control4D [51] propose to edit the contents of dynamic NeRFs. However, Dyn-E [63] can only edit the local appearance with user manipulation. Control4D [51] needs multi-view videos as input and can only handle videos with small motions and short video lengths. Recently, DynVideo-E [37] has attempted to edit monocular videos in 3D through text. However, it is not user-friendly for reconstruction and editing operations to take tens of hours to complete. Although these methods can produce high-fidelity results, their cost, size, and controlled environment are unfriendly to users. Instead, video or image editing frameworks are more likely to avoid these shortcomings. Our method introduces the motion priors to the human body. Then, we convert it into pixel position relationship conversion between each frame. While retaining the correctness of the 3D structure, it transforms the task into 2D image editing, which is also one of the primary motivations of our work. Our work focuses on proposing a new human-centric video representation to solve problems with the image or video editing task. Meanwhile, we will not compare these methods. [37, 51, 63]. due to some recent work not being open-source yet.

## 3 METHOD

Given a human-centric video, we aim to modify its visual attributes based on diverse user interactions while maintaining correct structural correspondence and temporal consistency. We tackle this problem with a multi-stage framework, namely reconstructing and editing. As shown in Fig. 2, HeroMaker introduces a novel video representation by leveraging the motion priors based on the SMPL-X [47], which establishes the transformation correspondence from the canonical field to each video frame. In the following, we first illustrate the motion priors in Sec.3.1, and then our novel video representation is elaborated in Sec.3.2, followed by details of the editing procedure and applications of the whole framework in Sec.3.3.

## 3.1 Preliminary: Motion Priors

As previously stated, the visual quality of video editing largely depends on the established video representation. To learn a better video representation for facilitating subsequent video editing, we resort to readily accessible motion priors. Specifically, we mitigate the deformation ambiguity by breaking it down into two components: known human motion warping and neural margin refinement. Thus, our first step is to obtain reliable motion priors.

As depicted in 2 (a), starting with a human-centric video $\{I_i\}_{i=0}^{N-1}$ consists of $N$ frames, we apply the off-the-shelf OSX [36] to predict the camera parameters $P_i$ and SMPL-X [47] coefficients due to its robustness toward partial observations and high efficiency. SMPL-X [47] is defined as a differentiable function $S(\beta, \theta, \psi) \longrightarrow (V, F)$ that outputs a 3D human body mesh with 10475 vertices $V \in \mathbb{R}^{10475 \times 3}$, 20908 faces $F \in \mathbb{R}^{20908 \times 3}$, where $\psi \in \mathbb{R}^{10}$ is the facial expression

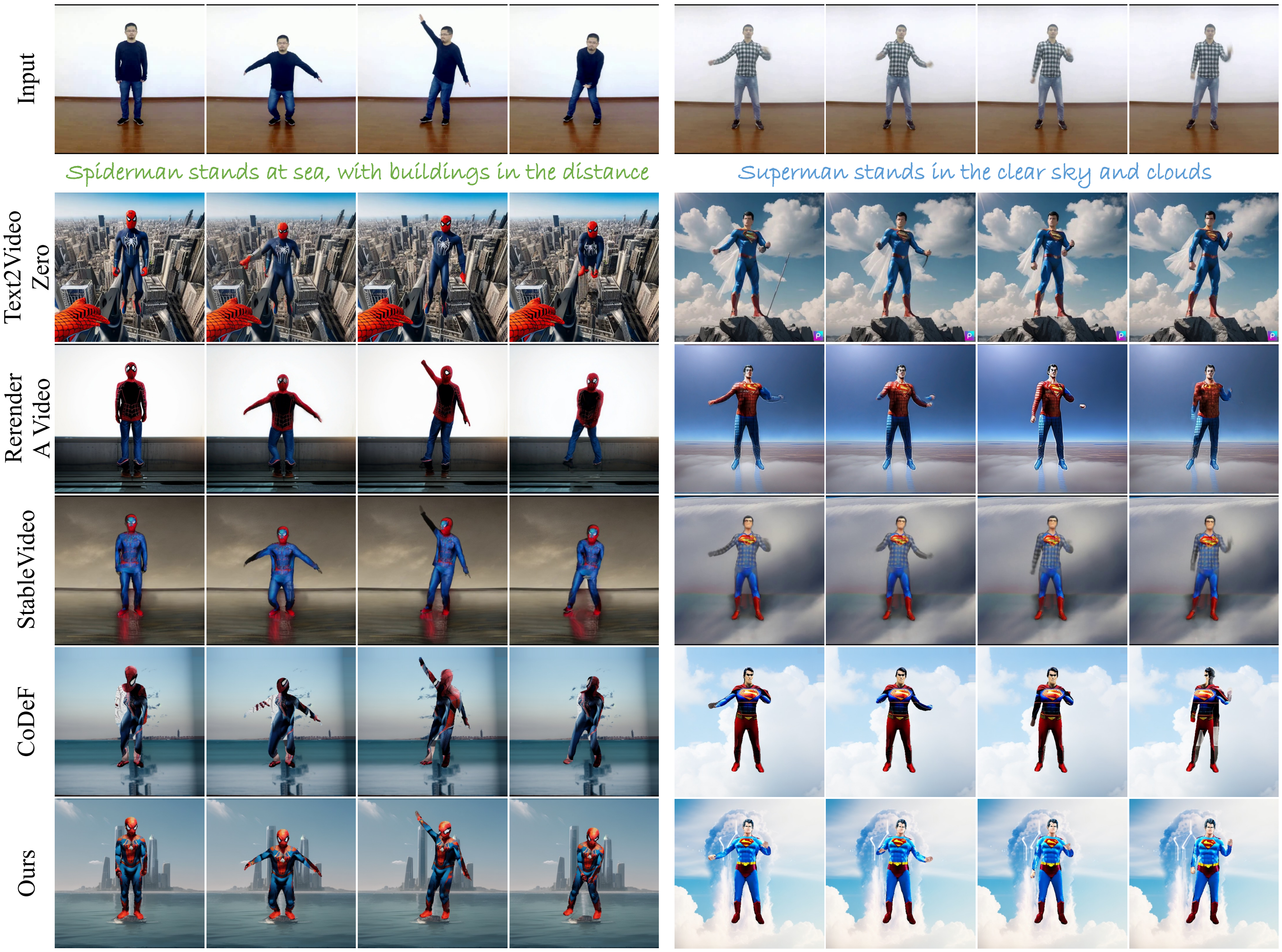

Input
Spiderman stands at sea, with buildings in the distance
Superman stands in the clear sky and clouds
Text2Video Zero
Rerender A Video
StableVideo
CoDeF
Ours

**Figure 3: Qualitative analysis of video editing by the prompt. We compare our method against baselines regarding video editing by a prompt. The first row is the input video, and the colorful description is the corresponding editing prompt. The results indicate that the first three methods suffer from self-occlusion and large-scale movements, thus producing temporal inconsistent results. CoDeF's canonical images differ from the natural ones, leading to results lacking semantics. The results of our method, in the last row, are temporally coherent and plausible.**

parameters, $\beta \in \mathbb{R}^{10}$ and $\theta \in \mathbb{R}^{22\times3}$ are the body shape parameters and pose parameters, respectively. In order to enhance the accuracy of the transformation correspondence, instead of directly utilizing the regression-based estimation provided by OSX [36], we adopt a two-step optimization strategy to obtain a more accurate SMPL-X [47] fit.

Firstly, we refine the SMPL-X coefficients with 2d keypoints. Specifically, we leverage mmpose [12] to attain 2D keypoints $P_{i(2\mathrm{D})}$ for each frame $i$. We optimize over the learnable parameters $\theta_{i=0}^{N-1}$ by minimizing the difference between estimated 2D keypoints $P_{i(2\mathrm{D})}$ and corresponding projected 2D joints $P_i(J_{\mathrm{sub}})$, where $P_i$ is the projection matrix. Additionally, we employ a temporal regularization term $\mathcal{L}^1_{\mathrm{reg}}$ on output mesh vertices $V^i_m$ to ensure continuity. The optimization objective of the first stage is:

$$\mathcal{L}^1 = \mathcal{L}_{\mathrm{kps}} + \lambda^1_{\mathrm{reg}}\mathcal{L}^1_{\mathrm{reg}} \tag{1}$$

$$\mathcal{L}_{\mathrm{kps}} = \|P_{i(2\mathrm{D})} - P_i(J_{\mathrm{sub}})\|^2_2 \tag{2}$$

$$\mathcal{L}^1_{\mathrm{reg}} = \|V_m^{[0:n-2]} - V_m^{[1:n-1]}\|^2_2 \tag{3}$$

To further improve the flexibility of the SMPL-X [47] model's expression ability, making it able to match the clothed human better in the video, rather than a skinned person. In the second stage, we added a per-vertex offset $D \in \mathbb{R}^{10475\times3}$, to capture the details of each frame and define the model as:

$$S(\beta, \theta, \psi, D) = \mathrm{LBS}(T(\beta, \theta, \psi, D), J(\beta), \theta, W) \tag{4}$$

$$T(\beta, \theta, \psi, D) = T(\beta, \theta, \psi) + D \tag{5}$$

$$T(\beta, \theta, \psi) = T + B_s(\beta) + B_p(\theta) + B_e(\psi) \tag{6}$$

where T is a mean shape template, $B_s, B_p$ and $B_e$ are shape, pose and expression blend shapes, respectively. LBS denotes linear blending skinning and $W$ is the vertices' skinning weights. We only optimize $D$ in this stage to avoid overfitting. Since we expect the rendered human outline to align with the foreground mask, we utilize mask loss as a supervision. We use SAM-Track [10] to get a per-frame binary mask $M_i$ of human and minimize the difference between the silhouette of the rendered body $R^s(P_i, \{V_i, F\})$ and the obtained mask $M_i$, where $R^s$ denotes the differentiable silhouette rasterizer. To ensure mesh smoothness, we regulate the offset with Laplacian smoothing loss [14, 45] $L_{\text{Laplacian}}$ and $L_2$ regularization. The optimization objective of the second stage is:

$$\mathcal{L}^2 = \mathcal{L}_{\text{silhouette}} + \lambda^2_{\text{reg}} \mathcal{L}^2_{\text{reg}} \tag{7}$$

$$\mathcal{L}_{\text{silhouette}} = \|R^s(P, \{V_i, F\}) - M_i\|^2_2 \tag{8}$$

$$\mathcal{L}^2_{\text{reg}} = L_{\text{Laplacian}}(D) + \gamma \|D\|^2_2 \tag{9}$$

Finally, we acquire motion priors that is sufficiently expressive for the video.

## 3.2 Video Reconstruction with Motion Priors

We find a powerful video presentation with better temporal continuity than the inter-frame attention model. With the motion priors described in Sec. 3.1, we target a more editing-friendly video representation, which could effectively convert the human-centric video editing problems into image editing problems. While overfitting the observed video using neural representation is relatively accessible, it often leads to a noisy canonical field due to the ill-posed nature of solving the deformation field. Intuitively, the prerequisite for promising editing is establishing a meaningful canonical field. Meanwhile, a well-defined deformation field can relieve the ambiguity in the canonical field, subsequently benefiting high-quality editing outcomes. The previous methods were mainly divided into two types. The first type [30] uses a UV mapping relationship between pixel space and layered neural atlas, which caused inconvenience during editing. The second type [46] wanted to compress video content onto images, but finding the correspondence between frames in large-scale human motion is challenging. Thus, we devise the canonical fields and decompose the temporal deformation in video based on the motion priors. As shown in 2, our video representation comprised of three components:

**Dual canonical fields.** We define the canonical human body as the A-posed SMPL-X+D, $S_c = S(\bar{\beta}, \theta_A, \bar{\psi}) + \bar{D}$ with mean estimated coefficients across video frames. Specifically, we adopt a dual canonical fields design in which we choose the front view $C_{\text{front}}$ and back view $C_{\text{back}}$ of the canonical human body for information complementarity. As for the network structure, our dual canonical fields are constructed using two 2D multi-resolution hash encodings, which map a 2D position $(x, y)$ to $(R, G, B)$ color.

**Human motion warping.** To alleviate the issue of overfitting resulting from directly learning a deformation field [46], we expect that explicit human motion warping dominates the overall deformation, while neural deformation serves as a refinement. Human motion warping is parametric-free, yet it provides semantic correspondences across frames in video representation.

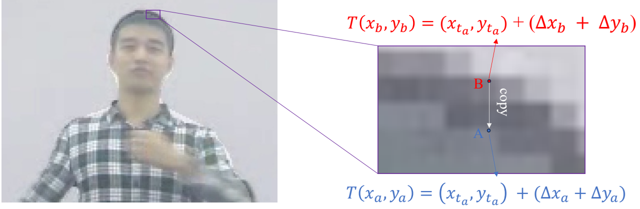

$T(x_b, y_b) = (x_{t_a}, y_{t_a}) + (\Delta x_b + \Delta y_b)$
B
copy
A
$T(x_a, y_a) = (x_{t_a}, y_{t_a}) + (\Delta x_a + \Delta y_a)$

**Figure 4: Margin refine method. Our method deals with the transformation relationships of points outside the mesh transformation matrix region and inside the human mask.**

Inspired by Liquid Warping GAN [39, 40], we build human motion warping using the Neural Mesh Renderer (NMR) [31]. To reconstruct a target frame $I_i$, we first query the canonical fields to acquire two canonical images $I_{\text{front}}$ and $I_{\text{back}}$, and then we embed them into texture space using a weak-perspective camera as $S_c$. Since our motion priors are topologically consistent, we can easily obtain the transformation $T^{\text{tr}}_{\text{front}}$ and $T^{\text{tr}}_{\text{back}}$, which warp the information from the canonical fields to the target frame. Moreover, we compute a mask to fuse the information from $C_{\text{front}}$ and $C_{\text{back}}$. For more details, please refer to the supplementary materials. It is worth noting that we define the deformation in a canonical-to-observation direction, which naturally prevents the drawback of backward deformation [9].

**Neural Margin refinement.** Now, we can make approximate transformations with human motion warping. However, be aware that human motion warping only accounts for rigid transformation, insufficient for clothes and fine-grained non-rigid deformations. In other words, the information in the position $I_i(x, y)$ is not solely determined by the transformation relationship $T^{tr}_{front|back}$. To this end, we design a small refinement field and padding strategy to refine the margin part of the human in a video.

As for the refinement field, we implement 2D multi-resolution hash encoding as the backbone. Specifically, we feed in a triplet $(x, y, i)$ into the refinement field and produce the residual $\Delta T$ : $(\Delta x, \Delta y)$. Formally, the full transformation from front or back-canonical image to each target frame $T = T^{\text{tr}}_{\text{front|back}} + \Delta T$, where $T^{\text{tr}}_{\text{front|back}}$ is estimated based on SMPL-X [47].

As shown in Fig. 4, we set the final transformation relationship to be related to the transformation relationship obtained by the human motion warping and margin refinement module.

For points within the transformation relationship $T^{tr}_{front|back}$, we set the full transformation relationship $T_{in}$ as the result of the current point's transformation relationship $T^{tr}_{front|back(in)}$ and add the result $\Delta T_{in}$ refined by the margin refinement module.

$$T_{in} = T^{tr}_{front|back(in)} + \Delta T_{in} \tag{10}$$

For points outside the transformation relationship $T^{tr}_{front|back}$ and inside the human mask $M$, we set the full transformation relationship $T_{out}$ as the transformation relationship of the point nearest to the effective transformation matrix $T_{front|back(nearest-in)}$, and add the result $\Delta T_{out}$ refined by the margin refinement module.

$$T_{out} = T^{tr}_{front|back(nearest-in)} + \Delta T_{out} \tag{11}$$

We select the two frames $I_{\text{nn\_front}}, I_{\text{nn\_back}}$ from the video closest to the front-view and back-view canonical field as regularization for training. Specifically, we additionally reconstruct the target frame $\hat{I'_i}$ with the information from $I_{\text{nn\_front}}$ or $I_{\text{nn\_back}}$ according to the orientation similarity of the pelvis. In this way, we can largely preserve the semantic information. Our representation is jointly trained by minimizing the reconstruction loss $L_{\text{rec}}$ between the predicted image $\hat{I}$ and the original one $I$ using mean square error. Moreover, we constrain the output of the deform field using $L_2$ norm empirically. Overall, the loss function of video reconstruction with motion priors can be written as:

$$\mathcal{L}^3 = \mathcal{L}_{\text{rec}}(\hat{I}_i, I_i) + \lambda_{\text{deform}} \|\Delta T\|_2^2 + \lambda^3_{\text{reg}} \mathcal{L}^3_{\text{reg}} \tag{12}$$

$$\mathcal{L}^3_{\text{reg}} = \sum_{I_i \in \text{front}} L_{\text{rec}}(\hat{I'_i}, I_i) + \sum_{I_i \in \text{back}} L_{\text{rec}}(\hat{I'_i}, I_i) \tag{13}$$

### 3.3 Video Editing Module

Upon the optimized video representation, we can obtain the trained front canonical image $I_{\text{front}}$ and back canonical image $I_{\text{back}}$ by querying the $C_{\text{front}}$ and $C_{\text{back}}$ with position $(x, y)$. Although our dual canonical fields design retains semantic and structural information, they also introduce a new challenge in terms of semantic consistency. For the randomness during the diffusing and denoising process, editing two canonical images separately using ControlNet [27] may yield inharmonious results. We suggest resolving this issue through a simple yet effective strategy to ensure editing coherence. Specifically, we concatenate the $I_{\text{front}}$ and $I_{\text{back}}$ along the width axis and feed it into ControlNet [27]. The self-attention mechanism implicitly builds the correlation between two canonical images.

As shown in Fig. 2 (c), we explore two different editing scenes: 1) **Video editing by prompt**. Users could modify the content of the human and background separately using text inputs. 2) **Video editing with user input**. Users could directly draw on the canonical images at their will. For example, they can attach a logo to their clothes and automatically propagate it throughout the video.

Moreover, HeroMaker supports editing a person individually within a multi-person video, which differs from most competing methods.

## 4 EXPERIMENTS

### 4.1 Experimental Setup

**Implementation Details.** HeroMaker is implemented in PyTorch. In the first stage, we optimize the mesh with the Adam optimizer($lr$ = 0.0001, $\beta = (0.9, 0.99)$) for 200 iterations. The regularization parameter, denoted as $\lambda^1_{reg} = 0.2$. In the second stage, We optimize for 20 iterations per frame with a learning rate of 0.0005, $\lambda^2_{reg} = 0.2$ and $\gamma = 10$. During video reconstruction, we jointly trained dual canonical fields and neural margin refinement field together with the Adam optimizer(lr=0.0001, $\beta = (\mathbf{0.9}, \mathbf{0.99})$) for 15000 iterations. We employ the pre-trained Stable Diffusion v1.5 model, and ControlNet [62] provides structure guidance regarding edges. For image editing, we implement 30 timesteps for DDIM sampling.

**Dataset.** We validate the effectiveness of our full pipeline using two datasets, including selected videos from the iPER [39, 40] and in-the-wild internet videos. These videos encompass individuals with diverse body shapes, each performing with different speeds and amplitudes. All videos consist of 50 to 200 frames, and we employ $2 \sim 4$ prompts during editing.

| Method | $E_{\text{vertices}} \downarrow$ | CLIP↑ |
|---|---|---|
| Text2Video-Zero [32] | 27.61 | 25.00 |
| Rerender-A-Video [58] | 25.85 | 26.05 |
| StableVideo[7] | 10.53 | 26.43 |
| CoDeF [46] | 26.22 | 27.48 |
| Ours | **7.81** | **27.70** |

**Table 1: Quantitative comparison on prompt-based video editing. We estimate and compute the average mesh vertices error as $E_{\text{vertices}}$ in the original and edited videos. For textual alignment, we report the average CLIP [50] score.**

| Method | Textual fidelity consistency ↑ | Shape preservation↑ | Visual effect↑ |
|---|---|---|---|
| Text2Video-Zero [32] | 0.531 | 0.500 | 0.469 |
| Rerender-A-Video [58] | 0.594 | 0.438 | 0.563 |
| StableVideo[7] | 0.375 | 0.500 | 0.469 |
| CoDeF [46] | 0.375 | 0.344 | 0.344 |
| Ours | **0.813** | **0.625** | **0.625** |

**Table 2: User study on prompt-based video editing.**

| Method | NLA [30] | CoDeF [46] | Ours |
|---|---|---|---|
| Visual effect ↑ | 0.375 | 0.365 | **0.750** |

**Table 3: User study on user interactive video editing.**

**Baselines.** We compare our method with five baselines: NLA [30], Text2Video-Zero [32], Rerender-A-Video [58], StableVideo [7] and CoDeF [46]. We compare our method with NLA [30] and CoDeF [46] in video editing with the user input task to validate the ability of the model to represent the video in a structure-aware correspondence. For video editing by prompt task, we compare our model with Text2Video-Zero [32], Rerender-A-Video [58], StableVideo [7] and CoDeF [46] to show the temporal consistency and ability to match the prompts.

**Evaluation Metrics.** Human-centric video editing aims to faithfully reflect the editing prompt while maintaining original shape coherency and temporal consistency. We further propose to measure the shape coherency. In detail, we estimate the human mesh in the original and edited video using OSX [36] and compute the average mesh vertices error as $E_{\text{vertices}}$. For textual alignment, we report the average CLIP [50] score, which computes the cosine similarity between the prompts' CLIP [50] embedding and each frame's image embedding in the edited video. However, simply using these metrics cannot fully represent the visual quality of edited videos. We thus conducted a user study. We show the textual descriptions and editing results of different methods, asking them to rate in three aspects: textual fidelity with temporal continuity, shape preservation, and comprehensive visual effect.

## 4.2 Comparison with Baselines

**Quantitative Comparison.** Following previous works, we evaluated our method and baselines with different metrics. As indicated in Table. 1, our method surpasses previous works in all metrics, demonstrating that our editing results align closely with the prompts and maintain the original body shape. We further conduct user studies as described in Sec.4.1. As shown in Table. 2 and Table. 3, the participants exhibit a clear preference for our results.

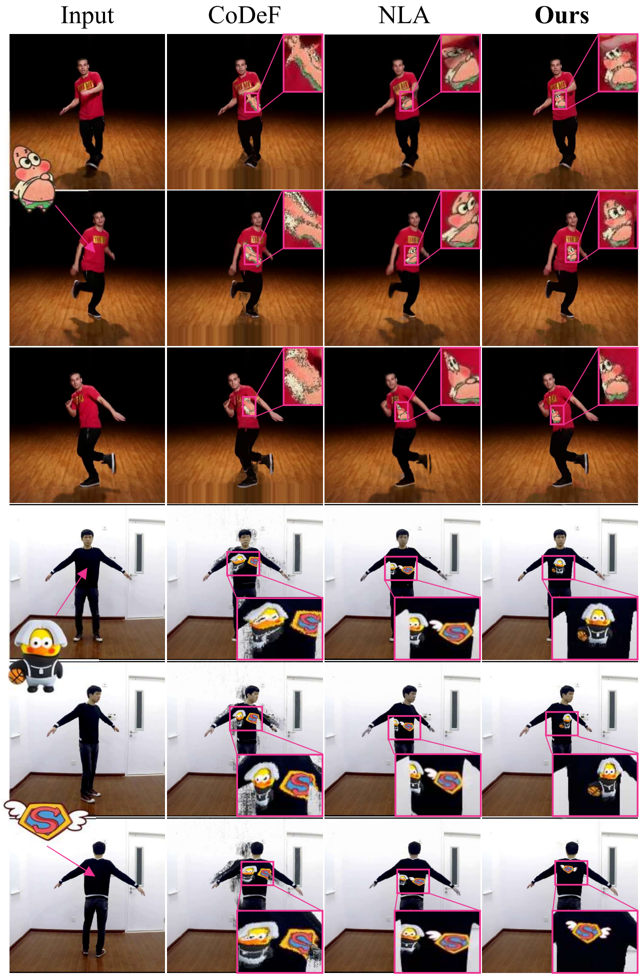


**Figure 5: Qualitative analysis of video editing with user input. Our method supports local editing and allows users to add customized icons accurately into the region of interest. We compare our method against NLA [30] and CoDeF [46].**

**Qualitative Comparison.** Fig.3 presents the visual results of prompt-based video editing. Text2Video-Zero [32] and Rerender-A-Video [58] generate outputs semantically aligned with the text description but fail to maintain temporal consistency. For instance, the body shapes are flickering, and the arms are distorted (see Spider-Man and Superman in Fig. 3). StableVideo [7] exhibits satisfactory temporal consistency. However, it is prone to generating outputs with reduced fidelity. Since CoDeF learns the deformation field and the canonical image without structure information, it generates different results from natural images when handling human-centric videos. According to Fig. 3, by leveraging the motion priors, our method successfully achieves temporal consistency while preserving fidelity.

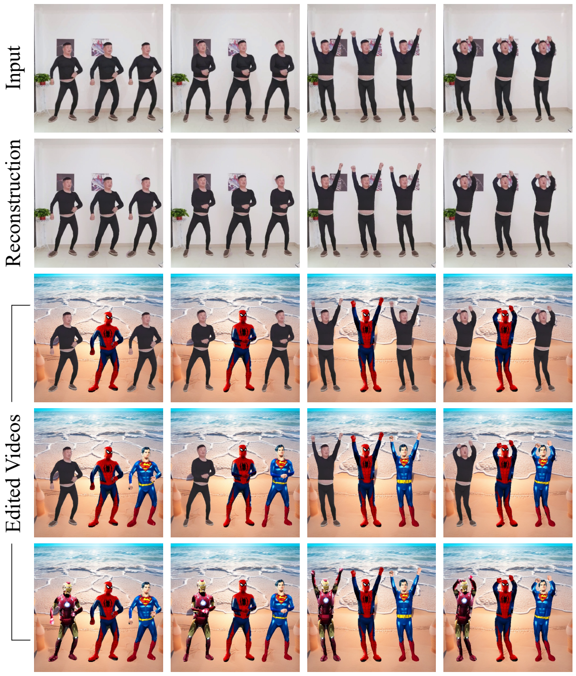


**Figure 6: Multiple people editing results. Our method can extend to the reconstruction and editing, where the input video sequence supports multiple people. In the subsequent stage, users can edit one or more people in the scene individually, providing greater flexibility.**

Furthermore, representing the video with motion priors allows our model to edit the human body locally. It enables users to edit regions of interest while maintaining the other parts. In Fig. 5, we compare our method to NLA [30] and CoDeF [46] in user interactive video editing. NLA [30] and CoDeF [46] models use optical flow to maintain the correspondence between frames. However, estimating optical flow for complex human motion is difficult, which leads to visual flaws. Although NLA [30] shows good textured results, it fails to maintain geometry consistency between human motion. CoDeF [46] leads to information losses when encoding the video into a canonical content field. In some cases, the correspondence of body deformation deviates, causing unpleasant editing results. In contrast, our model utilizes motion priors and thus learns human-aware canonical fields, ensuring that the editing contents are attached to the appropriate positions.

Additionally, unlike most previous methods, HeroMaker offers the ability to easily modify any character within a video containing multiple people, as illustrated in Fig. 6. This capability enhances the appeal and flexibility of human-centric video editing, providing users with a unique and engaging experience.

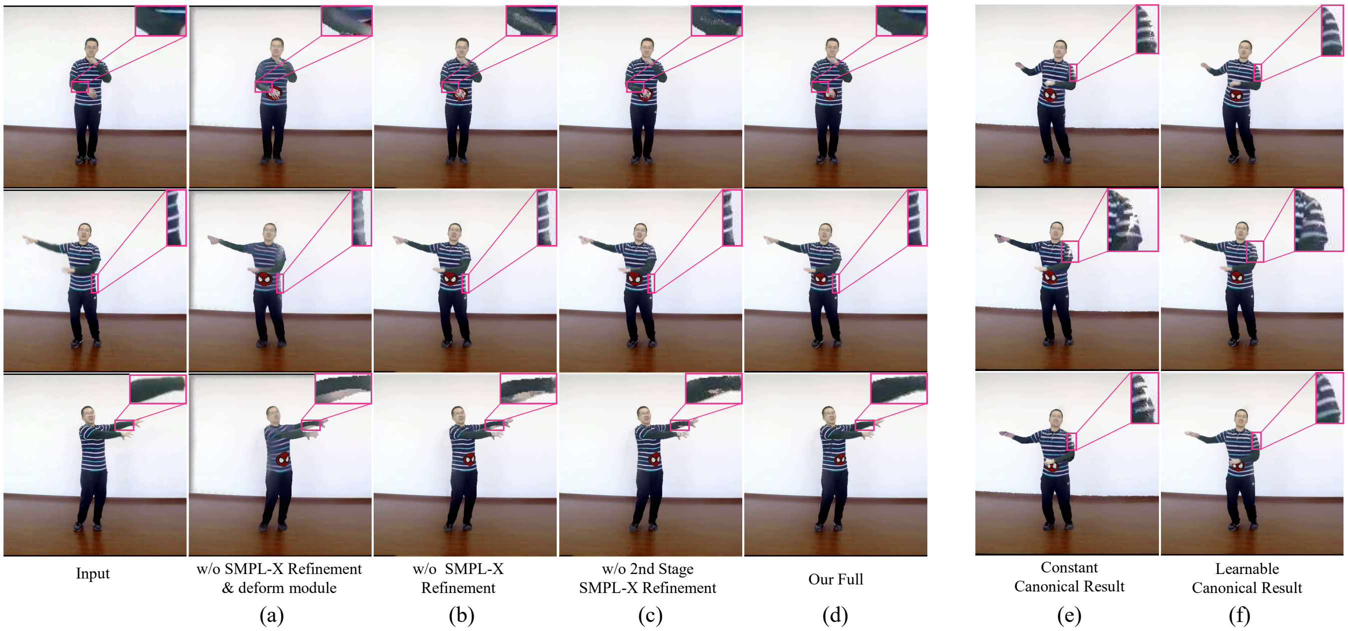

Input
w/o SMPL-X Refinement & deform module
(a)
w/o SMPL-X Refinement
(b)
w/o 2nd Stage SMPL-X Refinement
(c)
Our Full
(d)
Constant Canonical Result
(e)
Learnable Canonical Result
(f)

**Figure 7: Qualitative ablation results. Compared to method (a), method (b) benefits from a neural deformation module to correct estimation errors from the SMPL-X network. The method (c) further improves the results using first-stage SMPL-X refinement that effectively improves motion priors. Our full uses a neural deformation module and two stages of SMPL-X refinement to achieve precise mesh deformation. Additionally, we demonstrate the necessity of learnable canonical fields by comparing method (e) versus method (f).**

### 4.3 Ablation Studies

To verify the contributions of different modules to the overall performance, we systematically deactivate specific modules in our framework and present the visual comparison in Fig.7. In this section, we mainly analyzed the impact of varying degrees of SMPL-X refinement, whether to add a deform module and the necessity of learning canonical fields. We define method (a) as the model without SMPL-X refinement and the neural deformation module. Method (b) incorporates the neural deformation module into the baseline. Considering that our framework performs SMPL-X refinement in two stages, method (c) applies only the first refinement stage, whereas method (d) represents our full model.

**Neural deformation module.** The neural deformation module aims to correct estimation errors from the SMPL-X network by ensuring the alignment of frame images with the canonical fields under motion priors. The comparison between methods (a) and (b) demonstrates improvements in clarity and reduction in edge deviations, highlighting the module's efficacy in enhancing visual quality by learning accurate correspondences.

**SMPL-X refinement.** The refinement process mitigates mesh deviations detected by the SMPL-X network and utilizes 2D keypoints and inter-frame mesh deviations for optimization in the first stage. It is visible that the improvement in image quality in method (c) indicates that the correctness of motion priors has a positive impact on the results. The second refinement stage further rectifies edge artifacts, underscoring the refinement's critical role in achieving precise mesh deformation.

**Learnable canonical fields.** Additionally, to demonstrate the necessity of learnable canonical fields, we select the two images closest to the front and back view as canonical images and then optimize the deform network to obtain the final results. We believe that this can extract as much information as possible from the video while ensuring semantic information. However, as shown in Fig. 7, by comparing method (e) versus method (f), we observe that learnable canonical fields capture more detailed and relevant information from video sequences, thereby reducing reconstruction errors and improving edge definition. Constant canonical images, despite their simplicity, fail to accommodate the complexity and randomness of motion, leading to artifacts in reconstructed images.

## 5 CONCLUSION

In this paper, we present HeroMaker, an innovative human-aware framework that prioritizes human-centric video editing. Our approach utilizes motion priors based on human body mesh to establish the transformation correspondence from human-aware canonical fields to each video frame. Powered by our devised video representation, we maintain meaningful and structural canonical fields that facilitate the subsequent synthesis of temporal coherent and plausible results in response to diverse user interactions. Extensive experiments and visual results demonstrate the superior performance of our HeroMaker, while ablation studies confirm the effectiveness of our design.

## REFERENCES

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video Based Reconstruction of 3D People Models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. 2022. Blended Latent Diffusion. *arXiv preprint arXiv:2206.02779* (2022). arXiv:2206.02779.

[3] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. 2023. SINE: Semantic-driven Image-based NeRF Editing with Prior-guided Editing Field. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*. Springer, 707–723.

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. arXiv:2304.08818 [cs.CV]

[6] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. 2023. Pix2Video: Video Editing using Image Diffusion. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[7] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. 2023. StableVideo: Text-driven Consistency-aware Diffusion Video Editing. *arXiv preprint arXiv:2308.09592* (2023).

[8] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. 2023. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. arXiv:2305.13840 [cs.CV]

[9] Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger. 2021. SNARF: Differentiable Forward Skinning for Animating Non-Rigid Neural Implicit Shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11594–11604.

[10] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and Track Anything. *arXiv preprint arXiv:2305.06558* (2023).

[11] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. 2023. FLATTEN: Optical Flow-guided ATTENtion for Consistent Text-to-Video Editing. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[12] MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

[13] Paul Couairon, Clément Rambour, Jean-Emmanuel Haugeard, and Nicolas Thome. 2023. VidEdit: Zero-Shot and Spatially Aware Text-Driven Video Editing. arXiv:2306.08707 [cs.CV]

[14] Mathieu Desbrun, Mark Meyer, Peter Schröder, and Alan H. Barr. 1999. Implicit Fairing of Irregular Meshes using Diffusion and Curvature Flow. In *SIGGRAPH*.

[15] Zhongjie Duan, Lizhou You, Chengyu Wang, Cen Chen, Ziheng Wu, Weining Qian, and Jun Huang. 2023. DiffSynth: Latent In-Iteration Deflickering for Realistic Video Synthesis. arXiv:2308.03463 [cs.CV]

[16] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and Content-Guided Video Synthesis with Diffusion Models. arXiv:2302.03011 [cs.CV]

[17] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. 2023. CCEdit: Creative and Controllable Video Editing via Diffusion Models. arXiv:2309.16496 [cs.CV]

[18] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. 2022. Capturing and Animation of Body and Clothing from Monocular Video. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Article 45, 9 pages.

[19] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. *arXiv preprint arxiv:2307.10373* (2023).

[20] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 2023. Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations* (2024).

[22] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, and Qifeng Chen. 2023. Animate-A-Story: Storytelling with Retrieval-Augmented Video Generation. arXiv:2307.06940 [cs.CV]

[23] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626* (2022).

[24] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303 [cs.CV]

[25] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *arXiv:2204.03458* (2022).

[26] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *arXiv preprint arXiv:2311.17117* (2023).

[27] Zhihao Hu and Dong Xu. 2023. VideoControlNet: A Motion-Guided Video-to-Video Translation Framework by Using Diffusion Model with ControlNet. arXiv:2307.14073 [cs.CV]

[28] Jiahui Huang, Leonid Sigal, Kwang Moo Yi, Oliver Wang, and Joon-Young Lee. 2023. INVE: Interactive Neural Video Editing. arXiv:2307.07663 [cs.CV]

[29] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. 2022. SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. 2021. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–12.

[31] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Neural 3D Mesh Renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. *arXiv preprint arXiv:2303.13439* (2023).

[33] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. 2023. Shape-aware Text-driven Layered Video Editing. arXiv:2301.13173 [cs.CV]

[34] Zhenyi Liao and Zhijie Deng. 2023. LOVECon: Text-driven Training-Free Long Video Editing with ControlNet. arXiv:2310.09711 [cs.CV]

[35] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. 2023. MagicEdit: High-Fidelity and Temporally Coherent Video Editing. In *arXiv*.

[36] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. 2023. One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21159–21168.

[37] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. 2023. DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Human-Centric Video Editing. *arXiv preprint arXiv:2310.10624* (2023).

[38] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2023. Video-P2P: Video Editing with Cross-attention Control.

[39] Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019. Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*.

[40] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. 2021. Liquid warping GAN with attention: A unified framework for human image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[41] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. arXiv:2303.08320 [cs.CV]

[42] Aryan Mikaeili, Or Perel, Mehdi Safaee, Daniel Cohen-Or, and Ali Mahdavi-Amiri. 2023. SKED: Sketch-guided Text-based 3D Editing. *ICCV* (2023).

[43] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

[44] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. 2023. Dreamix: Video Diffusion Models are General Video Editors. arXiv:2302.01329 [cs.CV]

[45] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. 2006. Laplacian Mesh Optimization. In *GRAPHITE*.

[46] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. 2023. CoDeF: Content Deformation Fields for Temporally Consistent Video Processing. *arXiv preprint arXiv:2308.07926* (2023).

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[48] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. *arXiv:2303.09535* (2023).

[49] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. 2023. InstructVid2Vid: Controllable Video Editing with Natural Language Instructions. arXiv:2305.12328 [cs.CV]

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020* (2021).

[51] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. 2024. Control4D: Efficient 4D Portrait Editing with Text. (2024).

[52] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. arXiv:2209.14792 [cs.CV]

[53] Wen Wang, kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. 2023. Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models. *arXiv preprint arXiv:2303.17599* (2023).

[54] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16210–16220.

[55] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.

[56] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. 2023. SimDA: Simple Diffusion Adapter for Efficient Video Generation. *arXiv preprint arXiv:2308.09710* (2023).

[57] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. [n. d.]. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model.

[58] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*.

[59] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Gong Ming, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. 2023. NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation. arXiv:2303.12346 [cs.CV]

[60] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. 2023. Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation. *arXiv preprint arXiv:2309.15818* (2023).

[61] Jianfeng Zhang, Hanshu Yan, Zhongcong Xu, Jiashi Feng, and Jun Hao Liew. 2023. MagicAvatar: Multi-modal Avatar Generation and Animation. In *arXiv*.

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. [n. d.]. Adding Conditional Control to Text-to-Image Diffusion Models.

[63] Shangzhan Zhang, Sida Peng, Yinji ShenTu, Qing Shuai, Tianrun Chen, Kaicheng Yu, Hujun Bao, and Xiaowei Zhou. 2023. Dyn-E: Local Appearance Editing of Dynamic Neural Radiance Fields. arXiv:2307.12909 [cs.CV]

[64] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. 2023. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077* (2023).

[65] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2023. MagicVideo: Efficient Video Generation With Latent Diffusion Models. arXiv:2211.11018 [cs.CV]

[66] Qianshu Zhu, Chu Han, Guoqiang Han, Tien-Tsin Wong, and Shengfeng He. 2020. Video snapshot: Single image motion expansion via invertible motion embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (2020), 4491–4504.