Implicit Motion Function

Yue Gao Jiahao Li Lei Chu Yan Lu
Microsoft Research
{yuegao, li.jiahao, leichu, yanlu}@microsoft.com

Abstract

Recent advancements in video modeling extensively rely on optical flow to represent the relationships across frames, but this approach often lacks efficiency and fails to model the probability of the intrinsic motion of objects. In addition, conventional encoder-decoder frameworks in video processing focus on modeling the correlation in the encoder, leading to limited generative capabilities and redundant intermediate representations. To address these challenges, this paper proposes a novel Implicit Motion Function (IMF) method. Our approach utilizes a low-dimensional latent token as the implicit representation, along with the use of cross-attention, to implicitly model the correlation between frames. This enables the implicit modeling of temporal correlations and understanding of object motions. Our method not only improves sparsity and efficiency in representation but also explores the generative capabilities of the decoder by integrating correlation modeling within it. The IMF framework facilitates video editing and other generative tasks by allowing the direct manipulation of latent tokens. We validate the effectiveness of IMF through extensive experiments on multiple video tasks, demonstrating superior performance in terms of reconstructed video quality, compression efficiency and generation ability.

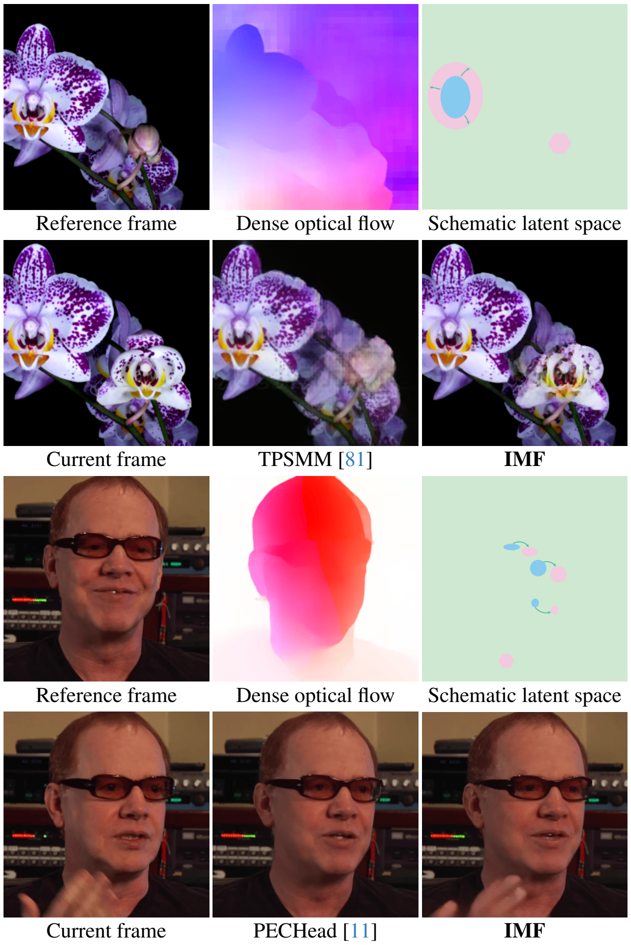

Figure 1. IMF represents the motion in latent space and model the correlation implicitly. In the diagram, the blue shapes represent tokens from the reference frame, while the pink shapes represent tokens from the current frame. Tokens that are unmatched in the reference frame, indicating new content, are depicted with pink polygons. Successfully matched tokens, such as changes of the head pose or the magnification of flower, are indicated with green arrows. Compared to explicit methods PECHead [11] and TPSMM [81], our IMF faithfully reconstructs the current frame.

1. Introduction

Recent years have witnessed an upsurge in the field of video modeling [44], processing [45, 62], compression [31, 38], prediction [19, 20], and generation [11, 64, 66]. A fundamental element in these diverse tasks is the use of optical flow, which encapsulates the relationship between consecutive frames. Existing approaches often employ readily available optical flow estimation models, such as RAFT [61], to facilitate this process. However, the definition of optical flow has inherent limitations. It lacks sparsity, capturing only the positional change of pixels, and fails to model the probabilistic aspects of object movements, thereby not comprehending the intrinsic motion and semantics of objects. Furthermore, the non-editability of explicit optical flow presents significant challenges in video generation tasks. This stems from its rigid structure, which lim-its the ability to manipulate or alter motion trajectories directly within the flow field. Consequently, this necessitates the integration of sophisticated neural networks for downstream generation tasks. These identified limitations underscore the need for an innovative approach in video modeling. It is in this context that we propose the Implicit Motion Function (IMF) method, a novel paradigm designed to address the aforementioned shortcomings by offering a more nuanced and editable representation of motion within video sequences.

Our design of the IMF is based on the observation that, contrasting with the aforementioned limitations brought by the explicit optical flow, modeling the frame relations in the latent space offers a more nuanced and adaptable approach. Specifically, by operating in this latent space, it becomes feasible to effectively distinguish different objects and new content between two frames. Such capability is derived from the inherent property of the latent space by encoding a richer and more abstract representation of motion and object characteristics. This encoding allows for a more generalized understanding of object dynamics and interactions, thereby enabling the IMF to adapt to and accurately represent a broader range of object movements and behaviors, including the newly appeared objects. It is worth noting that we generalize the concept of motion to include the entry and exit of objects from the frame as a form of motion, rather than solely encompassing pixel displacement.

The design of our pipeline, adopting an encoder-decoder style, can be highlighted in several aspects: (1) We employ a *dense feature encoder* and a *latent token encoder* to model dense features and latent tokens independently. This approach allows us to represent motion between two frames as a pair of *low-dimensional tokens*, significantly increasing the sparsity of the representation. (2) We design a *latent token decoder* to decode the tokens and a *implicit motion alignment* to model the correlation between two frames aligning the features from the reference frame to the current frame. They enable the model to understand the relationship between features across frames, thereby grasping the motion patterns of objects and increasing the sparsity of the motion representation. (3) By directly editing the latent token, we can facilitate editing and other generative tasks. In particular, the encoder independently extracts latent tokens from the current frame and the reference frame, so the frame-to-frame relationship is entirely modeled by the decoder. As shown in Fig. 1, we provide a schematic illustration of the correlation modeling process within the IMF latent space. Compared to explicit methods, our IMF allows for a faithful motion modeling of non-facial content (*e.g.*, flowers) and even new content (*e.g.*, appearing hand and blooming flower), demonstrating the semantic completeness is ensured by the proposed implicit modeling, for both talking head and general videos.

In implementing the IMF, two primary challenges emerge. The first concerns defining and extracting sparse tokens. Conventional approaches [11, 41, 57, 66, 81] rely on sparse keypoints transformed into dense optical flow, but this method proves suboptimal, particularly for data lacking strong priors. Our approach circumvents this limitation by allowing the encoder to extract latent tokens without tying them to physical coordinates, preserving semantic integrity while maintaining sparsity. The second challenge involves modeling correlations with latent tokens. Drawing parallels to language model advancements [43, 48], we leverage cross-attention [65] mechanism, a sophisticated weight-based matrix operation that exceeds the capabilities of optical flow-based grid sampling [41]. By utilizing a latent token decoder to generate motion feature maps from compact latent tokens and processing these features through cross-attention, we model frame relationships and align reference appearance features to the current frame. This approach enables refined motion feature generation and feature alignment, bolstering both the reconstructive and generative abilities of the decoder and offering significant potential when integrated with foundation models for video applications.

We conducted comprehensive experiments to validate the superiority of our IMF model in video reconstruction, video compression, and talking head editing. In video reconstruction, IMF outperformed existing benchmarks in both talking head and general datasets, demonstrating its ability to accurately reconstruct non-facial objects and new content. For video compression, IMF achieved a superior rate-distortion trade-off compared to current methods, including neural video codecs, highlighting the compactness of our latent tokens. In talking head editing, IMF surpassed state-of-the-art (SOTA) explicit editing methods, confirming the editability and robust generative capabilities of our latent tokens.

Our contributions can be summarized as follows:

- We propose the Implicit Motion Function (**IMF**), a novel framework for video modeling and video generation. The IMF is capable of faithfully reconstructing various videos at extremely low bit rates, while also enabling high-quality video generation.
- We incorporate the latent token decoder and the implicit motion alignment as the implicit motion function for correlation modeling. By integrating this step at the decoder, the generative capability of the decoder is unleashed.
- By manipulating the latent tokens, the proposed method can achieve high-quality editing results for talking head images, which also demonstrates the powerful expressive capability of our latent tokens.
- Extensive qualitative and quantitative results across several talking head and general datasets demonstrate the superiority of the our framework for video modeling and video generation.

2. Related Works

2.1. Image Animation and Video Generation

Image animation, which involves transferring motion between images, has traditionally relied on methods utilizing face meshes [13, 49], human keypoints [78], or action units [22, 47]. Recent trends, however, are leaning towards self-supervised techniques that require only video input. Notable pioneering works such as FOMM [57], and MRAA [58] introduced motion representation through sparse keypoints or regions employing local affine transformations. TPSMM [81] employs thin-plate spline transformation for motion estimation. LIA [67] uses latent vector decomposition but relies on optical flow for warping, and its limited editability leads to artifacts and facial interference. Furthermore, Mallya et al. [41] proposed a warping method based on cross-attention mechanisms [65]. However, its reliance on explicit keypoints as an intermediate representation imposes limitations on performance, as its effectiveness in modeling non-facial data remains relatively limited.

In contrast to methods targeting general data, the field of talking head video generation, which focuses on creating realistic videos of talking faces, has advanced significantly due to research in various areas. Unlike the image animation methods that do not need detailed facial models, talking head generation uses detailed information about the head to improve video quality. Progress in this field has been driven by large datasets of face images [8, 72, 73, 82], 3D models of faces [2, 10], face landmark detectors [6, 39], facial landmark detection [6, 39], neural radiance fields (NeRF) [14, 18, 28, 32, 42, 53, 75], and diffusion models [4, 15, 46, 51, 52, 59, 70, 79]. There are several approaches to generate talking head videos. (1). 3D Face Model-Based Methods [40, 63]: These methods use 3D models to track and recreate facial expressions. For example, Face2Face [63] tracks expressions from a source video and applies them to another face. (2). Direct Synthesis-Based Models [7, 76, 77]: These models create faces by transforming encoded representations of appearance and motion into images. They often use latent motion expressions but may also rely on explicit techniques like optical flow. (3). 3D Mesh-Based Methods [12, 26]: These methods use detailed 3D models to create lifelike head avatars from videos. (4). NeRF-Based Methods [14, 18, 18, 32, 53]: These utilize NeRF to represent head geometry and appearance in a novel 3D format. While effective, they can be complex and less versatile across different identities. (5). Warping-Based Methods [17, 50, 66, 69, 73]: These methods estimate motion fields to warp feature maps and generate images, often using learned keypoints. (6). Diffusion Model-Based Methods [9, 54, 60]: These methods create realistic videos from a single image and an audio sequence, enhancing lip-sync, video fidelity, and the naturalness of head movements and poses. However, they involve complex training and inference pipelines, which can be a challenge in terms of computational resources and efficiency. Despite the successes of these methods, there are still opportunities for improvement, especially in handling non-facial data and enhancing editability for talking head videos. This paper aims to explore new approaches in video modeling that maintain high quality and efficiency while being adaptable to different models.

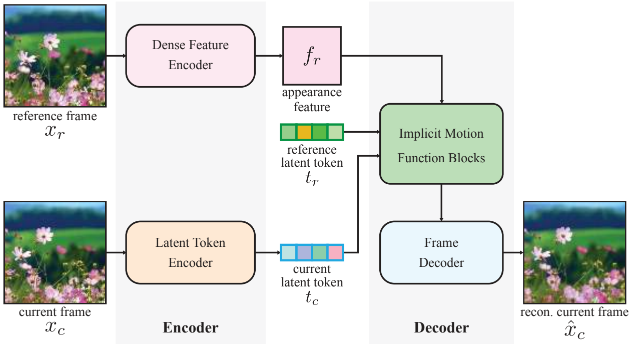

Figure 2. The overview of our method, which is encoder-decoder style. The reference latent token t_r is obtained with the Latent Token Encoder. The f_r^l and t_r are shared across the whole video.

2.2. Neural Video Codec

Many neural video codecs [3, 34, 36, 38, 74] adopt the widely-used residual coding-based framework, where the predicted frame is firstly generated via motion estimation and motion compensation, and then the residual is coded. They aim to improve the prediction accuracy as much as possible, via scale-space flow [3], block-level prediction[36], and so on. Instead of generating pixel-domain predicted frame, the emerging conditional coding-based codecs [27, 29–31, 37, 55] extract feature-domain temporal context to help encoding, decoding, and the entropy model. However, most of them still rely on optical flow to generate the prediction or temporal context. Coding the optical flow requires non-trivial bitrate cost, even with the help of the learned entropy model [29–31]. By contrast, our IMF learns sparse latent token, which is naturally compact and efficient.

3. Proposed Method

3.1. Overview

An overview of our framework is shown in Fig. 2, which follows the encoder-decoder style. The encoder contains the *dense feature encoder* E_F and the *latent token encoder* E_T. The decoder is composed of several *implicit motion function* (IMF) blocks and the *frame decoder* D_F. Given input video $\{x_c\}_{c\in\Omega}$ and a reference frame x_r, the appearance of the reference frame can be encoded into f_r by E_F, and the correlation between x_c an x_r can be represented by

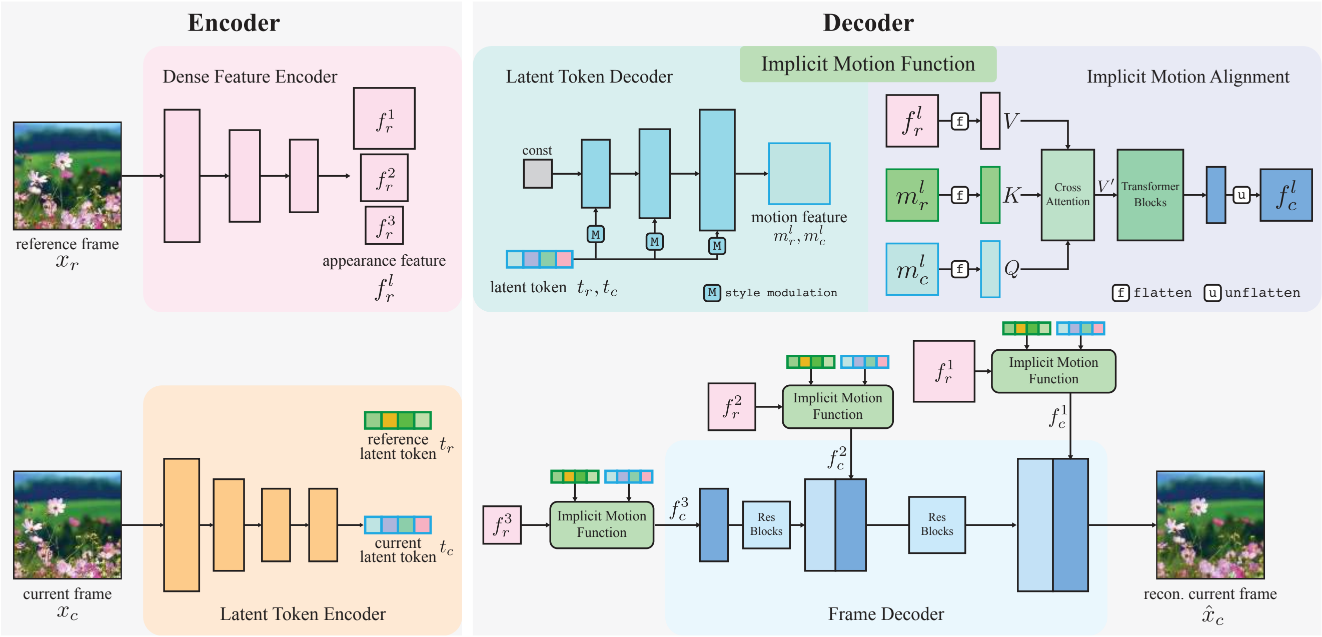

Figure 3. The illustration of our framework: a) the encoder contains the dense feature encoder E_F and the latent token encoder E_T; b) the decoder contains the implicit motion function IMF and the feature decoder D_F. IMF includes the latent token decoder IMF_D and the implicit motion alignment IMF_A. The motion features m_r^l and m_c^l are produced by IMF_D from t_r and t_c, independently.

the tokens pair (t_c, t_r) extracted respectively by E_T from x_c and x_r. The decoder D_F aims to faithfully reconstruct the $\hat{x}_c$ with (t_c, t_r, f_r). The pipeline can be formulated as:

$$\begin{aligned} &\hat{x}_c = D_F(IMF(t_c, t_r, f_r)) \quad \text{with} \\ &t_c = E_T(x_c), \quad t_r = E_T(x_r), \quad f_r = E_F(x_r). \end{aligned} \tag{1}$$

Note that the reference frame x_r is not necessarily selected from the input sequence $\{x_c\}$. Taking the talking head generation as an example, our pipeline can support the reference having a different ID with the input. Furthermore, different from the former works such as optical flow, the correlation modeling by (t_c, t_r) does not require the reference input and the current input has the same dimension, not even the same modality.

In our framework, to learn sparse yet essential information, we will constrain the information flow in t_c and t_r. The optimization target can be formulated as:

$$\begin{aligned} \min \quad & L(\hat{x}_c, x_c) \\ \text{s.t.} \quad & |t| < \epsilon, \ t \in \{t_c, t_r\}, \end{aligned} \tag{2}$$

where L is the function that measures the similarity between $\hat{x}_c$ and x_c. $|t|$ is the size of the latent tokens, and ϵ is the size limitation. To achieve this objective, we specially design the IMF at decoder to obtain high-fidelity $\hat{x}_c$. A detailed illustration of our full pipeline is shown in Fig. 3. The remaining part of this section will be organized as follows: In Sec. 3.2, we briefly introduce the encoder details of E_F and E_T. In Sec. 3.3, we describe the proposed IMF. Sec. 3.4 demonstrates the editability of the latent tokens on talking head generation. In Sec. 3.5, we delve into the design and underlying rationale of our proposed framework.

3.2. Dense Feature & Latent Token Encoder

Dense Feature Encoder. Given a reference frame $x_r \in \mathbb{R}^{H \times W \times 3}$, the dense feature encoder E_F extracts multi-scale feature $f_r^l \in \mathbb{R}^{h_l \times w_l \times d_l}, l \in \{1, L\}$, where l is the layer index and L is number of layers. h_l, w_l is spatial size of the feature. d_l is the depth dimension of the feature. The E_F is only performed on the reference frame, which means the features f_r^l are shared across the whole video if the reference frame is constant.

Latent Token Encoder. The latent token encoder E_T encodes the reference frame and the current frame to t_r and t_c, independently. It maps the space $\mathbb{R}^{H \times W \times 3} \Rightarrow \mathbb{R}^{\mathbf{d}}$. Similar to f_r, when the reference frame is constant, t_r is also shared across the entire video. When $\mathbf{d} = K \times 2$, where K is the number of keypoints, our tokens can be viewed as the explicit coordinates representations from FOMM [57]. However, due to the limited information expressed by coordinates, the expressive capability of this kind of representation is limited, especially in terms of semantic integrity, *i.e.* the completeness to faithfully reconstruct the original frame. By contrast, we choose $\mathbf{d} = 1 \times d_m$, a latent vector to ensure both completeness and compactness. We further discuss in Sec. 3.5 the choice between explicit and implicit representations of t_c and reveal the pros and cons of each in the ablation experiments.

3.3. Implicit Motion Function

With the latent token t_r and the appearance feature f_r extracted from the reference frame, the target of IMF is to obtain the appearance feature f_c of the current frame from its latent token t_c. Specifically, the IMF is mainly composed of two parts. First, we design a *latent token decoder* IMF_D module to transform the highly compact latent tokens into spatially aligned motion features that correspond with the frame. Subsequently, we utilize *implicit motion alignment* IMF_A module to align and refine the appearance feature of the reference frame to the current frame.

Latent Token Decoder. The IMF_D decodes the compact tokens t_r, t_c to multiple motion features $m_r^l \in \mathbb{R}^{h_l \times w_l \times d_l}$ and $m_c^l \in \mathbb{R}^{h_l \times w_l \times d_l}$ where l is the layer index. The "style modulation" utilizes Weight Modulation and Demodulation technique of StyleGAN2 [25], scales the convolution weights with the latent token, and normalizes them to unit standard deviation. Owing to the varying granularities, our latent tokens effectively compress multi-scale information, serving as a comprehensive representation. Furthermore, the fully implicit nature of our representation allows for flexible adjustment of the latent token dimension to accommodate different scenarios. Different from keypoint-based explicit representation [41, 57, 66], where the motion features are Gaussian heatmaps converted from the keypoints, our design enjoys better scalability and capability due to the latent token being directly learned by the encoder instead of coordinates with a limited value range.

Implicit Motion Alignment. As shown in Fig.3, with the motion features m_r^l and m_c^l, the IMF_A will align the reference appearance features f_r^l to the current frame. The IMF_A contains two parts including the cross-attention [65] and transformer. The cross-attention module is implemented with scaled dot-product cross-attention. This attention module takes motion features m_c^l, m_r^l, and appearance features f_r^l as Q, K, and V, respectively. The features are first flattened, then the positional embeddings P_q, P_k are added to the queries and keys, We compute the dot products of the queries with keys, divide each by $\sqrt{d_k}$, and apply a `softmax` function to obtain the weights on the values. Then the output-aligned values are computed through matrix multiplication. With the aligned values V', we further refine them using multi-head self-attention and feed-forward network-based Transformer blocks [65], and finally obtain the appearance features f_c^l of the current frame.

3.4. Token Manipulation

In contrast to the explicit optical flow method, the implicit representation offers a distinct advantage in terms of editability. As the latent token is not task-specific, for a new controllable generation task, the decoder can be fixed and just train a small adapter with a small cost. Taking talking head generation as an example, in the training of the latent space, the full network is only trained to reconstruct the current frame. After the latent space is trained, we can edit learned tokens through another independent token manipulation network. Formally, with an editing module ψ, source frame x_s, and control condition h, we can rewrite Eq. 1 to obtain the edited feature map $\tilde{f}_s$ as:

$$\tilde{f}_s = \mathit{IMF}\big(\psi(t_s|h), t_s, f_s\big). \tag{3}$$

Passing $\tilde{f}_s$ to the frame decoder trained before, we can get the edited frame $\tilde{x}_s$. In our experiments, we implement ψ with two MLP encoders to encode the source token and control condition (*e.g.*, the 3DMM face coefficients), and one MLP decoder to output the edited token.

3.5. Discussion

Explicit or Implicit. Based on the foundational works MonkeyNet [56], FOMM [57], and MRAA [58], introduced by Siarohin *et al.*, several works employ explicit keypoints [17, 66], landmarks [11, 81], regions [58] or vectors [67] as intermediate representations for direct optical flow estimation. Given the pivotal role of optical flow in elucidating relationships between video frames, this explicit approach has become the predominant methodology in video modeling. However, explicit representations are inherently constrained. Optical flow, being a dense representation, poses optimization challenges due to the per-pixel operations in grid sampling. While keypoints offer sparsity, they compromise on detailed motion information. Additionally, keypoints are effective for quasi-rigid structures like human heads but falter in general video tasks due to the diverse shapes and motion patterns, which are difficult to encapsulate using explicit methods. Another notable drawback of keypoints is their limited scalability, where increased point usage for finer motion details often leads to redundancy and artifact-prone outputs.

Encoder-centric (EC) or Decoder-centric (DC). Here, we discuss the placement of correlation modeling, whether it is within the encoder or the decoder. Most existing video compression and video modeling methods belong to EC, which requires a complex design for encoder. For example, the functioning of a video codec encoder is contingent on the decoder output to obtain the prediction or context. Such intertwined framework necessitates meticulous tuning. For instance, the previous SOTA codec DCVC-DC [31, 55] incorporates upwards of ten distinct complicate strategies during training phases. In contrast, recent advancements in Large Language Models (LLMs) [48] usually utilize a DC architecture to support diverse tasks. This paradigm shift prompts reconsideration of similar frameworks for video codec. Actually, Wyner-Ziv coding [71] has given the theoretical support that the rate of DC framework can be identical with that of the EC framework. Our IMF design aligns with successful DC-based LLMs, achieving significant enhancements in video modeling.

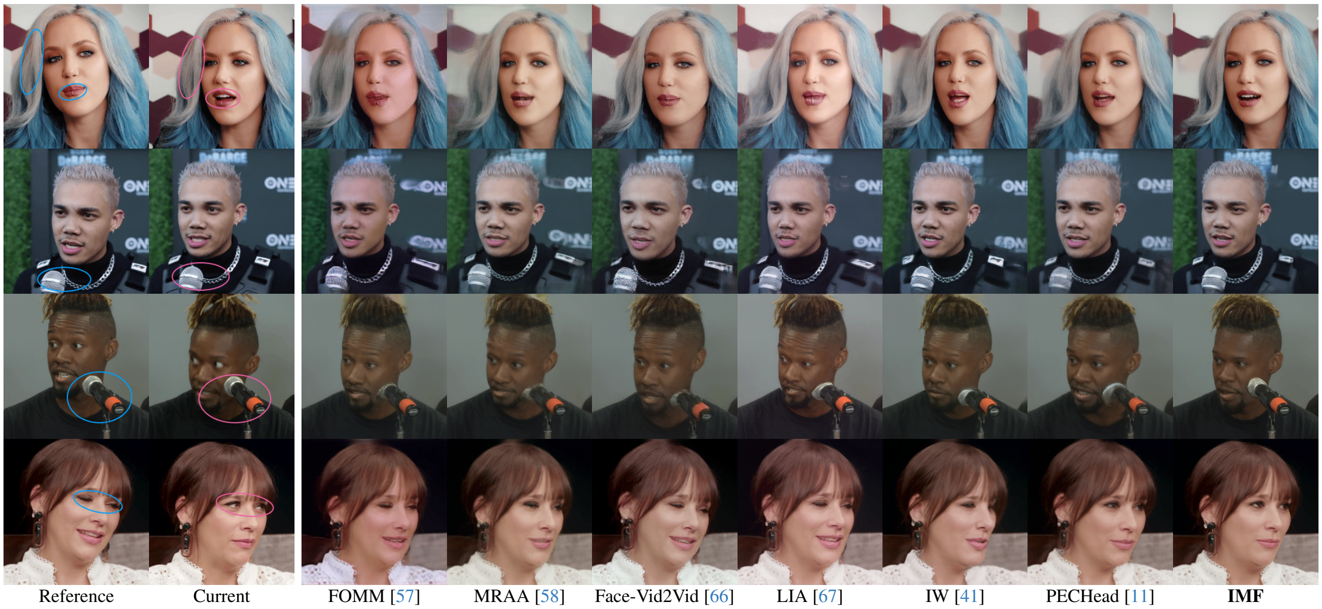

Figure 4. Comparison of talking head video reconstruction results obtained by the proposed method and previous SOTA approaches. The main differences between reference frames and current frames are highlighted in circles.

Table 1. Quantitative results of video reconstruction on talking head datasets.

Methods	CelebV-HQ [82]					VFHQ [72]				
	L_1	SSIM	PSNR	FID	LPIPS	L_1	SSIM	PSNR	FID	LPIPS
MRAA [58]	0.0568	0.777	22.33	64.23	0.1539	0.0454	0.812	22.60	40.17	0.1117
F-Vid2Vid [66]	0.0589	0.746	21.56	67.40	0.1889	0.0491	0.804	21.79	41.95	0.1054
LIA [67]	0.0654	0.754	20.75	65.15	0.1541	0.0537	0.815	21.47	42.27	0.1099
IW [41]	0.0539	0.801	22.67	57.65	0.1287	0.0426	0.862	23.46	34.05	0.0823
PECHead [11]	0.0552	0.803	24.29	56.68	0.1372	0.0435	0.859	23.03	31.20	0.0847
IMF	**0.0417**	**0.849**	**25.28**	**53.22**	**0.1057**	**0.0283**	**0.918**	**26.86**	**23.35**	**0.0540**

Table 2. Quantitative results of video reconstruction on general datasets.

Methods	Flower			Wavecloth			Foliage		
	L_1	SSIM	LPIPS	L_1	SSIM	LPIPS	L_1	SSIM	LPIPS
FOMM [57]	0.0837	0.620	0.2559	0.0749	0.684	0.2009	0.0801	0.6881	0.1761
MRAA [58]	0.0792	0.632	0.2460	0.0665	0.710	0.1798	0.0757	0.7180	0.1925
TPSMM [81]	0.0642	0.704	0.2097	0.0625	0.716	0.1537	0.0626	0.7771	0.1545
LIA [67]	0.0848	0.589	0.2453	0.0733	0.682	0.1868	0.0826	0.6619	0.1679
IW [41]	0.0639	0.736	0.1983	0.0566	0.790	0.1380	0.0697	0.7427	0.1479
IMF	**0.0544**	**0.761**	**0.1746**	**0.0473**	**0.804**	**0.1265**	**0.0548**	**0.8235**	**0.1209**

4. Experiments

Datasets. For talking head datasets, CelebVHQ [82] and VFHQ [72] are used. In addition, to verify the universality of our model, we also test three general datasets: Flower, Wavecloth and Foliage[1]. For visual results of the non-face datasets, we obtain in-the-wild video from the Internet to verify the generalization ability.

Implementation details. We use 256 × 256 as the frame spatial size. The latent token dimension is $d_m = 32$ for all experiments except for ablation studies. During the training, the perceptual Loss [23] and GAN loss [21, 33] are also used. More training details and network structure of each module are in supplementary materials.

Baselines. We compare four representative works: FOMM [57], MRAA [58], Face-Vid2Vid [66], and PECHead [11] on talking head data. Besides FOMM [57] and MRAA [58], we compare TPSMM [81], LIA [67] and IW [41] on general data. These methods were primarily proposed for face data, but can also be applied for general data because the models do not explicitly rely on prior knowledge of face.

Metrics. We use L_1, SSIM [68] and PSNR to evaluate the low-level similarity. FID [16] and LPIPS [80] are used to assess the perceptual quality. For video compression, BPP (bits per pixel) is used to measure the bit-stream size.

4.1. Video Reconstruction

Tab. 1 and Tab. 2 show the quantitative comparisons with previous SOTA models. Qualitative comparisons are presented in the Fig. 4 and Fig. 5. More results and videos are in supplemental material. These results show that our method significantly outperforms previous SOTA methods. For talking head data, although previous methods can approximately reconstruct the correct facial area, only our model accurately reconstructs the non-facial areas, such as hair and microphones. Our model also performs best for local motion changes, like blinking. For general data, our method is able to precisely reconstruct various contents with complex motion, such as blooming. However, the methods based on explicit representation and alignment [56, 57, 81] perform poorly in cases of drastic content changes and also have inferior modeling capabilities for subtle motions, such as minor positional changes in leaves and flowers. Implicit warping-based method IW [41] is better than previous methods, but the keypoints are still used. It

[1] For more information about these datasets, please contact us.

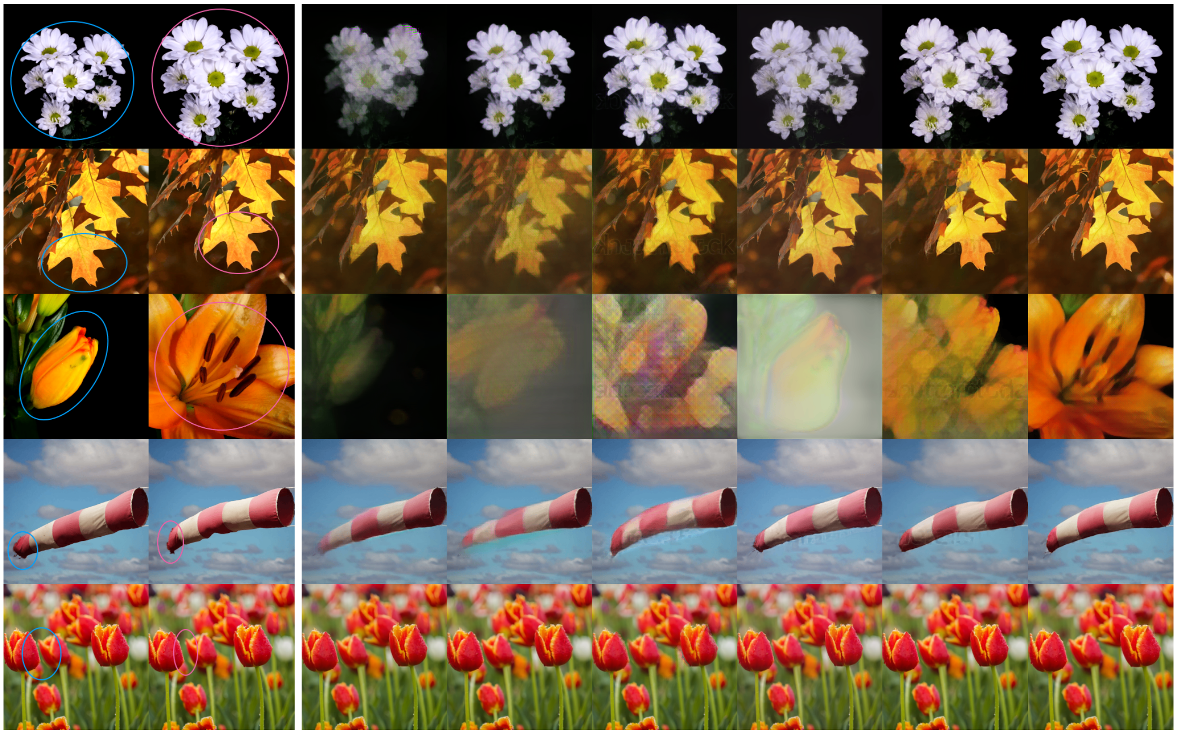
Reference Current FOMM [57] MRAA [58] TPSMM [81] LIA [67] IW [41] **IMF**

Figure 5. Comparison of general video reconstruction results obtained by the proposed method and previous SOTA approaches. The main differences between reference frames and current frames are highlighted in circles.

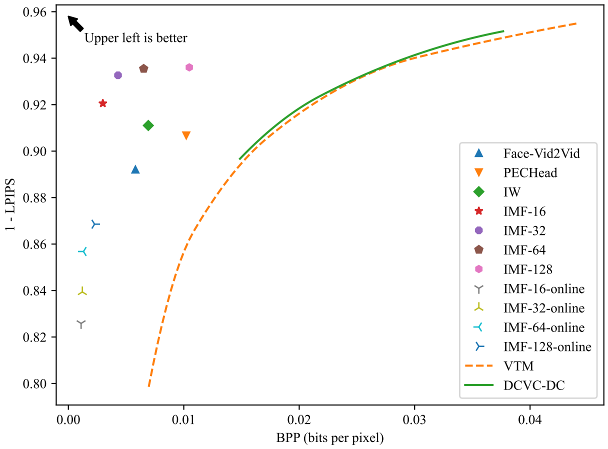

Figure 6. The rate-distortion curve on video compression.

lacks robustness, leading to the appearance of jitter artifacts in the videos.

4.2. Video Compression

To verify the efficiency and semantic completeness of our latent tokens, we also test the video compression task using VFHQ [72] dataset. The rate-distortion curves are presented in Fig. 6. Besides the existing methods of talking head generation, we also compare the best standard codec H.266/VTM [1, 5] and previous SOTA neural video codec DCVC-DC [31]. For the results of all methods, we only calculate the current BPP (bits per pixel). Our first approach is directly applying an offline array compressor [35] to compress the latent tokens, and we test different token sizes $d_m \in \{16, 32, 64, 128\}$, denoted as IMF-$d_m$. To verify the compactness of the latent token, we also follow neural codec and implement an online learned entropy model [29, 31], where the token of the previous frame is used to predict the distribution of that of the current frame and then the arithmetic coding is applied. It is denoted as IMF-d_m-online. From these results, we can see that our method is significantly better than all previous methods in terms of rate-distortion trade-off. Furthermore, the comparison with the online entropy model also reveal that our latent tokens are naturally compact, because the online entropy model only bring a very small bitrate reduction but with non-trivial quality degradation.

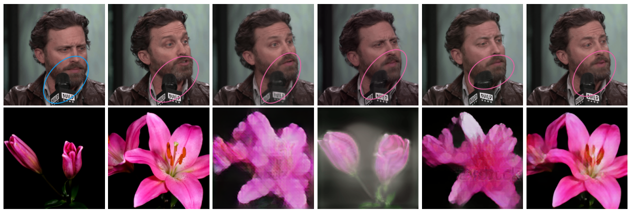
Reference Current TPS [81] LIA [67] IW [41] **IMF**

Figure 7. Comparison with explicit methods on both talking head and general datasets.

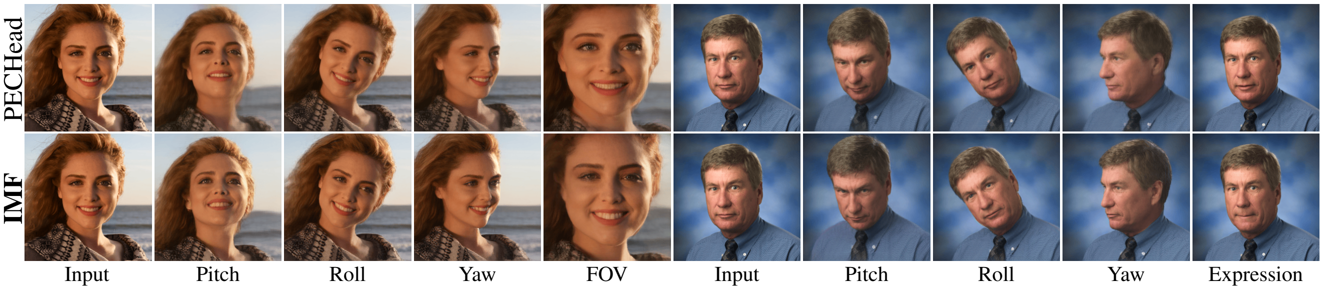

Figure 8. Comparison of talking head editing results obtained by the proposed method and the previous SOTA PECHead [11].

Table 3. Quantitative results for ablation studies. IR is implicit motion representation and IA is implicit motion alignment.

Settings		Flower				VFHQ [72]			
IR	IA	L_1	SSIM	PSNR	LPIPS	L_1	SSIM	FID	LPIPS
		0.0642	0.704	22.23	0.2097	0.0435	0.859	31.20	0.0847
✓		0.0604	0.693	21.58	0.1921	0.0396	0.877	29.84	0.0612
	✓	0.0578	0.758	22.46	0.1762	0.0366	0.885	30.08	0.0713
✓	✓	**0.0544**	**0.761**	**23.33**	**0.1746**	**0.0283**	**0.918**	**23.35**	**0.0540**

4.3. Comparison with Explicit Methods

As shown in Fig. 7, we compare previous SOTA explicit methods on both talking head and general datasets. The optical flow-based TPS [81] cannot handle the correlation of facial and non-facial content, or new content, as the microphone and flower are not correctly modeled. The latent motion decomposition-based LIA [67] containing explicit flow in the model, fails when the corresponding regions are small or the motion is complex. The keypoints-based IW [41] cannot handle the local motion and can also cause jittering in the videos.

4.4. Token Editing

In our framework, token manipulation facilitates the editing of facial features, enabling precise control over aspects such as head pose and facial expression. To demonstrate the generalization capability of our method, we applied token editing to images from the FFHQ dataset [24]. We benchmark our approach against previous SOTA explicit model for talking head editing, PECHead [11]. Qualitative results, as depicted in Fig. 8, demonstrate that our method not only parallels the editing performance of PECHead [11] but also surpasses it in terms of realism and naturalness, especially in intricate areas such as the eyes, mouth, and ears.

4.5. Ablation Studies

We compare the implicit motion representation and alignment with the explicit motion representation (*i.e.*, keypoints) and alignment (*i.e.*, optical flow-based grid sample). For explicit motion representation and alignment method on general dataset Flower, we use the TPSMM [81] as the baseline model. For the talking head dataset VFHQ [72], we use the PECHead [11] as the baseline model. For implicit motion representation and explicit motion alignment, we keep the latent token encoder and decoder, and replace the implicit motion alignment with an optical flow estimator and flow-based grid sample to align the motion feature. For explicit motion representation and implicit motion alignment, we replace the latent token encoder with a keypoints extractor and the latent motion decoder with explicit keypoints to Gaussian heatmap [57], while keeping the implicit motion alignment. The results are shown in the Tab. 3 and Fig. 9. The results indicate that explicit representation can only model the face for facial data and is incapable of modeling hands: it either fails to generate results with hands, or it does not correctly reconstruct the position of the hands. The results on flower data also demonstrate the shortcomings of explicit representation and alignment: they either fail to capture complex motions or are unable to accurately reconstruct the frames.

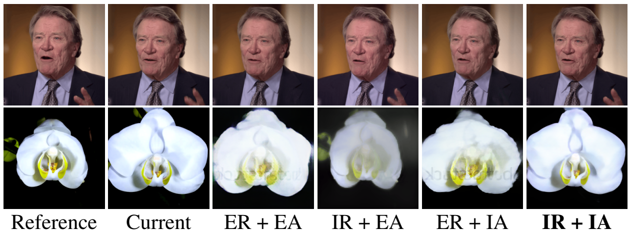

Figure 9. Ablation studies on both talking head and general datasets. The IR and ER stand for implicit and explicit motion representation. The IA and EA stand for implicit and explicit motion alignment. The IR + IA is our full model.

5. Conclusion

This paper introduces the Implicit Motion Function (IMF), an innovative approach that advances beyond traditional optical flow in video modeling and generation. The implicit methodology of IMF not only achieves high-fidelity video reconstruction, especially in non-facial and new content regions, but also demonstrates a remarkable compression ratio improvement. Its capacity for directly editable latent tokens paves the way for superior editing performance. We believe IMF represents a meaningful step forward in video modeling, potentially inspiring further research and development in video reconstruction and generation.

References

[1] VTM-17.0. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/. 7

[2] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Audio-visual face reenactment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5178–5187, 2023. 3

[3] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8503–8512, 2020. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[5] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 7

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 3

[7] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 3

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 3

[9] Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4281–4289, 2023. 3

[10] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 3

[11] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5609–5619, 2023. 1, 2, 5, 6, 8

[12] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 3

[13] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time. *arXiv preprint arXiv:2006.10962*, 2020. 3

[14] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 3

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[17] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 3, 5

[18] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 3

[19] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2023. 1

[20] Zhihao Hu, Guo Lu, Jinyang Guo, Shan Liu, Wei Jiang, and Dong Xu. Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5921–5930, 2022. 1

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[22] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 3

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 8

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 5

[26] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 345–362. Springer, 2022. 3

[27] Théo Ladune, Pierrick Philippe, Wassim Hamidouche, Lu Zhang, and Olivier Déforges. Conditional coding for flexible learned video compression. In *Neural Compression: From Information Theory to Applications – Workshop @ ICLR*, 2021. 3

[28] Seoyoung Lee, Seongsu Ha, and Joonseok Lee. Disentangled audio-driven nerf: Talking head generation with detailed identity-specific microexpressions. 3

[29] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 7

[30] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[31] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, Canada, June 18-22, 2023*, 2023. 1, 3, 5, 7

[32] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978, 2023. 3

[33] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 6

[34] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-LVC: multiple frames prediction for learned video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[35] Peter Lindstrom. Fixed-Rate Compressed Floating-Point Arrays, 2014. 7

[36] Bowen Liu, Yu Chen, Rakesh Chowdary Machineni, Shiyu Liu, and Hun-Seok Kim. Mmvc: Learned multi-mode video compression with block-based prediction mode selection and density-adaptive entropy coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18487–18496, 2023. 3

[37] Jerry Liu, Shenlong Wang, Wei-Chiu Ma, Meet Shah, Rui Hu, Pranaab Dhawan, and Raquel Urtasun. Conditional entropy coding for efficient video compression. In *European Conference on Computer Vision*, pages 453–468. Springer, 2020. 3

[38] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. DVC: an end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 1, 3

[39] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017. 3

[40] Luming Ma and Zhigang Deng. Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 1–10, 2019. 3

[41] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022. 2, 3, 5, 6, 7, 8

[42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[43] OpenAI. Chatgpt. https://chat.openai.com, 2023. 2

[44] Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2806–2826, 2020. 1

[45] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 1

[46] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[47] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 3

[48] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2, 5

[49] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018. 3

[50] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven

generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[53] Shuai Shen, Wanhua Li, Xiaoke Huang, Zheng Zhu, Jie Zhou, and Jiwen Lu. Sd-nerf: Towards lifelike talking head animation via spatially-adaptive dual-driven nerfs. *IEEE Transactions on Multimedia*, 2023. 3

[54] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 3

[55] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 2022. 3, 5

[56] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 5, 6

[57] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 4, 5, 6, 7, 8

[58] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 3, 5, 6, 7

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[60] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023. 3

[61] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 1

[62] A Murat Tekalp. *Digital video processing*. Prentice Hall Press, 2015. 1

[63] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[64] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 1

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3, 5

[66] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1, 2, 3, 5, 6

[67] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 3, 5, 6, 7, 8

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[69] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 3

[70] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023. 3

[71] Aaron Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on information Theory*, 22(1):1–10, 1976. 5

[72] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 3, 6, 7, 8

[73] Kewei Yang, Kang Chen, Yuan-Chen Guo, Daoliang Guo, Song-Hai Zhang, and Weidong Zhang. Face2face$^{\rho}$: Real-time high-resolution one-shot face reenactment. In *European conference on computer vision*. Springer, 2022. 3

[74] Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte. Learning for video compression with recurrent autoencoder and recurrent probability model. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):388–401, 2021. 3

[75] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 3

[76] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 3

[77] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 3

[78] Jing Zhang, Zhe Chen, and Dacheng Tao. Towards high performance human keypoint detection. *International Journal of Computer Vision*, 129(9):2639–2662, 2021. 3

[79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[81] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 1, 2, 3, 5, 6, 7, 8

[82] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. *arXiv preprint arXiv:2207.12393*, 2022. 3, 6