Velocity Disambiguation for Video Frame Interpolation

Zhihang Zhong, Yiming Zhang, Wei Wang, Xiao Sun, Yu Qiao, Gurunandan Krishnan, Sizhuo Ma†, and Jian Wang†

***Abstract*—Existing video frame interpolation (VFI) methods blindly predict where each object is at a specific timestep t ("time indexing"), which struggles to predict precise object movements. Given two images of a baseball, there are infinitely many possible trajectories: accelerating or decelerating, straight or curved. This often results in blurry frames as the method averages out these possibilities. Instead of forcing the network to learn this complicated time-to-location mapping implicitly together with predicting the frames, we provide the network with an explicit hint on how far the object has traveled between start and end frames, a novel approach termed "distance indexing". This method offers a clearer learning goal for models, reducing the uncertainty tied to object speeds. We further observed that, even with this extra guidance, objects can still be blurry especially when they are equally far from both input frames (*i.e.*, halfway in-between), due to the directional ambiguity in long-range motion. To solve this, we propose an iterative reference-based estimation strategy that breaks down a long-range prediction into several short-range steps. When integrating our plug-and-play strategies into state-of-the-art learning-based models, they exhibit markedly sharper outputs and superior perceptual quality in arbitrary time interpolations, using a uniform distance indexing map in the same format as time indexing without requiring extra computation. Furthermore, we demonstrate that if additional latency is acceptable, a continuous map estimator can be employed to compute a pixel-wise dense distance indexing using multiple nearby frames. Combined with efficient multi-frame refinement, this extension can further disambiguate complex motion, thus enhancing performance both qualitatively and quantitatively. Additionally, the ability to manually specify distance indexing allows for independent temporal manipulation of each object, providing a novel tool for video editing tasks such as re-timing. The code is available at https://zzh-tech.github.io/InterpAny-Clearer/.**

***Index Terms*—Video frame interpolation, Temporal super-resolution, Disambiguation, Video editing**

I. Introduction

VIDEO frame interpolation (VFI) plays a crucial role in creating slow-motion videos [1], video generation [2], prediction [3], and compression [4]. Directly warping the starting and ending frames using the optical flow between them can only model linear motion, which often diverges from actual motion paths, leading to artifacts such as holes. To solve this, learning-based methods have emerged as leading solutions to VFI, which aim to develop a model, represented as $\mathcal{F}$, that uses a starting frame I_0 and an ending frame I_1 to generate a frame for a given timestep, described by:

$$I_t = \mathcal{F}\left(I_0, I_1, t\right). \tag{1}$$

Two paradigms have been proposed: In fixed-time interpolation [1], [5], the model only takes the two frames as input and always tries to predict the frame at $t = 0.5$. In arbitrary-time interpolation [6], [7], the model is further given a user-specified timestep $t \in [0, 1]$, which is more flexible at predicting multiple frames in-between.

Yet, in both cases, the unsampled blank between the two frames, such as the motion between a ball's starting and ending points, presents infinite possibilities. The velocities of individual objects within these frames remain undefined, introducing a *velocity ambiguity*, a myriad of plausible time-to-location mappings during training. Incorporating additional neighboring frames as input [8] can partially restrict the solution space by imposing constraints on motion trajectories, but it cannot fully resolve the ambiguity. We observed that velocity ambiguity is a primary obstacle hindering the advancement of learning-based VFI: Models trained using aforementioned *time indexing* receive identical inputs with differing supervision signals during training. As a result, they tend to produce blurred and imprecise interpolations, as they average out the potential outcomes.

Could an alternative indexing method minimize such conflicts? One straightforward option is to provide the optical flow at the target timestep as an explicit hint on object motion. However, this information is unknown at inference time, which has to be approximated by the optical flow between I_0 and I_1, scaled by the timestep. This requires running optical flow estimation on top of VFI, which may increase the computational complexity and enforce the VFI algorithm to rely on the explicitly computed but approximate flow. Instead, we propose a more flexible *distance indexing* approach. In lieu of an optical flow map, we employ a *distance ratio* map D_t, where each pixel denotes *how far the object has traveled between start and end frames*, within a normalized range of $[0, 1]$,

$$\boxed{I_t = \mathcal{F}\left(I_0, I_1, \text{motion hint}\right) \;\Rightarrow\; I_t = \mathcal{F}\left(I_0, I_1, D_t\right).} \tag{2}$$

During training, D_t is derived from optical flow ratios computed from ground-truth frames. During inference, it is sufficient to provide a uniform map as input, in the exactly same way as time indexing methods, *i.e.*, $D_t(x, y) = t, \forall x, y$. However, the semantics of this indexing map have

† denotes corresponding authors.

Z. Zhong is with the School of Artificial Intelligence, Shanghai Jiao Tong University. Y. Zhang is with Cornell University. W. Wang, X. Sun and Y. Qiao. are with Shanghai Artificial Intelligence Laboratory. G. Krishnan is with OtoNexus Medical Technologies. S. Ma and J. Wang. are with Snap Inc.

Part of the work was done while Z. Zhong and G. Krishnan were at Snap.

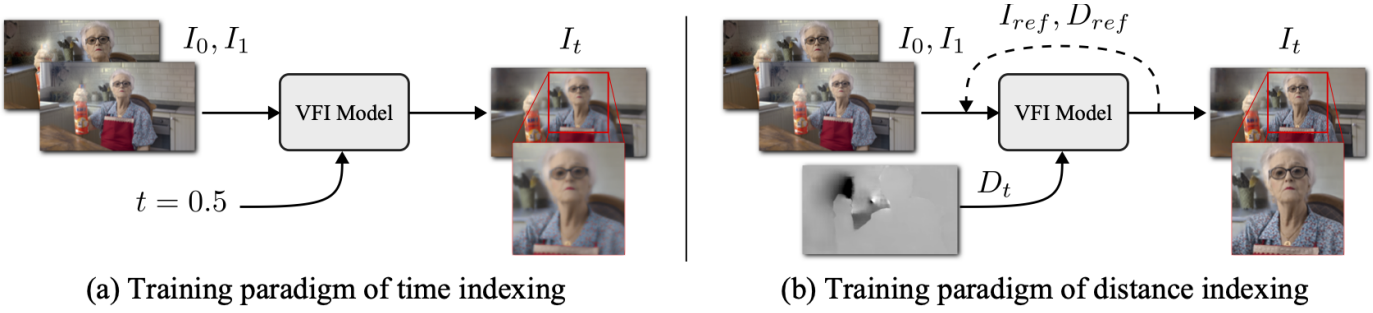


Fig. 1. Comparison of time indexing and distance indexing training paradigms. (a) Time indexing uses the starting frame I_0, ending frame I_1, and a scalar variable t as inputs. (b) Distance indexing replaces the scalar with a distance map D_t and optionally incorporates iterative reference-based estimation (I_{ref}, D_{ref}) to address velocity ambiguity, resulting in a notably sharper prediction.

shifted from an uncertain timestep map to a more deterministic motion hint. Through distance indexing, we effectively solve the one-to-many time-to-position mapping dilemma, fostering enhanced convergence and interpolation quality.

Although distance indexing addresses the scalar *speed ambiguity*, the *directional ambiguity* of motion remains a challenge. Empirically we found that, although learning-based video frame interpolation can handle minor directional uncertainty at a small timestep, the ambiguity becomes pronounced at the temporal midway between the two input frames, *i.e.*, t=0.5, as illustrated in Fig. 3(b). Inspired by iterative inference paradigms in optical flow [9] and image generation [10], we introduce an iterative reference-based estimation strategy. Rather than estimating the full motion field at once, our approach decomposes the problem into incremental distance steps. By propagating estimates from nearby to farther points, we constrain the search space at each iteration, thereby minimizing directional uncertainty and enhancing synthesis quality.

Our approach addresses challenges that are not bound to specific network architectures. Indeed, it can be applied as a plug-and-play strategy that requires only modifying the input channels for each model, as demonstrated in Fig. 1. We conducted extensive experiments on four existing VFI methods to validate the effectiveness of our approach, which produces frames of markedly improved perceptual quality. Moreover, instead of using a uniform map, it is also possible to use a spatially-varying 2D map as input to manipulate the motion of objects. Paired with state-of-the-art segmentation models such as Segment Anything Model (SAM) [11], this empowers users to freely control the interpolation of any object, *e.g.*, making certain objects backtrack in time.

When using more than two input frames [8], nearby frames offer additional constraints that facilitate the computation of a pixel-wise dense distance map. The methods discussed so far employ a uniform distance map because a deterministic distance map cannot be derived from only two frames. Inspired by continuous parametric optical flow estimation [12], we utilize cubic B-splines and neural ordinary differential equations to estimate a dense distance map from four input frames, which improves our model's performance across both perceptual and pixel-centric metrics. Furthermore, we make a trainable copy of the original interpolator architecture to refine the initial two-frame interpolation results using information from two additional frames I_{-1} and I_2. This multi-frame refiner module is intended to fully harness the potential of additional frames, thereby enhancing multi-frame interpolation quality.

In summary, our key contributions are: 1) Proposing distance indexing and iterative reference-based estimation to address the velocity ambiguity and enhance the capabilities of arbitrary time interpolation models; 2) Presenting an unprecedented manipulation method that allows for customized interpolation of any object. 3) Adopting a continuous distance map estimator and proposing a multi-frame fusion architecture to enhance interpolation quality across both perceptual and pixel-centric metrics.

A preliminary version of this work was presented in [13], where we focus on using a uniform distance map during inference because an accurate, pixel-wise distance map cannot be reliably calculated from two frames. Although using a uniform distance map produced plausible results with better perceptual quality, the predictions did not align with the ground truth, resulting in lower performance on metrics such as PSNR and SSIM. In this paper, we address this limitation by using multiple frames (more than two) as input and introduce a continuous distance map estimator that approximates the map from nearby frames. We also present a simple yet effective multi-frame refiner architecture for video frame interpolation. Extensive experiments demonstrate that this approach significantly enhances performance.

II. Related Work

A. Video frame interpolation

1) General overview: Numerous VFI solutions rely on optical flows to predict latent frames. Typically, these methods warp input frames forward or backward using flow calculated by off-the-shelf networks like [9], [14]–[16] or self-contained flow estimators like [7], [17], [18]. Networks then refine the warped frame to improve visual quality. SuperSlomo [6] uses a linear combination of bi-directional flows for intermediate flow estimation and backward warping. DAIN [1] introduces a depth-aware flow projection layer for advanced intermediate flow estimation. AdaCoF [19] estimates kernel weights and offset vectors for each target pixel, while BMBC [20] and ABME [21] refine optical flow estimation. Large motion interpolation is addressed by XVFI [22] through a recursive multi-

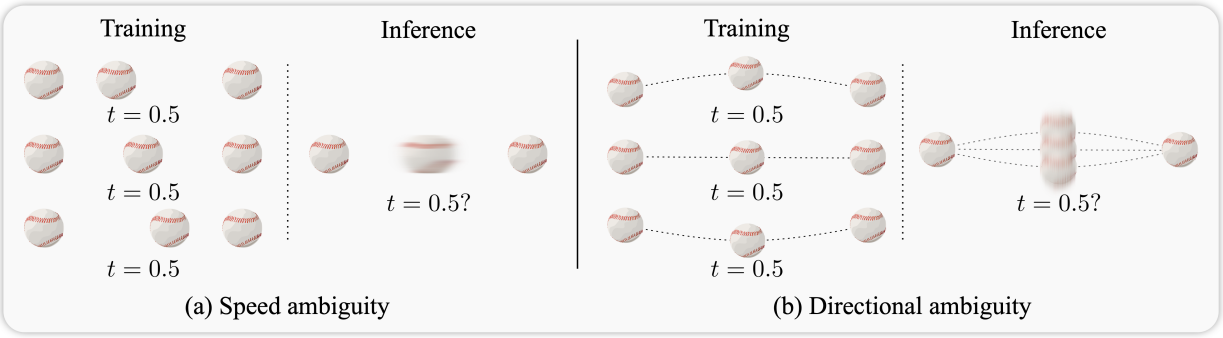


Fig. 2. Velocity ambiguity. (a) Speed ambiguity. (b) Directional ambiguity.

scale structure. VFIFormer [23] employs Transformer architectures to model long-range pixel correlations, while VFI-Mamba [24] adopts a Mamba-based architecture for efficient sequence modeling. IFRNet [25], RIFE [7], and UPR-Net [26] employ efficient pyramid network designs for high-quality, real-time interpolation, with IFRNet and RIFE using leakage distillation losses for flow estimation. Recently, more advanced network modules and operations are proposed to push the upper limit of VFI performance, such as the transformer-based bilateral motion estimator of BiFormer [27], a unifying operation of EMA-VFI [17] to explicitly disentangle motion and appearance information, and bi-directional correlation volumes for all pairs of pixels of AMT [18]. On the other hand, SoftSplat [28] and M2M [29] actively explore the forward warping operation for VFI. To further improve perceptual clarity in video frame interpolation, FILM [30] introduces perceptual losses during training, while uncertainty-aware and adaptive interpolation methods have also been proposed to better handle ambiguous regions [31], [32]. More recently, diffusion-based video frame interpolators, such as LDMVFI [33] and SVDKFI [34], have leveraged generative priors to enhance visual quality but still face challenges such as high computational costs and slow runtimes.

Other contributions to VFI come from various perspectives. For instance, Xu *et al*. [8], [35] leverage acceleration information from nearby frames, VideoINR [36] is the first to employ an implicit neural representation, and Lee *et al*. [37] explore and address discontinuity in video frame interpolation using figure-text mixing data augmentation and a discontinuity map. Flow-free approaches have also attracted interest. SepConv [38] integrates motion estimation and pixel synthesis, CAIN [39] employs the PixelShuffle operation with channel attention, and FLAVR [40] utilizes 3D space-time convolutions. Additionally, specialized interpolation methods for anime, which often exhibit minimal textures and exaggerated motion, are proposed by AnimeInterp [41] and Chen *et al*. [42]. On the other hand, motion induced blur [43]–[45], shutter mode [46]–[48], and event camera [49], [50] are also exploited to achieve VFI. For a more comprehensive overview of recent advances in video frame interpolation, we refer readers to the survey by Kye *et al*. [51].

2) Learning paradigms: One major thread of VFI methods train networks on triplet of frames, always predicting the central frame. Iterative estimation is used for interpolation ratios higher than $\times 2$. This *fixed-time* method often accumulates errors and struggles with interpolating at arbitrary continuous timesteps. Hence, models like SuperSloMo [6], DAIN [1], BMBC [20], EDSC [52], RIFE [7], IFRNet [25], EMA-VFI [17], and AMT [18] have adopted an *arbitrary time* interpolation paradigm. While theoretically superior, the arbitrary approach faces challenges of more complicated time-to-position mappings due to the velocity ambiguity, resulting in blurred results. This study addresses velocity ambiguity in arbitrary time interpolation and offers solutions.

Prior work by Zhou *et al*. [53] identified motion ambiguity and proposed a texture consistency loss to implicitly ensure interpolated content resemblance to given frames. In contrast, we explicitly address velocity ambiguity and propose solutions. These innovations not only enhance the performance of arbitrary time VFI models but also offer advanced manipulation capabilities. Additionally, we demonstrate that leveraging information from nearby frames through a multi-frame refiner module, combined with continuous indexing map estimation, can further improve interpolation quality.

3) Segment anything: The emergence of Segment Anything Model (SAM) [11] has marked a significant advancement in the realm of zero-shot segmentation, enabling numerous downstream applications including video tracking and segmentation [54], breakthrough mask-free inpainting techniques [55], and interactive image description generation [56]. By specifying the distance indexing individually for each segment, this work introduces a pioneering application to this growing collection: Manipulated Interpolation of Anything.

III. Velocity Ambiguity

In this section, we begin by revisiting the time indexing paradigm. We then outline the associated velocity ambiguity, which encompasses both speed and directional ambiguities.

Fig. 2 (a) shows the example of a horizontally moving baseball. Given a starting frame and an ending frame, along with a time indexing variable $t = 0.5$, the goal of a VFI model is to predict a latent frame at this particular timestep, in accordance with Eq. (1).

Although the starting and ending positions of the baseball are given, its location at $t = 0.5$ remains ambiguous due to an unknown speed distribution: The ball can be accelerating

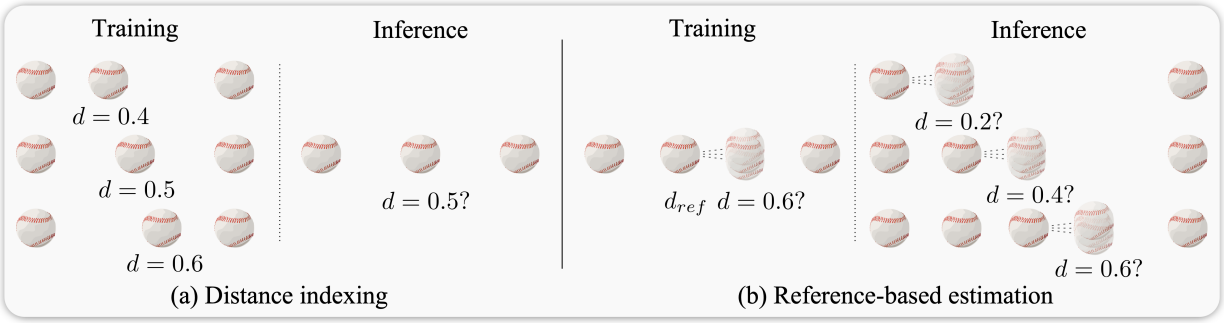


Fig. 3. Disambiguation strategies for velocity ambiguity. (a) Distance indexing. (b) Iterative reference-based estimation.

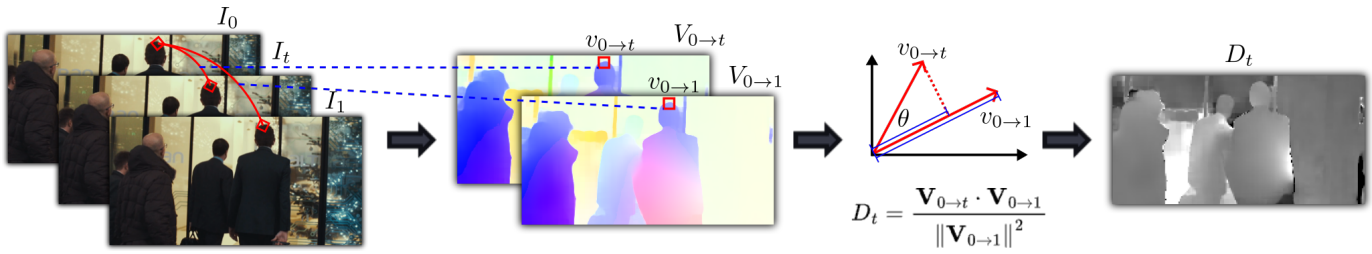


Fig. 4. Calculation of distance map for distance indexing. $V_{0\to t}$ is the estimated optical flow from I_0 to I_t by RAFT [9], and $V_{0\to 1}$ is the optical flow from I_0 to I_1.

or decelerating, resulting in different locations. This ambiguity introduces a challenge in model training as it leads to multiple valid supervision targets for the identical input. Contrary to the deterministic scenario illustrated in Eq. (1), the VFI function $\mathcal{F}$ is actually tasked with generating a sampled sequence of plausible frames within the *distribution* derived from the same input frames and time indexing. This can be expressed as:

$$\left\{I_t^1, I_t^2, \ldots, I_t^n\right\} = \mathcal{F}(I_0, I_1, t), \tag{3}$$

where n is the number of plausible frames. Empirically, the model, when trained with this ambiguity, tends to produce a weighted average of possible frames during inference. While this minimizes the loss during training, it results in blurry frames that are perceptually unsatisfying to humans, as shown in Fig. 1 (a). This blurry prediction $\hat{I}_t$ can be considered as an average over all the possibilities if an L_2 loss is used:

$$\hat{I}_t = \mathbb{E}_{I_t \sim \mathcal{F}(I_0, I_1, t)}[I_t]. \quad \text{(See details in Appendix)} \tag{4}$$

For other losses, Eq. (4) no longer holds, but we empirically observe that the model still learns an aggregated mixture of training frames which results in blur (RIFE [7] and EMA-VFI [17]: Laplacian loss, *i.e.*, L1 loss between the Laplacian pyramids of image pairs; IFRNet [25] and AMT [18]: Charbonnier loss).

Indeed, not only the speed but also the direction of motion remains indeterminate, leading to what we term as "directional ambiguity." This phenomenon is graphically depicted in Fig. 2 (b). This adds an additional layer of complexity in model training and inference. We collectively refer to speed ambiguity and directional ambiguity as velocity ambiguity.

So far, we have been discussing the ambiguity for the fixed time interpolation paradigm, in which t is set by default to 0.5. For arbitrary time interpolation, the ambiguity becomes more pronounced: Instead of predicting a single timestep, the network is expected to predict a continuum of timesteps between 0 and 1, each having a multitude of possibilities. This further complicates learning. Moreover, this ambiguity is sometimes referred to as *mode averaging*, which has been studied in other domains [57]. See Appendix for details.

IV. Disambiguation Strategies

In this section, we introduce two innovative strategies, namely distance indexing and iterative reference-based estimation, aimed at addressing the challenges posed by the velocity ambiguity. Designed to be plug-and-play, these strategies can be seamlessly integrated into any existing VFI models without necessitating architectural modifications, as shown in Fig. 1 (b).

In traditional time indexing, models intrinsically deduce an uncertain time-to-location mapping, represented as $\mathcal{D}$:

$$I_t = \mathcal{F}(I_0, I_1, \mathcal{D}(t)). \tag{5}$$

This brings forth the question: Can we guide the network to interpolate more precisely without relying on the ambiguous mapping $\mathcal{D}(t)$ to decipher it independently? To address this, we introduce a strategy to diminish speed uncertainty by directly specifying a distance ratio map (D_t) instead of the uniform timestep map. This is termed as distance indexing. Consequently, the model sidesteps the intricate process of deducing the time-to-location mapping:

$$I_t = \mathcal{F}(I_0, I_1, D_t). \tag{6}$$

A. Distance indexing

We utilize an off-the-shelf optical flow estimator, RAFT [9], to determine the pixel-wise distance map, as shown in Fig. 4. Given an image triplet $\{I_0, I_1, I_t\}$, we first calculate the optical flow from I_0 to I_t, denoted as $\mathbf{V}_{0\to t}$, and from I_0 to I_1 as $\mathbf{V}_{0\to 1}$. At each pixel (x, y), we project the motion vector $\mathbf{V}_{0\to t}(x, y)$ onto $\mathbf{V}_{0\to 1}(x, y)$. The distance map is then defined as the ratio between the projected $\mathbf{V}_{0\to t}(x, y)$ and $\mathbf{V}_{0\to 1}(x, y)$:

$$D_t(x, y) = \frac{\mathbf{V}_{0\to t}(x, y) \cdot \mathbf{V}_{0\to 1}(x, y)}{\|\mathbf{V}_{0\to 1}(x, y)\|^2}, \tag{7}$$

where θ denotes the angle between the two. By directly integrating D_t, the network achieves a clear comprehension of distance during its training phase, subsequently equipping it to yield sharper frames during inference, as showcased in Fig. 3 (a).

During inference, the algorithm does not have access to the exact distance map since I_t is unknown. In practice, we notice it is usually sufficient to provide a uniform map $D_t = t$, similar to time indexing. Physically this encourages the model to move each object at constant speeds along their trajectories. We observe that constant speed between frames is a valid approximation for many real-world situations. In Section VI, we show that even though this results in pixel-level misalignment with the ground-truth frames, it achieves significantly higher perceptual scores and is strongly preferred in the user study. Precise distance maps can be computed from multiple frames, which can potentially further boost the performance. See a detailed discussion in Appendix.

B. Iterative reference-based estimation

While distance indexing addresses speed ambiguity, it omits directional information, leaving directional ambiguity unresolved. Our observations indicate that, even with distance indexing, frames predicted at greater distances from the starting and ending frames remain not clear enough due to this ambiguity. To address this, we propose an iterative reference-based estimation strategy, which divides the complex interpolation for long distances into shorter, easier steps. This strategy enhances the traditional VFI function, $\mathcal{F}$, by incorporating a reference image, I_{ref}, and its corresponding distance map, D_{ref}. Specifically, the network now takes the following channels as input:

$$I_t = \mathcal{F}(I_0, I_1, D_t, I_{\text{ref}}, D_{\text{ref}}). \tag{8}$$

In the general case of N steps, the process of iteration is as follows:

$$I_{(i+1)t/N} = \mathcal{F}(I_0, I_1, D_{(i+1)t/N}, I_{t/N}, D_{t/N}), \tag{9}$$

where $i \in \{0, 1, \ldots, N-1\}$. For example, if we break the estimation of a remote step t into two steps:

$$I_{t/2} = \mathcal{F}(I_0, I_1, D_{t/2}, I_0, D_0), \tag{10}$$

$$I_t = \mathcal{F}(I_0, I_1, D_t, I_{t/2}, D_{t/2}). \tag{11}$$

Importantly, in every iteration, we consistently use the starting and ending frames as reliable appearance references, preventing divergence of uncertainty in each step. By dividing a long step into shorter steps, the uncertainty in each step is reduced, as shown in Fig. 3 (b). While fixed time models also employ an iterative method in a bisectioning way during inference, our strategy progresses from near to far, ensuring more deterministic trajectory interpolation. This reduces errors and uncertainties tied to a single, long-range prediction. **See more on the rationale for solving ambiguities in Appendix.**

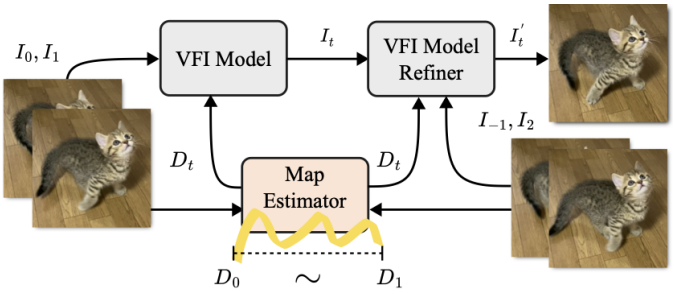


Fig. 5. Multi-frame fusion architecture with continuous map estimator.

V. Leveraging Information from Nearby Frames

In this section, we first introduce a continuous indexing map estimator to achieve pixel-wise distance map estimation. Next, we demonstrate how to reuse the original interpolation architecture to fuse nearby frames for enhancing interpolation quality. The complete framework is illustrated in Fig. 5.

A. Pixel-wise distance map estimation

To achieve a continuous indexing map estimator, we adopt the pretrained model introduced in CPFlow [12]. CPFlow takes a sequence of images as input (four in our case) and predicts the optical flow from the initial frame to any arbitrary timestamp within the sequence. Specifically, it models each pixel's motion trajectory using cubic B-splines, enabling dense and temporally continuous flow estimation. Given normalized time $t \in [0, 1]$, the displacement of a pixel is defined as:

$$\mathbf{V}_{0\to t} = \sum_{i=0}^{N-1} B_{i,k}(t)\mathbf{P}_i, \tag{12}$$

where $\mathbf{P}_i$ are learnable control points and $B_{i,k}(t)$ are spline basis functions defined recursively [12]. To enhance temporal consistency, the model employs a Neural Ordinary Differential Equation (NODE) module in combination with ConvGRU, where the hidden feature state $h(t)$ evolves according to:

$$\frac{dh(t)}{dt} = f(h(t), t), \quad h(t) = h(t_0) + \int_{t_0}^{t} f(h(\tau), \tau)\, d\tau. \tag{13}$$

The hidden state $h(t)$ is refined with frame-specific features ε_t via ConvGRU: $\tilde{h}(t) = \text{ConvGRU}(h(t), \varepsilon_t)$. The model then computes multi-scale correlation volumes between the reference feature $\tilde{h}(0)$ and $\tilde{h}(t)$, denoted by $C(t) = \text{Corr}(\tilde{h}(0), \tilde{h}(t))$. These correlations are used in an iterative decoder to update the control points of the spline as:

$$\mathbf{P}_i^{(s+1)} = \mathbf{P}_i^{(s)} + \Delta\mathbf{P}_i^{(s)}(C(t)), \tag{14}$$

where s is the iteration index. The final continuous optical flow $\mathbf{V}_{0\to t}$ is reconstructed from the refined control points using the spline formulation above. This continuous formulation allows the model to produce high-fidelity indexing maps using Eq. (7) for arbitrary interpolation times $D_t = \|\mathbf{V}_{0\to t}\| \cos\theta / \|\mathbf{V}_{0\to 1}\|$.

B. Multi-frame refiner

We further design a multi-frame fusion module to leverage the relevant pixel appearance information beyond the two adjacent frames. As a brief review, flow-based VFI models [7] typically adhere to the following formulation:

$$I_t^+, I_t^-, M = \mathcal{F}(I_0, I_1, D_t), \tag{15}$$

where M is a one-channel blending mask, and I_t^+ and I_t^- denote the warped images derived from I_0 and I_1 using the predicted optical flows. The final interpolated frame I_t is obtained by:

$$I_t = M \odot I_t^+ + (1 - M) \odot I_t^- \tag{16}$$

An additional residual term is often included which is omitted here. To adapt the model to accept four consecutive frames as input while maintaining plug-and-play compatibility with various VFI models, we introduce a simple yet effective framework: we create a trainable copy of the two-frame VFI model $\mathcal{F}'$, which accepts the additional frames I_{-1} and I_2 along with a new distance map D_t' computed relative to I_{-1} and I_2. Additionally, the initially interpolated frame I_t is provided, enabling the network to refine the result by utilizing this supplementary information:

$$I_t^{'+}, I_t^{'-}, M' = \mathcal{F}'(I_{-1}, I_2, D_t', I_t), \tag{17}$$

where $I_t^{'+}, I_t^{'-}$ are warped version of I_{-1}, I_2 respectively. $M' = [M_1, M_2, M_3]$ is a three-channel blending mask such that $M_1 + M_2 + M_3 = 1$ at each pixel, corresponding to $I_t^{'+}$, I_t, and $I_t^{'-}$. The final refined frame I_t' is then computed as:

$$I_t' = M_1 \odot I_t^{'+} + M_2 \odot I_t + M_3 \odot I_t^{'-} \tag{18}$$

Notice that we directly use the interpolated frame I_t instead of latent features as network input to ensure compatibility with different VFI models.

C. Model tuning

During training, we first freeze the parameters of the two-frame VFI model $\mathcal{F}$ and the continuous map estimator. Only the refiner $\mathcal{F}'$ is updated during this stage, using the same training configuration as the original model. This setup enables the refiner to learn how to enhance interpolation by leveraging information from more distant adjacent frames. Next, we freeze the parameters of the pretrained map estimator and jointly optimize both the VFI model $\mathcal{F}$ and the proposed refiner $\mathcal{F}'$. This approach enables the entire system to adapt to indexing maps derived from optical flow predicted by CPFlow. We experimented with various training strategies and loss functions to evaluate their effectiveness (see Section VI-I).

VI. Experiment

A. Implementation

We leveraged the plug-and-play nature of our distance indexing and iterative reference-based estimation strategies to seamlessly integrate them into influential arbitrary time VFI models such as RIFE [7] and IFRNet [25], and state-of-the-art models including AMT [18] and EMA-VFI [17]. To further strengthen the empirical evidence, we additionally validated the effectiveness of distance indexing on a diffusion-based model, LDMVFI [33], and a representative multi-frame method, VFI-Transformer [58]. We adhere to the original hyperparameters for each model for a fair comparison and implement them with PyTorch [59]. For training, we use the septuplet dataset from Vimeo90K [60]. The septuplet dataset comprises 91,701 seven-frame sequences at 448×256, extracted from 39,000 video clips. For evaluation, we use both pixel-centric metrics like PSNR and SSIM [61], and perceptual metrics such as reference-based LPIPS [62] and non-reference NIQE [63]. Concerning the iterative reference-based estimation strategy, D_{ref} during training is calculated from the optical flow derived from ground-truth data at a time point corresponding to a randomly selected reference frame, like $t/2$. In the inference phase, we similarly employ a uniform map for reference, for example, setting $D_{ref} = t/2$.

B. Qualitative comparison

1) Qualitative analysis: We conducted a qualitative analysis on different variants of each arbitrary time VFI model. We evaluate the base model, labeled as $[T]$, against its enhanced versions, which incorporate distance indexing ($[D]$), iterative reference-based estimation ($[T, R]$), or a combination of both ($[D, R]$), as shown in Fig. 6. We observe that the $[T]$ model yields blurry results with details difficult to distinguish. Models with the distance indexing ($[D]$) mark a noticeable enhancement in perceptual quality, presenting clearer interpolations than $[T]$. In most cases, iterative reference ($[T, R]$) also enhances model performance, with the exception of AMT-S. As expected, the combined approach $[D, R]$ offers the best quality for all base models including AMT-S. This highlights the synergistic potential of distance indexing when paired with iterative reference-based estimation. Overall, our findings underscore the effectiveness of both techniques as plug-and-play strategies, capable of significantly elevating the qualitative performance of cutting-edge arbitrary time VFI models.

2) User study: To validate the effectiveness of our proposed strategies, we further conducted a user study with 30 anonymous participants. Participants were tasked with ranking the interpolation quality of frames produced by four model variants: $[T]$, $[D]$, $[T, R]$, and $[D, R]$. See details of user study UI in Appendix. The results, presented in Fig. 7, align with our qualitative and quantitative findings. The $[D, R]$ model variant emerged as the top-rated, underscoring the effectiveness of our strategies.

C. Quantitative comparison

1) Convergence curves: To further substantiate the efficacy of our proposed strategies, we also conducted a quantitative

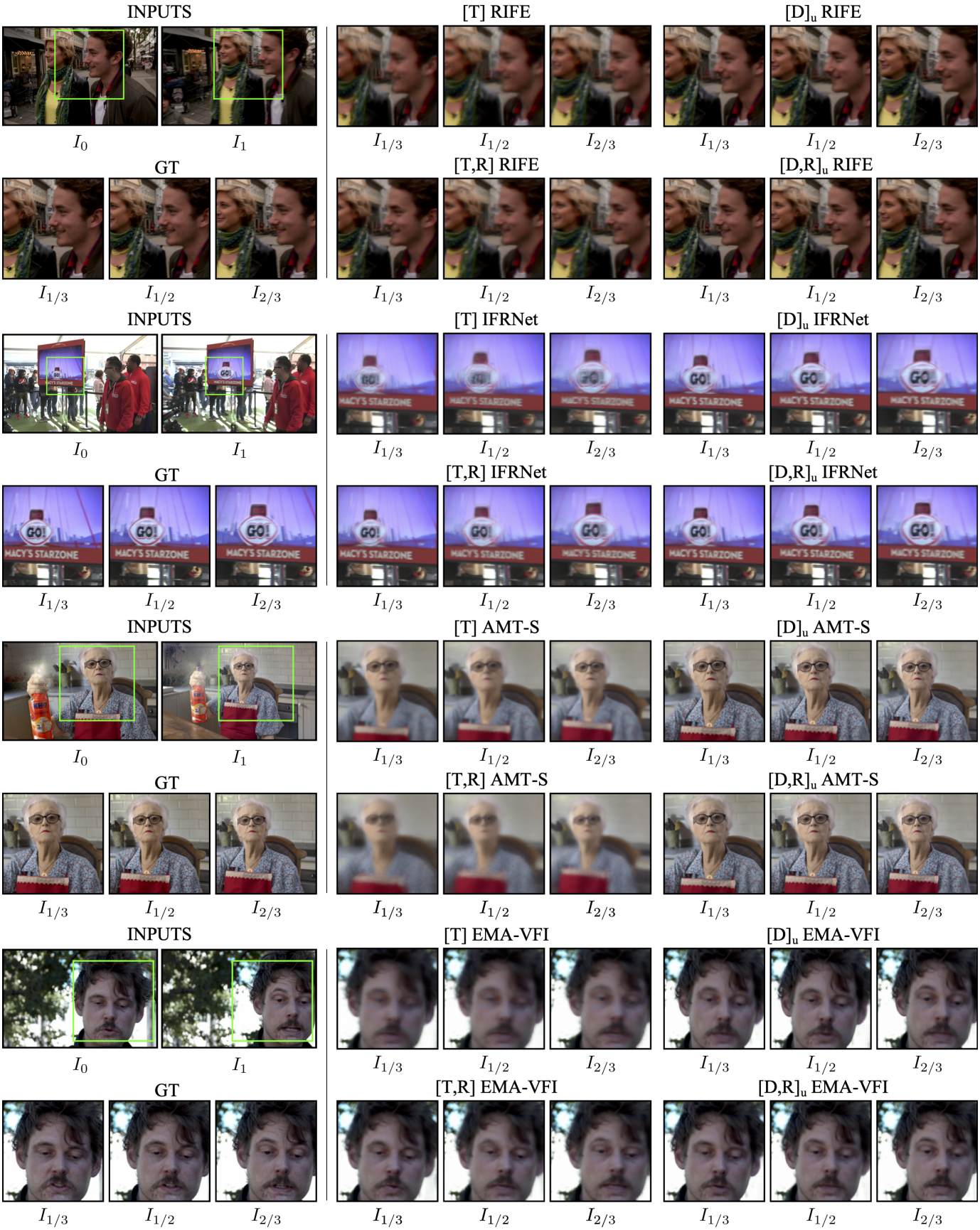


Fig. 6. Comparison of qualitative results. $[T]$: original arbitrary time VFI models using time indexing. $[D]_u$: models trained using our distance indexing, then inference using uniform maps. $[T, R]$: models using time indexing with iterative reference-based estimation. $[D, R]_u$: models trained using both distance indexing and iterative reference-based estimation, then inference using uniform maps.

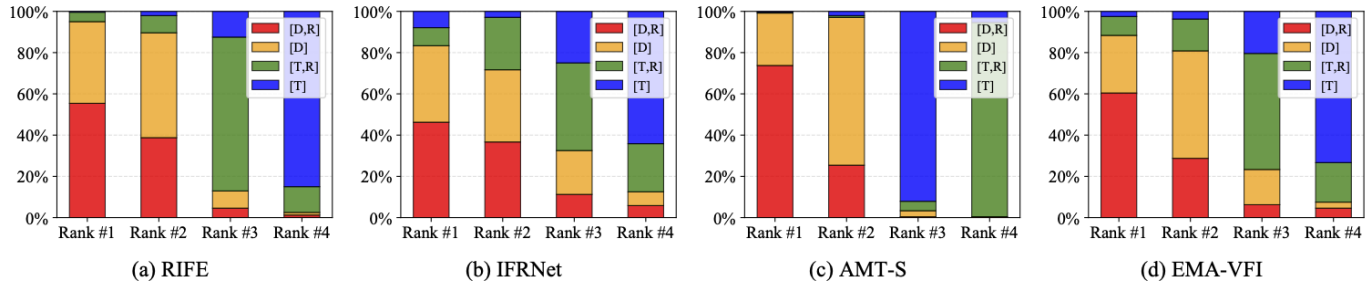


Fig. 7. User study. The horizontal axis represents user rankings, where #1 is the best and #4 is the worst. The vertical axis indicates the percentage of times each model variant received a specific ranking. Each model variant was ranked an equal number of times. The model $[D, R]$ emerged as the top performer. All models use uniform maps.

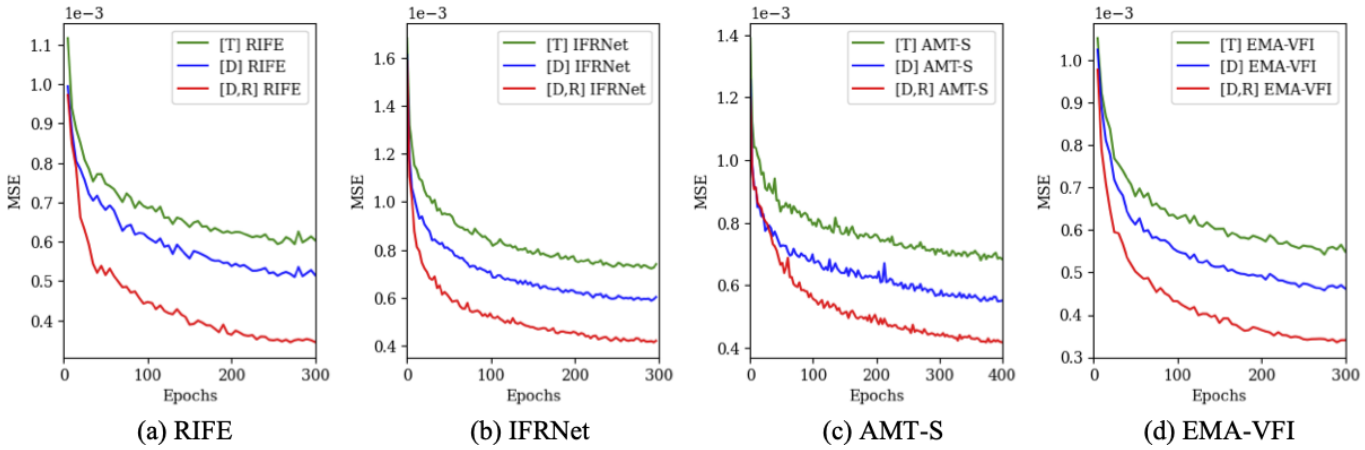


Fig. 8. Convergence curves. $[T]$ denotes traditional time indexing. $[D]$ denotes the proposed distance indexing. $[R]$ denotes iterative reference-based estimation.

TABLE I

COMPARISON ON VIMEO90K SEPTUPLET DATASET. $[T]$ DENOTES THE METHOD TRAINED WITH TRADITIONAL ARBITRARY TIME INDEXING PARADIGM. $[D]$ AND $[R]$ DENOTE THE DISTANCE INDEXING PARADIGM AND ITERATIVE REFERENCE-BASED ESTIMATION STRATEGY, RESPECTIVELY. $[R]$ USES 2 ITERATIONS BY DEFAULT. $[\cdot]_u$ DENOTES INFERENCE WITH UNIFORM MAP AS TIME INDEXES. WE UTILIZE THE FIRST AND LAST FRAMES AS INPUTS TO PREDICT THE REST FIVE FRAMES. THE **BOLD FONT** DENOTES THE BEST PERFORMANCE IN CASES WHERE COMPARISON IS POSSIBLE. WHILE THE GRAY FONT INDICATES THAT THE SCORES FOR PIXEL-CENTRIC METRICS, PSNR AND SSIM, ARE NOT CALCULATED USING STRICTLY ALIGNED GROUND-TRUTH AND PREDICTED FRAMES.

	RIFE [7]			IFRNet [25]			AMT-S [18]			EMA-VFI [17]		
	$[T]$	$[D]$	$[D, R]$	$[T]$	$[D]$	$[D, R]$	$[T]$	$[D]$	$[D, R]$	$[T]$	$[D]$	$[D, R]$
PSNR↑	28.22	**29.20**	28.84	28.26	**29.25**	28.55	28.52	**29.61**	28.91	29.41	**30.29**	25.10
SSIM↑	0.912	**0.929**	0.926	0.915	**0.931**	0.925	0.920	**0.937**	0.931	0.928	**0.942**	0.858
LPIPS↓	0.105	0.092	**0.081**	0.088	0.080	**0.072**	0.101	0.086	**0.077**	0.086	**0.078**	0.079
NIQE↓	6.663	6.475	**6.286**	6.422	6.342	**6.241**	6.866	6.656	**6.464**	6.736	6.545	**6.241**
	$[T]$	$[D]_u$	$[D, R]_u$	$[T]$	$[D]_u$	$[D, R]_u$	$[T]$	$[D]_u$	$[D, R]_u$	$[T]$	$[D]_u$	$[D, R]_u$
PSNR↑	**28.22**	27.55	27.41	**28.26**	27.40	27.13	**28.52**	27.33	27.17	**29.41**	28.24	24.73
SSIM↑	**0.912**	0.902	0.901	**0.915**	0.902	0.899	**0.920**	0.902	0.902	**0.928**	0.912	0.851
LPIPS↓	0.105	0.092	**0.086**	0.088	0.083	**0.078**	0.101	0.090	**0.081**	0.086	**0.079**	0.081
NIQE↓	6.663	6.344	**6.220**	6.422	6.196	**6.167**	6.866	6.452	**6.326**	6.736	6.457	**6.227**

analysis. Fig. 8 shows the convergence curves for different model variants, *i.e.*, $[T]$, $[D]$, and $[D, R]$. The observed trends are consistent with our theoretical analysis from Section IV, supporting the premise that by addressing velocity ambiguity, both distance indexing and iterative reference-based estimation can enhance convergence limits.

2) Comparison on Vimeo90K septuplet dataset: In Table I, we provide a performance breakdown for each model variant. The models $[D]$ and $[D, R]$ in the upper half utilize ground-truth distance guidance, which is not available at inference in practice. The goal here is just to show the achievable upper-bound performance. On both pixel-centric metrics such as PSNR and SSIM, and perceptual measures like LPIPS

TABLE II
ABLATION STUDY OF THE NUMBER OF ITERATIONS ON VIMEO90K SEPTUPLET DATASET. $[\cdot]^{\#}$ DENOTES THE NUMBER OF ITERATIONS USED FOR INFERENCE.

	RIFE [7]			IFRNet [25]			AMT-S [18]			EMA-VFI [17]		
$[D,R]_u$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$
LPIPS↓	0.093	0.086	**0.085**	0.085	**0.078**	0.078	0.086	**0.081**	0.081	0.084	0.081	**0.080**
NIQE↓	6.331	6.220	**6.186**	6.205	6.167	**6.167**	6.402	**6.326**	6.327	6.303	6.227	**6.211**
$[T,R]$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$
LPIPS↓	0.103	0.087	**0.087**	0.091	0.084	**0.084**	**0.106**	0.135	0.157	0.088	**0.083**	0.085
NIQE↓	6.551	6.300	**6.206**	6.424	6.347	**6.314**	**6.929**	7.246	7.502	6.404	6.280	**6.246**

and NIQE, the improved versions $[D]$ and $[D,R]$ outperform the base model $[T]$. Notably, the combined model $[D,R]$ using both distance indexing and iterative reference-based estimation strategies performs superior in perceptual metrics, particularly NIQE. The superior pixel-centric scores of model $[D]$ compared to model $[D,R]$ can be attributed to the indirect estimation (2 iterations) in the latter, causing slight misalignment with the ground-truth, albeit with enhanced details.

In realistic scenarios where the precise distance map is inaccessible at inference, one could resort to a uniform map akin to time indexing. The bottom segment of Table I shows the performance of the enhanced models $[D]$ and $[D,R]$, utilizing identical inputs as model $[T]$. Given the misalignment between predicted frames using a uniform distance map and the ground-truth, the enhanced models do not outperform the base model on pixel-centric metrics. However, we argue that in most applications, the goal of VFI is not to predict pixel-wise aligned frames, but to generate plausible frames with high perceptual quality. Furthermore, pixel-centric metrics are less sensitive to blur [62], the major artifact introduced by velocity ambiguity. The pixel-centric metrics are thus less informative and denoted in gray. On perceptual metrics (especially NIQE), the enhanced models significantly outperforms the base model. This consistency with our qualitative observations further validates the effectiveness of distance indexing and iterative reference-based estimation.

3) Ablation study of the number of iterations: Table II offers an ablation study on the number of iterations and the efficacy of a pure iterative reference-based estimation strategy. The upper section suggests that setting iterations at two strikes a good trade-off between computational efficiency and performance. The lower segment illustrates that while iterative reference-based estimation generally works for time indexing, there are exceptions, as observed with AMT-S. However, when combined with distance indexing, iterative reference-based estimation exhibits more stable improvement, as evidenced by the results for $[D,R]_u$. This is consistent with qualitative comparison.

4) Comparison on other benchmarks: The septuplet set of Vimeo90K [60] is large enough to train a practical video frame interpolation model, and it represents the situations where the temporal distance between input frames is large. Thus, Vimeo90K (septuplet) can well demonstrate the velocity ambiguity problem that our work aims to highlight. We further show $\times 16$ interpolation on X4K1000FPS with larger temporal distance in Table III. The results highlight that the benefits of our strategies are more pronounced with increased temporal distances.

TABLE III
COMPARISON ON X4K1000FPS [22] FOR $\times 16$ INTERPOLATION WITH RIFE [7].

	$[T]$	$[D]_u$	$[D,R]_u$
PSNR ↑	31.04	**31.60**	31.52
SSIM ↑	0.910	0.914	**0.922**
LPIPS ↓	0.104	0.094	**0.079**
NIQE ↓	7.215	6.953	**6.927**

TABLE IV
COMPARISON ON VIMEO90K [60] USING GMFLOW [64] FOR DISTANCE MAP CALCULATION WITH RIFE [7].

	$[T]$	$[D]_u$	$[D,R]_u$
PSNR ↑	**28.22**	27.29	26.96
SSIM ↑	**0.912**	0.898	0.895
LPIPS ↓	0.105	0.101	**0.092**
NIQE ↓	6.663	6.449	**6.280**

5) Other optical flow estimator: We also employ GMFlow [64] for the precomputation of distance maps, enabling an analysis of model performance when integrated with alternative optical flow estimations. The results are as shown in Table IV. Our strategies still lead to consistent improvement on perceptual metrics. However, this more recent and performant optical flow estimator does not introduce improvement compared to RAFT [9]. A likely explanation is that since we quantify the optical flow to $[0,1]$ scalar values for better generalization, our training strategies are less sensitive to the precision of the optical flow estimator.

6) Comparison of using perceptual loss: In addition to training with traditional pixel losses based on L1 and L2 losses, we present the results of employing the more recent LPIPS loss [62] with a VGG backbone [65], as shown in Table V. The non-reference perceptual quality metric, NIQE, shows notable improvement across all variants. The results also consistently demonstrate the effectiveness of our strategies in resolving velocity ambiguity. Besides, due to the direct optimization of LPIPS loss, the assumption of constant speed in uniform maps affects the performance for this metric. This

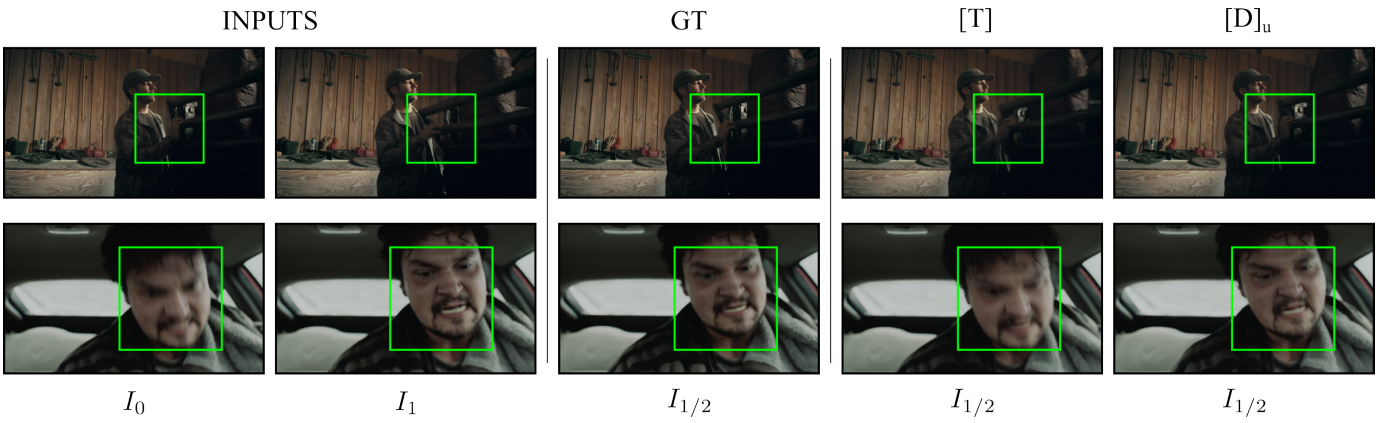


Fig. 9. Comparison on Vimeo90K Septuplet with LDMVFI [33].

TABLE V
COMPARISON ON THE SEPTUPLET OF VIMEO90K [60] USING LPIPS LOSS [62]. WE USE RIFE [7] AS A REPRESENTATIVE EXAMPLE.

	$[T]$	$[D]_u$	$[D, R]_u$
PSNR ↑	**27.19**	26.71	26.72
SSIM ↑	**0.898**	0.889	0.890
LPIPS ↓	**0.061**	0.065	0.064
NIQE ↓	6.307	5.901	**5.837**

is why in the test results, $[T]$ has a lower LPIPS, while $[D]_u$ and $[D, R]_u$ are slightly higher.

7) Temporal consistency: We have further evaluated models using FVD [66] and VBench [67]. As shown in Table VI, replacing time indexing with distance indexing consistently improves FVD scores as well as VBench metrics related to subject consistency, background consistency, and overall imaging quality. We notice that the *motion smoothness* metric in VBench [67] shows limited discriminative power for video frame interpolation. These results indicate that reducing velocity ambiguity not only enhances per-frame perceptual quality but also leads to more temporally coherent and stable video interpolation.

D. Evaluating distance indexing on diffusion-based baseline

We further evaluate diffusion-based VFI models, which are known for their strong generative capacity. Since SVD-KFI [34] directly generates a fixed sequence of multiple frames in a single pass, we adopt LDMVFI [33] for evaluation, whose formulation is more compatible with our interpolation setting. As reported in Table VII, even though diffusion models may partially mitigate ambiguity through generative priors, replacing time indexing with distance indexing still leads to further improvements. We also provide a qualitative comparison in Fig. 9. These results indicate that velocity ambiguity remains relevant for diffusion-based approaches and that distance indexing provides complementary benefits.

E. Evaluating distance indexing on multi-frame baseline

We also evaluate a representative multi-frame input method, Video Frame Interpolation Transformer (VFI-Transformer) [58]. As shown in Table VIII and Fig. 10, distance indexing also improves performance in the multi-frame setting, confirming that additional temporal context alone does not fully eliminate velocity ambiguity and that the proposed strategy remains beneficial.

F. 2D manipulation of frame interpolation

Beyond simply enhancing the performance of VFI models, distance indexing equips them with a novel capability: tailoring the interpolation patterns for each individual object, termed as "manipulated interpolation of anything". Fig. 11 demonstrates the workflow. The first stage employs SAM [11] to produce object masks for the starting frame. Users can then customize the distance curve for each object delineated by the mask, effectively controlling its interpolation pattern, *e.g.*, having one person moving backward in time. The backend of the application subsequently generates a sequence of distance maps based on these specified curves for interpolation. One of the primary applications is re-timing specific objects **(See the supplementary video)**.

G. Multi-frame qualitative comparison

As shown in Fig. 12, we present a qualitative comparison of different variants of each VFI model under multi-frame setup. $[D, M]_e$ denotes the model that incorporate both the indexing map with multi-frame refiner and continuous indexing map estimator. As anticipated, the $[D, M]_e$ configuration consistently yields the highest visual quality across all base models, demonstrating the effectiveness of both the multi-frame refiner and the continuous indexing map estimator.

H. Multi-frame quantitative comparison

In Table IX, we present a quantitative performance comparison across various model variants. The configurations $[T]$, $[D]_u$, and $[D]$ retain the same definitions as previously described. $[D]_e$ denotes the model augmented with a continuous

TABLE VI
TEMPORAL CONSISTENCY EVALUATION ON VIMEO90K SEPTUPLET USING FVD [66] AND VBENCH [67].

	RIFE [7]		IFRNet [25]		AMT-S [18]		EMA-VFI [17]	
	$[T]$	$[D]_u$	$[T]$	$[D]_u$	$[T]$	$[D]_u$	$[T]$	$[D]_u$
FVD ↓	0.0174	**0.0137**	0.0142	**0.0137**	0.0167	**0.0137**	0.0119	**0.0108**
Subject Consistency ↑	0.957	**0.959**	0.958	**0.961**	0.958	**0.960**	0.962	**0.967**
Background Consistency ↑	0.952	**0.955**	0.952	**0.956**	0.953	**0.955**	0.955	**0.958**
Imaging Quality ↑	0.448	**0.475**	0.460	**0.481**	0.462	**0.485**	0.468	**0.486**
Motion Smoothness ↑	**0.996**	**0.996**	**0.996**	**0.996**	**0.996**	**0.996**	**0.996**	**0.996**

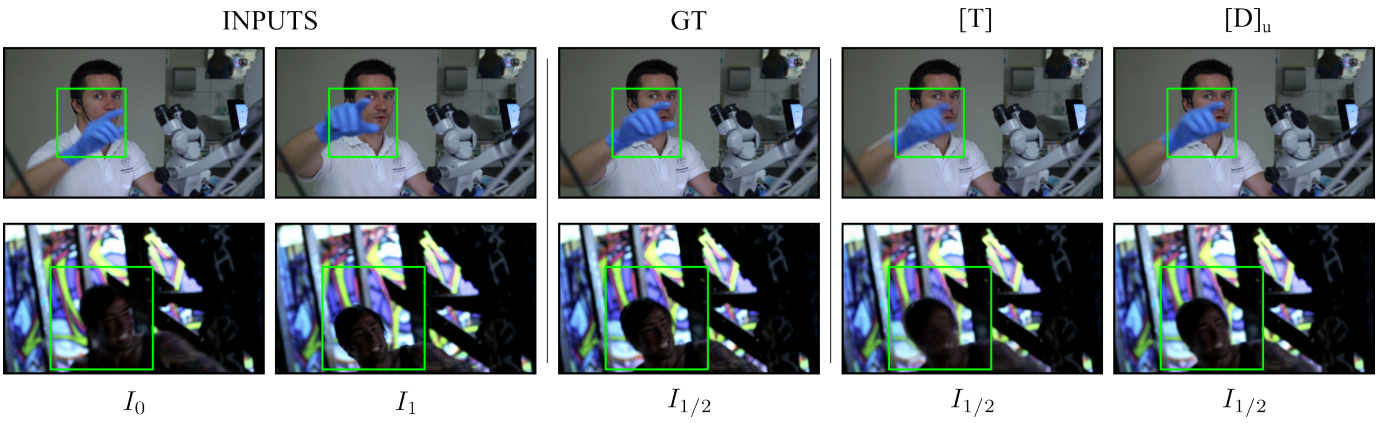


Fig. 10. Comparison on Vimeo90K Septuplet with VFI-Transformer [17].

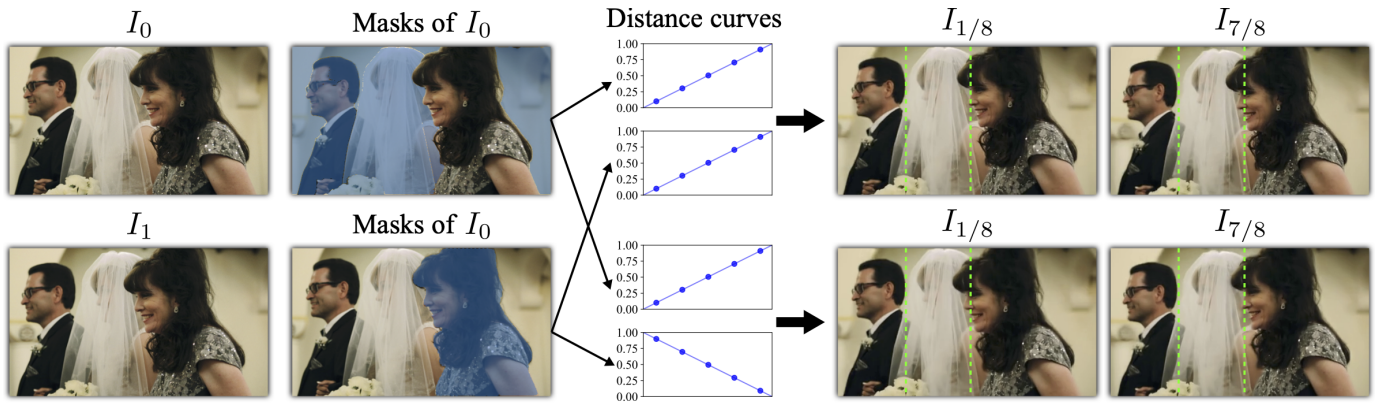


Fig. 11. Manipulated interpolation of anything. Leveraging Segment-Anything [11], users can tailor distance curves for selected masks. Distinct masks combined with varying distance curves generate unique distance map sequences, leading to diverse interpolation outcomes.

TABLE VII
COMPARISON ON VIMEO90K SEPTUPLET WITH LDMVFI [33].

	$[T]$	$[D]$	$[D]_u$
PSNR↑	26.83	**27.12**	25.94
SSIM↑	0.904	**0.906**	0.893
LPIPS↓	0.098	**0.086**	0.090
NIQE↓	6.652	6.481	**6.451**

TABLE VIII
COMPARISON ON VIMEO90K SEPTUPLET WITH VFI-TRANSFORMER [17].

	$[T]$	$[D]$	$[D]_u$
PSNR↑	36.09	**36.46**	35.43
SSIM↑	0.974	**0.976**	0.964
LPIPS↓	0.084	**0.079**	0.081
NIQE↓	6.251	**6.082**	6.197

indexing map estimator, $[D, M]$ includes the proposed multi-frame fusion with the refiner model, and $[D, M]_e$ incorporates both the refiner and the map estimator. $[T, M]$ denotes multi-frame fusion using the refiner model trained with the time indexing map. On pixel-level metrics such as PSNR and SSIM, the enhanced variant $[D, M]_e$ consistently outperforms the base model $[T]$ as well as $[D]_u$, which lacks a map estimator and relies on uniform indexing maps. The improvement from

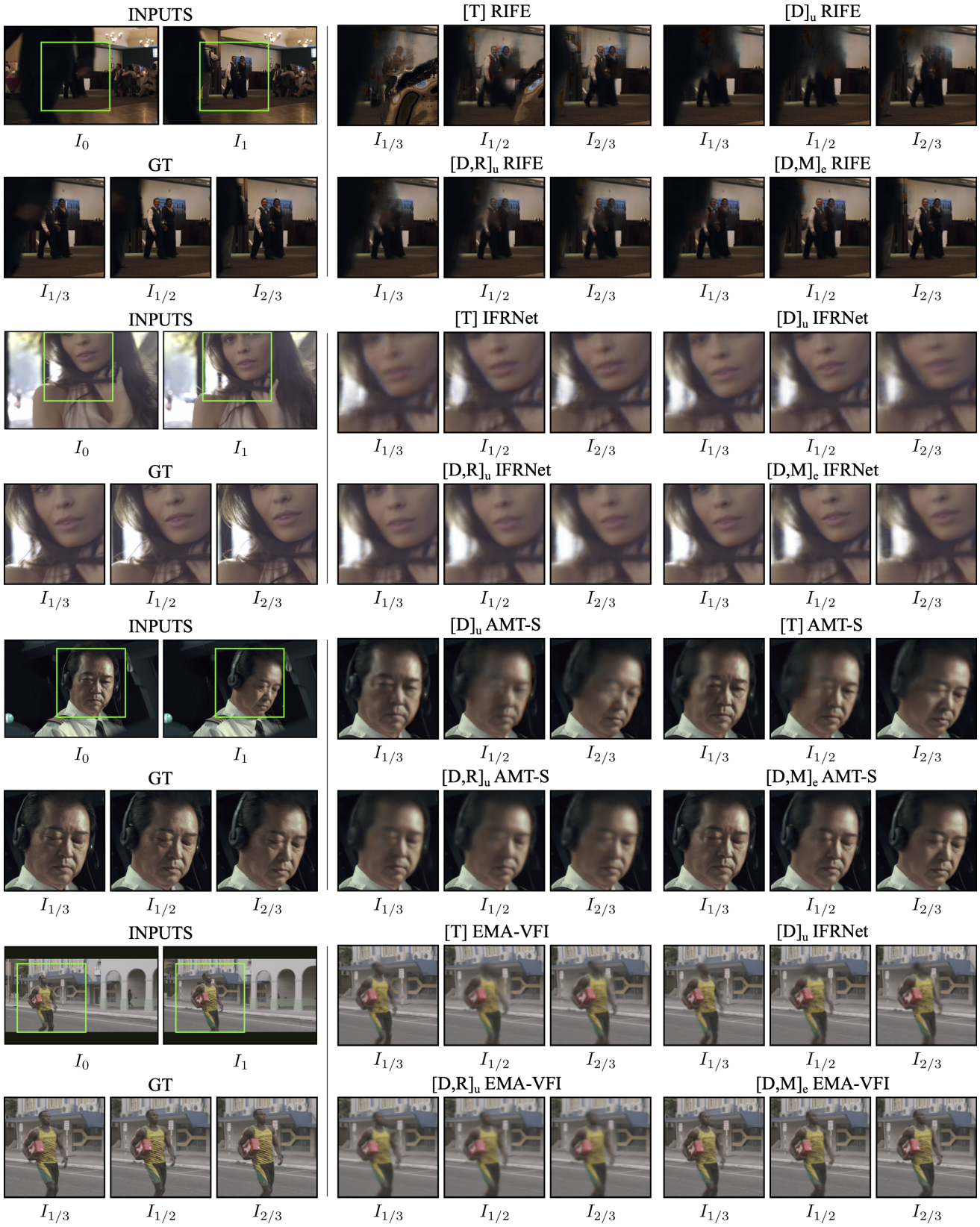


Fig. 12. Qualitative comparison under multi-frame setting. $[T]$: original arbitrary time VFI models using time indexing. $[D]_u$: models trained using distance indexing, then inference using uniform maps. $[D, R]_u$: models trained using both distance indexing and iterative reference-based estimation, then inference using uniform maps. $[D, M]_e$: models trained using distance indexing with multi-frame fusion, then inference using estimated maps. Due to space limitations, the two adjacent input frames are omitted in the visualization, while the actual input consists of four frames.

$[D]_u$ to $[D, M]_e$ highlights the effectiveness of integrating the continuous indexing map estimator. Meanwhile, performance gains from $[D]$ to $[D, M]$, and from $[D]_e$ to $[D, M]_e$, demonstrate the additional benefit provided by the multi-frame refiner. Moreover, the comparison between time-indexed ($[T, M]$) and distance-indexed ($[D, M]_e$, $[D, M]$) multi-frame video frame interpolation models again shows that our proposed distance indexing yields better interpolation results.

TABLE IX
MULTI-FRAME COMPARISON ON VIMEO90K SEPTUPLET DATASET. $[M]$ DENOTES THE MULTI-FRAME FUSION. $[\cdot]_e$ DENOTES INFERENCE WITH ESTIMATED INDEXING MAP. OTHER NOTATIONS HAVE THE SAME MEANING AS IN THE PREVIOUS EXPERIMENTS.

RIFE [7]	$[T]$	$[T, M]$	$[D]_u$	$[D]_e$	$[D, M]_e$	$[D]$	$[D, M]$
PSNR ↑	28.22	**28.84**	27.55	28.25	28.34	29.20	**31.63**
SSIM ↑	0.912	0.922	0.902	0.923	**0.928**	0.929	**0.952**
LPIPS ↓	0.105	0.097	0.092	0.099	**0.089**	0.092	**0.062**
NIQE ↓	6.663	6.518	6.344	6.554	**6.173**	6.475	**5.990**
IFRNet [25]	$[T]$	$[T, M]$	$[D]_u$	$[D]_e$	$[D, M]_e$	$[D]$	$[D, M]$
PSNR ↑	28.26	**28.75**	27.40	27.63	28.28	29.25	**32.11**
SSIM ↑	0.915	0.918	0.902	0.907	**0.919**	0.931	**0.958**
LPIPS ↓	0.088	0.085	0.083	0.087	**0.083**	0.080	**0.074**
NIQE ↓	6.422	6.388	**6.196**	6.414	6.249	6.342	**5.957**
AMT-S [18]	$[T]$	$[T, M]$	$[D]_u$	$[D]_e$	$[D, M]_e$	$[D]$	$[D, M]$
PSNR ↑	28.52	**28.91**	27.33	27.60	28.80	29.61	**31.80**
SSIM ↑	0.920	**0.924**	0.902	0.908	0.922	0.937	**0.955**
LPIPS ↓	0.101	0.094	0.090	0.098	**0.084**	0.086	**0.072**
NIQE ↓	6.866	6.598	6.382	6.452	**6.223**	6.656	**6.056**
EMA-VFI [17]	$[T]$	$[T, M]$	$[D]_u$	$[D]_e$	$[D, M]_e$	$[D]$	$[D, M]$
PSNR ↑	29.41	**29.80**	28.24	28.67	29.45	30.29	**31.31**
SSIM ↑	0.928	0.930	0.912	0.919	**0.932**	0.942	**0.951**
LPIPS ↓	0.086	0.086	**0.079**	0.082	0.084	0.078	**0.069**
NIQE ↓	6.736	6.597	6.457	6.609	**6.313**	6.545	**6.146**

TABLE X
MULTI-FRAME ABLATION STUDY

	FE			*VFI*	
	$\mathcal{L}_D$	$\mathcal{L}_{V_{b+f}}$	$\mathcal{L}_{D+V_{b+f}}$	$\mathcal{L}_{V_f}$	$\mathcal{L}_{V_{b+f}}$
PSNR ↑	26.10	27.22	27.34	30.89	**31.63**
SSIM ↑	0.891	0.898	0.901	0.938	**0.952**
LPIPS ↓	0.111	0.104	0.097	0.069	**0.062**
NIQE ↓	6.627	6.575	6.509	6.087	**5.990**

I. Multi-frame tuning strategy

In this section, we evaluate different training strategies for multi-frame fusion using RIFE [7] as the baseline for ablation studies. All experiments follow the same setting of previous experiments on Vimeo90k. We find that jointly tuning both refiner and map estimator leads to instability and failure to converge. Thus, we split the tuning process in two stages. In the first stage, we first finetune the multi-frame refiner $\mathcal{F}'$ alone using original VFI loss. In the second stage, there are two possible choices. Since jointly tuning with the map estimator remains unstable, we either finetune the flow estimator only to adapt it to VFI (referred to as *FE*), or finetune the VFI modules $\mathcal{F}$ and $\mathcal{F}'$ to adapt them to the pretrained flow estimator (referred to as *FE*), as shown in Table X. The loss term $\mathcal{L}_D$ is calculated by $\mathcal{L}_D = \|D_t - D_t^{\text{RAFT}}\|_2^2$, denoting the L2 loss between the indexing maps from CPFlow and RAFT. The setting $\mathcal{L}_{V_f}$ corresponds to training only the VFI refiner with its original loss, while $\mathcal{L}_{V_{b+f}}$ involves jointly training both the base VFI model and the refiner with their original loss. Among the tested strategies, jointly tuning both the VFI model and the refiner under $\mathcal{L}_{V_{b+f}}$ achieves the best performance. This improvement is attributed to the model's ability to adapt both components to the updated indexing map derived from the newly predicted optical flow.

TABLE XI
COSTS COMPARISON INCLUDING RUNTIME AND NUMBER OF PARAMETERS. WE EVALUATE THE COMPUTATIONAL OVERHEAD ON AN NVIDIA A100 GPU USING IMAGES WITH A RESOLUTION OF 448×256.

	$[D]$		$[D, M]$		$[D, M]_e$	
	Sec.	MB	Sec.	MB	Sec.	MB
RIFE [7]	0.03	10.21	0.06	20.46	0.10	30.68
IFRNet [25]	0.14	4.73	0.17	14.91	0.21	25.13
AMT-S [18]	0.20	2.86	0.22	15.01	0.26	25.23
EMA-VFI [17]	2.35	62.62	2.38	74.84	2.42	85.04

J. Computational costs of the proposed framework

a) Distance indexing: Transitioning from time indexing ($[T]$) to distance indexing ($[D]$) does not introduce extra computational costs during the inference phase, yet significantly enhancing image quality. In the training phase, the primary requirement is a one-time offline computation of distance maps for image triplets.

b) Iterative reference-based estimation: Given that the computational overhead of merely expanding the input channel, while keeping the rest of the structure unchanged, is negligible, the computational burden during the training phase remains equivalent to that of the $[D]$ model. Regarding inference, the total consumption is equal to the number of iterations $\times$ the consumption of the $[D]$ model. We would like to highlight that this iterative strategy is optional: Users can adopt this strategy at will when optimal interpolation results are demanded and the computational budget allows.

c) Continuous indexing map estimation: The runtime and the number of parameters are reported in Table XI. The estimation process introduces an additional latency of approximately 0.04 seconds per inference, which is relatively low. The number of parameters for the map estimator is 10.2M. Overall, the computational overheads introduced by indexing map estimation are acceptable in practical scenarios.

d) Multi-frame refiner: We also evaluate the computational overhead introduced by the multi-frame refiner as reported in Table XI. The refiner adds approximately 0.03 seconds per inference and requires an additional 10.3 million parameters, which is practical for deployment.

K. Limitations

While the proposed distance indexing strategy effectively stabilizes motion interpolation under large temporal gaps,

it does not explicitly address content ambiguity caused by occlusion or severe lighting variations, where visual information is partially missing or altered. The proposed distance indexing relies on distance maps estimated from optical flow. In scenarios involving severe occlusion, curved motion, or significant lighting variations, inaccuracies in optical flow estimation may lead to imperfect distance maps, which can negatively affect the learning process. While our experiments show that distance indexing remains robust to moderate estimation errors, its performance may be limited when such errors become dominant. Nevertheless, we believe that future advances in optical flow estimation will further enhance the effectiveness of distance-indexed video frame interpolation.

VII. Conclusion

We challenge the traditional time indexing paradigm and address its inherent uncertainties related to velocity distribution. Through the introduction of distance indexing and iterative reference-based estimation strategies, we offer a transformative paradigm to VFI. Our innovative plug-and-play strategies not only improves the performance in video interpolation but also empowers users with granular control over interpolation patterns across varied objects. We also propose a continuous distance map estimator to accurately predict distance maps when using multi-frame inputs. Additionally, a multi-frame refiner is integrated into the interpolation pipeline for further enhancement. While the proposed framework significantly improves both pixelwise and perceptual metrics, it still faces challenges in learning and representing diverse interpolation trajectories. Extreme large motions may require generative priors from models like diffusion models. The insights gleaned from our strategies have potential applications across a range of tasks that employ time indexing, such as space-time super-resolution, future predictions, blur interpolation and more.

Acknowledgments

This work was partially supported by the National Key R&D Program of China (No. 2022ZD0160104). We thank Dorian Chan, Zhirong Wu, and Stephen Lin for their insightful feedback and advice. Our thanks also go to Vu An Tran for developing the web application.

References

[1] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3703–3712.

[2] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[3] Y. Wu, Q. Wen, and Q. Chen, "Optimizing video prediction via video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 814–17 823.

[4] C.-Y. Wu, N. Singhal, and P. Krahenbuhl, "Video compression through image interpolation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 416–431.

[5] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4463–4471.

[6] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008.

[7] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Springer, 2022, pp. 624–642.

[8] X. Xu, L. Siyao, W. Sun, Q. Yin, and M.-H. Yang, "Quadratic video interpolation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[12] J. Luo, Z. Wan, B. Li, Y. Dai *et al.*, "Continuous parametric optical flow," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] Z. Zhong, G. Krishnan, X. Sun, Y. Qiao, S. Ma, and J. Wang, "Clearer frames, anytime: Resolving velocity ambiguity in video frame interpolation," in *European Conference on Computer Vision*. Springer, 2025, pp. 346–363.

[14] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[15] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[16] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[17] G. Zhang, Y. Zhu, H. Wang, Y. Chen, G. Wu, and L. Wang, "Extracting motion and appearance via inter-frame attention for efficient video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5682–5692.

[18] Z. Li, Z.-L. Zhu, L.-H. Han, Q. Hou, C.-L. Guo, and M.-M. Cheng, "Amt: All-pairs multi-field transforms for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9801–9810.

[19] H. Lee, T. Kim, T.-y. Chung, D. Pak, Y. Ban, and S. Lee, "Adacof: Adaptive collaboration of flows for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5316–5325.

[20] J. Park, K. Ko, C. Lee, and C.-S. Kim, "Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 109–125.

[21] J. Park, C. Lee, and C.-S. Kim, "Asymmetric bilateral motion estimation for video frame interpolation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 539–14 548.

[22] H. Sim, J. Oh, and M. Kim, "Xvfi: Extreme video frame interpolation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 489–14 498.

[23] L. Lu, R. Wu, H. Lin, J. Lu, and J. Jia, "Video frame interpolation with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3532–3542.

[24] G. Zhang, C. Liu, Y. Cui, X. Zhao, K. Ma, and L. Wang, "Vfimamba: Video frame interpolation with state space models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 225–107 248, 2024.

[25] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, C. Wang, and J. Yang, "Ifrnet: Intermediate feature refine network for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1969–1978.

[26] X. Jin, L. Wu, J. Chen, Y. Chen, J. Koo, and C.-h. Hahm, "A unified pyramid recurrent network for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1578–1587.

[27] J. Park, J. Kim, and C.-S. Kim, "Biformer: Learning bilateral motion estimation via bilateral transformer for 4k video frame interpolation,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1568–1577.
[28] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5437–5446.
[29] P. Hu, S. Niklaus, S. Sclaroff, and K. Saenko, "Many-to-many splatting for efficient video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3553–3562.
[30] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, and B. Curless, "Film: Frame interpolation for large motion," in *European Conference on Computer Vision*. Springer, 2022, pp. 250–266.
[31] M. Plack, K. M. Briedis, A. Djelouah, M. B. Hullin, M. Gross, and C. Schroers, "Frame interpolation transformer and uncertainty guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9811–9821.
[32] K. M. Briedis, A. Djelouah, R. Ortiz, M. Meyer, M. Gross, and C. Schroers, "Kernel-based frame interpolation for spatio-temporally adaptive rendering," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.
[33] D. Danier, F. Zhang, and D. Bull, "Ldmvfi: Video frame interpolation with latent diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1472–1480.
[34] X. Wang, B. Zhou, B. Curless, I. Kemelmacher-Shlizerman, A. Holynski, and S. M. Seitz, "Generative inbetweening: Adapting image-to-video models for keyframe interpolation," *arXiv preprint arXiv:2408.15239*, 2024.
[35] M. Hu, K. Jiang, Z. Zhong, Z. Wang, and Y. Zheng, "Iq-vfi: Implicit quadratic motion estimation for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6410–6419.
[36] Z. Chen, Y. Chen, J. Liu, X. Xu, V. Goel, Z. Wang, H. Shi, and X. Wang, "Videoinr: Learning video implicit neural representation for continuous space-time super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2047–2057.
[37] S. Lee, H. Lee, C. Shin, H. Son, and S. Lee, "Exploring discontinuity for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9791–9800.
[38] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 670–679.
[39] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 10 663–10 671.
[40] T. Kalluri, D. Pathak, M. Chandraker, and D. Tran, "Flavr: Flow-agnostic video representations for fast frame interpolation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2071–2082.
[41] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C. C. Loy, and Z. Liu, "Deep animation video interpolation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6587–6595.
[42] S. Chen and M. Zwicker, "Improving the perceptual quality of 2d animation interpolation," in *European Conference on Computer Vision*. Springer, 2022, pp. 271–287.
[43] W. Shen, W. Bao, G. Zhai, L. Chen, X. Min, and Z. Gao, "Blurry video frame interpolation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5114–5123.
[44] Z. Zhong, X. Sun, Z. Wu, Y. Zheng, S. Lin, and I. Sato, "Animation from blur: Multi-modal blur decomposition with motion guidance," in *European Conference on Computer Vision*. Springer, 2022, pp. 599–615.
[45] Z. Zhong, M. Cao, X. Ji, Y. Zheng, and I. Sato, "Blur interpolation transformer for real-world motion from blur," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5713–5723.
[46] B. Fan and Y. Dai, "Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4228–4237.
[47] Z. Zhong, M. Cao, X. Sun, Z. Wu, Z. Zhou, Y. Zheng, S. Lin, and I. Sato, "Bringing rolling shutter images alive with dual reversed distortion," in *European Conference on Computer Vision*. Springer, 2022, pp. 233–249.
[48] X. Ji, Z. Wang, Z. Zhong, and Y. Zheng, "Rethinking video frame interpolation from shutter mode induced degradation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 259–12 268.
[49] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time lens: Event-based video frame interpolation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 155–16 164.
[50] G. Lin, J. Han, M. Cao, Z. Zhong, and Y. Zheng, "Event-guided frame interpolation and dynamic range expansion of single rolling shutter image," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3078–3088.
[51] D. Kye, C. Roh, S. Ko, C. Eom, and J. Oh, "Acevfi: A comprehensive survey of advances in video frame interpolation," *arXiv preprint arXiv:2506.01061*, 2025.
[52] X. Cheng and Z. Chen, "Multiple video frame interpolation via enhanced deformable separable convolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7029–7045, 2021.
[53] K. Zhou, W. Li, X. Han, and J. Lu, "Exploring motion ambiguity and alignment for high-quality video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 169–22 179.
[54] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," *arXiv preprint arXiv:2304.11968*, 2023.
[55] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen, "Inpaint anything: Segment anything meets image inpainting," *arXiv preprint arXiv:2304.06790*, 2023.
[56] T. Wang, J. Zhang, J. Fei, Y. Ge, H. Zheng, Y. Tang, Z. Li, M. Gao, S. Zhao, Y. Shan *et al.*, "Caption anything: Interactive image description with diverse multimodal controls," *arXiv preprint arXiv:2305.02677*, 2023.
[57] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.
[58] Z. Shi, X. Xu, X. Liu, J. Chen, and M.-H. Yang, "Video frame interpolation transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 482–17 491.
[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[60] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.
[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
[63] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.
[64] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8121–8130.
[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[66] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.
[67] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818.

Zhihang Zhong is an Associate Professor with the School of Artificial Intelligence, Shanghai Jiao Tong University, since 2026. He has also been a Researcher at Shanghai AI Laboratory since 2023. In 2018, he received the B.E. degree in mechatronics engineering with Chu Kochen Honors from Zhejiang University. He received the M.E. degree in precision engineering and the Ph.D. degree in computer science from the University of Tokyo in 2020 and 2023, respectively. He has published papers in top-tier conferences and journals, such as CVPR, ECCV, ICCV, IJCV, *etc*. His current research interests include neural rendering, spatial agents, and computational photography.

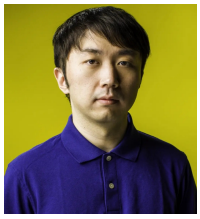
Yiming Zhang is currently a master student at Cornell University and works as an intern at Shanghai AI Laboratory. He received his B.E. degree in Electrical and Computer Engineering from Shanghai Jiao Tong University in 2023. His research interests include diffusion generative models, multi-agent intelligence, and trajectory analysis.

Wei Wang is currently a Research Engineer at Shanghai Artificial Intelligence Laboratory. His research interests include object detection, human pose estimation and 3D reconstruction. He received the B.S. from University of Chinese Academy of Sciences in 2020, and the M.S. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2023.

Xiao Sun is a scientist at Shanghai AI Laboratory, where he leads diverse R&D groups on AI for sports, healthcare, and robotics. Before that, he served as a Senior Researcher at the Visual Computing Group, Microsoft Research Asia (MSRA), from Feb. 2016 to Jul. 2022. Xiao received the B.S. and M.S. degrees in Information Engineering from South China University of Technology, China, in 2011 and 2014, respectively. His research interests include computer vision, machine learning, and computer graphics.

Yu Qiao is the lead scientist with Shanghai Artificial Intelligence Laboratory, a researcher, and the honored director of the Multimedia Laboratory with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (SIAT). He served as an assistant professor with the Graduate School of Information Science and Technology, University of Tokyo from 2009 to 2010. He has been working on deep learning since 2006, and he is one of the earliest people to introduce deep learning to video understanding. He and his team invented center loss and temporal segment networks. He has published more than 400 articles in top-tier conferences and journals in computer science with more than 80,000 citations. He received the CVPR 2023 Best Paper Award and AAAI 2021 Outstanding Paper Award. His research interests revolve around foundation models, computer vision, deep learning, robotics, and AI applications.

Gurunandan Krishnan is an SVP of OtoNexus Medical Technologies. He obtained his master's degree from Columbia University and bachelor's degree from Visvesvaraya Technological University. His research interests are in Computer Vision, Computer Graphics, and Computational Imaging. He has published papers in top-tier conferences and journals, such as Siggraph, CVPR, *etc*.



Sizhuo Ma is a Senior Research Scientist at Snap Inc. He obtained a Ph.D. degree from University of Wisconsin-Madison in 2022. He received his bachelor's degree from Shanghai Jiao Tong University. His research interests include computational imaging, computational photography and low-level vision. He has published papers at prestigious journals and conferences including CVPR, ECCV, Siggraph, MobiCom, ISMAR, *etc*.

Jian Wang is a Staff Research Scientist at Snap Inc., focusing on computational photography and imaging. He has published in top-tier venues such as CVPR, MobiCom, and SIGGRAPH, and has contributed numerous features to production. He has received Best Paper awards at SIGGRAPH Asia 2024 and the 4th IEEE International Workshop on Computational Cameras and Displays, as well as Best Poster awards at the IEEE Conference on Computational Photography 2022 and the Responsible Imaging workshop at ICCV 2025. He has served as an Area Chair for CVPR, NeurIPS, ICLR, ICML, *etc*. Jian holds a Ph.D. from Carnegie Mellon University.