LivePhoto: Real Image Animation with Text-guided Motion Control

Xi Chen[1] Zhiheng Liu[2] Mengting Chen[2] Yutong Feng[2]
Yu Liu[2] Yujun Shen[3] Hengshuang Zhao[1⋆]

[1]The University of Hong Kong [2]Alibaba Group [3]Ant Group

Abstract. Despite the recent progress in text-to-video generation, existing studies usually overlook the issue that only spatial contents but not temporal motions in synthesized videos are under the control of text. Towards such a challenge, this work presents a practical system, named `LivePhoto`, which allows users to animate an image of their interest with text descriptions. We first establish a strong baseline that helps a well-learned text-to-image generator (*i.e.*, Stable Diffusion) take an image as a further input. We then equip the improved generator with a motion module for temporal modeling and propose a carefully designed training pipeline to better link texts and motions. In particular, considering the facts that (1) text can only describe motions roughly (*e.g.*, regardless of the moving speed) and (2) text may include both content and motion descriptions, we introduce a motion intensity estimation module as well as a text re-weighting module to reduce the ambiguity of text-to-motion mapping. Empirical evidence suggests that our approach is capable of well decoding motion-related textual instructions into videos, such as actions, camera movements, or even conjuring new contents from thin air (*e.g.*, pouring water into an empty glass). Interestingly, thanks to the proposed intensity learning mechanism, our system offers users an additional control signal (*i.e.*, the motion intensity) besides text for video customization. Project page is xavierchen34.github.io/LivePhoto-Page.

Keywords: Image Animation · Image-to-Video · Text-to-Video

1 Introduction

Image and video content synthesis has become a burgeoning topic with significant attention and broad real-world applications. Fueled by the diffusion model and extensive training data, image generation has witnessed notable advancements through powerful text-to-image models [4,35,37,48] and controllable downstream applications [6, 18, 23, 24, 28, 36, 51]. In the realm of video generation, a more complex task requiring spatial and temporal modeling, text-to-video has steadily improved [2, 10, 19, 40, 49]. Various works [3, 8, 22, 43, 45] also explore enhancing controllability with sequential inputs like optical flows, motion vectors, *etc.*

⋆ Corresponding author.

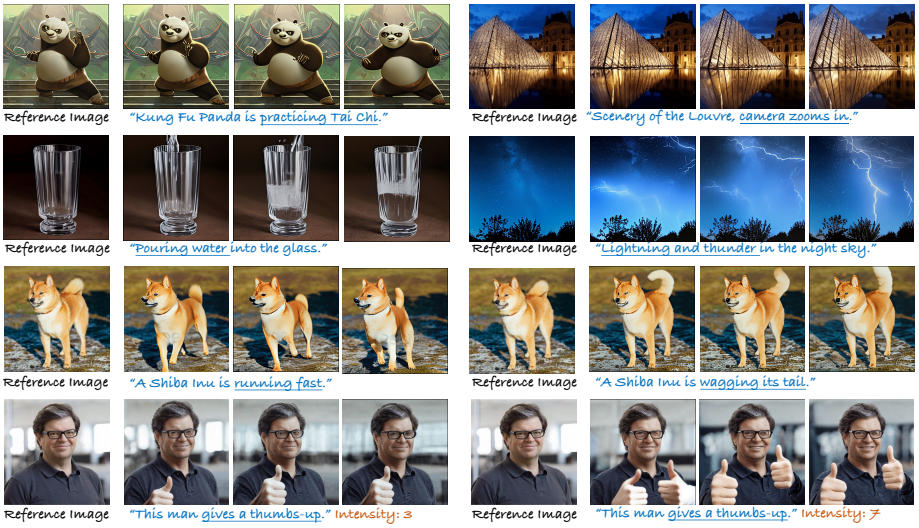

Fig. 1: Zero-shot real image animation with text control. Besides adequately decoding motion descriptions like actions and camera movements (row 1), `LivePhoto` could also conjure new contents from thin air (row 2). Meanwhile, `LivePhoto` is highly controllable, supporting users to customize the animation by inputting various texts (row 3) and adjusting the degree of motion intensity (row 4).

This work explores utilizing a real image as the initial frame to guide the "content" and employ the text to control the "motion" of the video. This topic holds promising potential for a wide range of applications, including meme generation, production advertisement, film making, *etc.* Previous image-to-video methods [5, 15, 17, 25, 41, 50, 52] mainly focus on specific subjects like humans or could only animate synthetic images. GEN-2 [34] and Pikalabs [33] animate real images with optional text input, however, an issue is that the text could only enhance the content but exerts limited control for the motions.

Facing this challenge, we propose `LivePhoto`, an image animation framework that truly listens to the text instructions. We first establish a powerful image-to-video baseline. The initial step is to equip a text-to-image model (*i,e.*, Stable Diffusion) with the ability to refer to a real image. Specifically, we concatenate the image latent with input noise to provide pixel-level guidance. In addition, a content encoder is employed to extract image patch tokens, which are injected via cross-attention to guide the global identity. During inference, a noise inversion of the reference image is introduced to offer content priors. Afterward, following the contemporary methods [2, 10, 45], we freeze stable diffusion models and insert trainable motion layers to model the inter-frame temporal relations.

Although the text branch is maintained in this strong image-to-video baseline, the model seldom listens to the text instructions. The generated videos usually remain nearly static, or sometimes exhibit overly intense movements,

deviating from the text. We identify two key issues for the problem: firstly, the text is not sufficient to describe the desired motion. Phrases like "shaking the head" or "camera zooms in" lack important information like moving speed or action magnitude. Thus, a starting frame and a text may correspond to diverse motions with varying intensities. This ambiguity leads to difficulties in linking text and motion. Facing this challenge, we parameterize the motion intensity using a single coefficient, offering a supplementary condition. This approach eases the optimization, significantly improves the motion quality, and allows users to adjust motion intensity during inference conveniently. Another issue arises from the fact that the text contains both content and motion descriptions. The content descriptions translated by stable diffusion may not perfectly align with the desired video, while the reference image is prioritized for content control. Consequently, when the content descriptions are learned to be suppressed to mitigate conflicts, motion descriptions are simultaneously under-weighted. To address this concern, we design a prompt adapter, which learns to accentuate the motion descriptions, enabling the text to work compatibly with the image for better motion control.

As shown in Fig. 1, equipped with motion intensity guidance and prompt adapter, `LivePhoto` demonstrates impressive abilities for text-guided motion control. `LivePhoto` is able to deal with real images from versatile domains and subjects, and adequately decodes the motion descriptions like actions and camera movements. Besides, it shows fantastic capacities of conjuring new contents from thin air, like "pouring water into a glass" or simulating "lightning and thunder". In addition, with motion intensity guidance, `LivePhoto` supports users to customize the motion with the desired intensity.

2 Related Work

Image animation. To realize content controllable video synthesis, image animation takes a reference image as content guidance. Most of the previous works [7, 38, 39, 54, 55] depend on another video as a source of motion, transferring the motion to the image with the same subject. Other works focus on specific categories like fluide [13, 26, 29] or nature objects [16, 21]. Make-it-Move [15] uses text control but it only manipulates simple geometries like cones and cubes. Recently, human pose transfer methods [5, 17, 42, 50] convert the human images to videos with extra controls like dense poses, depth maps, etc. VideoComposer [43] could take image and text as controls, however, the text shows limited controllability for the motion and it usually requires more controls like sketches and motion vectors. In general, existing work either requires more controls than text or focuses on a specific subject. In this work, we explore constructing a generalizable framework for universal domains and use the most flexible control (text) to customize the generated video.

Text-to-video generation. Assisted by the diffusion model [11], the field of text-to-video has progressed rapidly. Early attempts [12, 40, 49] train the entire parameters, making the task resource-intensive. Recently, researchers have

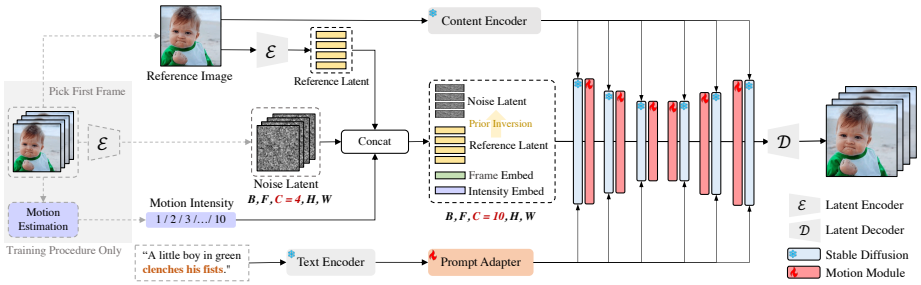

Fig. 2: Overall pipeline of `LivePhoto`. Besides taking the reference image and text as input, `LivePhoto` leverages the motion intensity as a supplementary condition. The image and the motion intensity (from level 1 to 10) are obtained from the ground truth video during training and customized by users during inference. The reference latent is first extracted as local content guidance. We concatenate it with the noise latent, a frame embedding, and the intensity embedding. This 10-channel tensor is fed into the UNet for denoising. During inference, we use the inversion of the reference latent instead of the pure Gaussian to provide content priors. At the top, a content encoder extracts the visual tokens to provide global content guidance. At the bottom, we introduce the prompt adapter, which learns to emphasize the motion-related part of the text embedding for better text-motion mapping. The visual and textual tokens are injected into the UNet via cross-attention. For the UNet, we freeze the pre-trained stable diffusion and insert motion modules to capture the inter-frame relations. Symbols of **flames** and **snowflakes** denote trainable and frozen parameters respectively.

turned to leveraging the frozen weights of pre-trained text-to-image models tapping into robust priors. Tune-A-Video [45] inflates the text-to-video model and tuning attention modules to construct an inter-frame relationship with a one-shot setting. Align-Your-Lantens [2] inserts newly designed temporal layers into frozen text-to-image models to make video generation. AnimateDiff [10] proposes to freeze the stable diffusion [35] blocks and add learnable motion modules, enabling the model to incorporate with subject-specific LoRAs [14] to make customized generation. A common issue is that the text could only control the spatial content of the video but exert limited effect for controlling the motions.

3 Method

We first give a brief introduction to the preliminary knowledge for diffusion-based image generation in Sec. 3.1. Following that, our comprehensive pipeline is outlined in Sec. 3.2. Afterward, Sec. 3.3 delves into image content guidance to make the model refer to the image. In Sec. 3.4 and Sec. 3.5, we elaborate on the novel designs of motion intensity guidance and prompt adapter to better align the text conditions with the video motion.

3.1 Preliminaries

Text-to-image with diffusion models. Diffusion models [11] show promising abilities for both image and video generation. In this work, we opt for the widely used Stable Diffusion [35] as the base model, which adapts the denoising procedure in the latent space with lower computations. It initially employs VQ-VAE [20] as the latent encoder to transform an image $\mathbf{x}_0$ into the latent space: $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. During training, Stable Diffusion transforms the latent into Gaussian noise as follows:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where the noise $\epsilon \sim \mathcal{U}([0,1])$, and $\bar{\alpha}_t$ is a cumulative products of the noise coefficient α_t at each step. Afterward, it learns to predict the added noise as:

$$\mathbb{E}_{\mathbf{z},\mathbf{c},\epsilon,t}(\|\epsilon_\theta(\mathbf{z}_t,\mathbf{c},t) - \epsilon\|_2^2). \tag{2}$$

t is the diffusion timestep, $\mathbf{c}$ is the condition of text prompts. During inference, Stable Diffusion is able to recover an image from Gaussian noise step by step by predicting the noise added for each step. The denoising results are fed into a latent decoder to recover the colored images from latent representations as $\hat{\mathbf{x}}_0 = \mathcal{D}(\hat{\mathbf{z}}_0)$.

3.2 Overall Pipeline

The framework of `LivePhoto` is demonstrated in Fig. 2. The model takes a reference image, a text, and the motion intensity as input to synthesize the desired video. When the ground truth video is provided during training, the reference image is picked from the first frame, and the motion intensity is estimated from the video. During inference, users could customize the motion intensity or directly use the default level. `LivePhoto` utilizes a 4-channel tensor of $\mathbf{z}^{B\times F\times C\times H\times W}$ to represent the noise latent of the video, where the dimensions mean batch, frame, channel, height, and width, respectively. The reference latent is extracted by VAE encoder [20] to provide local content guidance. Meanwhile, the motion intensity is transformed to a 1-channel intensity embedding. We concatenate the noise latent, the reference latent, the intensity embedding, and a frame embedding to form a 10-channel tensor for the input of UNet. At the same time, we use a content encoder to extract the visual tokens of the reference image and inject them via cross-attention. A prompt adapter is added after the text encoder [32], which learns to assign different weights to each part of the text to accentuate the motion descriptions of the text. Following modern text-to-video models [2, 10]. We freeze the stable diffusion [35] blocks and add learnable motion modules [10] at each stage to capture the inter-frame relationships. To generate high-quality results, we follow previous works [40, 52] to add a spatial upscaler to recover the fine details.

3.3 Image Content Guidance

The most essential step is enabling `LivePhoto` to keep the identity of the reference image. Thus, we collect local guidance by concatenating the reference latent at the input. Moreover, we employ a content encoder to extract image tokens for global guidance. Additionally, we introduce the image inversion in the initial noise to offer content priors.

Reference latent. We extract the reference latent and incorporate it at the UNet input to provide pixel-level guidance. Simultaneously, a frame embedding is introduced to impart temporal awareness to each frame. Thus, the first frame could totally trust the reference latent as it is required to re-generate the given reference image. Subsequent frames make degenerative references and exhibit distinct behavior. The frame embedding is represented as a 1-channel map, with values linearly interpolated from zero (first frame) to one (last frame).

Content encoder. The reference latent effectively guides the initial frames due to their higher pixel similarities. However, as content evolves in subsequent frames, understanding the image and providing high-level guidance becomes crucial. Drawing inspiration from [6], we employ a frozen DINOv2 [30] to extract patch tokens from the reference image. We add a learnable linear layer after DINOv2 to project these tokens, which are then injected into the UNet through newly added cross-attention layers.

Prior inversion. Previous methods [19, 25, 27, 41, 45] prove that using an inverted noise of the reference image, rather than a pure Gaussian noise, could effectively provide appearance priors. During inference, we add the inversion of the reference latent $\mathbf{r}_0$ to the noise latent $\mathbf{z}_T^n$ of frame n at the initial denoising step (T), following Eq. (3).

$$\tilde{\mathbf{z}}_T^n = \alpha^n \cdot \mathrm{Inv}(\mathbf{r}_0) + (1 - \alpha^n) \cdot \mathbf{z}_T^n, \tag{3}$$

where α^n is a descending coefficient from the first frame to the last frame. We set α^n as a linear interpolation from 0.033 to 0.016 by default.

3.4 Motion Intensity Guidance

It is challenging to align the motion coherently with the text. We analyze the core issue is that the text lacks descriptions for the motion speed and magnitude. Thus, the same text leads to various motion intensities, creating ambiguity in the optimization process. To address this, we leverage the motion intensity as an additional condition. We parameterize the motion intensity using a single coefficient. Thus, the users could adjust the intensity conveniently by sliding a bar or directly using the default value.

In our pursuit of parameterizing motion intensity, we experimented with various methods, such as calculating optical flow magnitude, computing mean square error between adjacent frames, and leveraging CLIP/DINO similarity between frames. Ultimately, we found that Structural Similarity (SSIM) [44] produces results the most aligned with human perceptions. Concretely, given

a training video clip $\mathbf{X}^n$ with n frames, we determine its motion intensity $\mathbf{I}$ by computing the average value for the SSIM [44] between each adjacent frame. The structure similarity considers the luminance, contrast, and structure differences between two images.

We compute the motion intensity on the training data to determine the overall distribution and categorize the values into 10 levels. We create a 1-channel map filled with the level numbers and concatenate it with the input of UNet. During inference, users can utilize level 5 as the default intensity or adjust it between levels 1 to 10. Throughout this paper, unless specified, we use level 5 as the default. The motion intensity guidance simplifies the learning of text-motion alignment and brings significantly better motion quality.

3.5 Prompt Adapter

Another challenge in instructing video motions arises from the fact that the text prompt encompasses both "content descriptions" and "motion descriptions". The "content descriptions", translated by the frozen Stable Diffusion, often fail to perfectly align with the ground truth videos. The text descriptions could only guide the semantic and color consistency, but hard to depict the identity and fine-grained layouts. This issue commonly exists in the training data. Considering that the text-to-image model is frozen, these content conflicts could not be solved by training. In image-to-video generation, as the reference image provides more consistent content guidance compared with the text, the model learns to trust the reference image. Thus, the whole text tends to be overlooked.

The problem of content conflicts also exists in text-to-video generation, freezing the text-to-image modules causes inferior performance. To alleviate the conflict, AnimateDiff adds a domain adapter in the frozen Stable Diffusion. This solution eases the optimzation, but the trained model always requires an additional Lora to control the content.

Instead of adding adapters in the frozen Stable Diffusion module, we explore manipulating the CLIP text embeddings to suppress the content description and accentuate the motion descriptions. Recognizing that directly tuning the text encoder on limited samples might impact generalization, we investigate adjusting the weights of each embedding without disrupting the CLIP feature space. Concretely, we add three trainable transformer layers and a linear projection layer after the CLIP text embeddings. Afterward, the predicted weights are normed from 0 to 1 with a sigmoid function. These weights are then multiplied with the corresponding text embeddings, thereby providing guidance that focuses on directing the motions. The comprehensive structure of the prompt adapter and actual examples are depicted in Fig. 3. The numerical results prove that the module successfully learns to emphasize the "motion descriptions". This allows signals from images and texts to integrate more effectively, resulting in stronger text-to-motion control.

Fig. 3: Demonstrations for the structure of prompt adapter. We use three transformer encoder layers and a frame-specific linear layer to predict the weight of each text token. Examples are given on the right. In cases where multiple tokens correspond to a single word, we calculate the average weight for better visualization. The words with the maximum weight are underlined.

4 Experiments

4.1 Implementation Details

Detailed configurations. We implement `LivePhoto` based on the frozen Stable Diffusion v1.5 [35]. The structure of our Motion Module aligns with AnimateDiff [10]. Our model is trained on the WebVID [1] dataset employing 8 A100 GPUs. We sample training videos with 16 frames, perform center-cropping, and resize each frame to 256×256 pixels. A commonly used MSE loss is leveraged to train the model. During inference, the default output resolution aligns with previous works [43,46] as 256×256, we also provide an option for users to upscale the video into 1024×1024 using the $4\times$upscaler provided by Stable Diffusion.
Evaluation protocols. We conduct user studies to compare our approach with previous methods and analyze our newly designed modules. To validate the generalization ability, we gather images from various domains encompassing real images and cartoons including humans, animals, still objects, natural sceneries, *etc.* For quantitative assessment, we utilize the validation set of WebVID [1] and MSR-VTT [47]. The first frame and prompt are used as controls to generate videos. Following image customization methods [6, 9, 36], we assess the ID-preserving ability using DINO similarities between the reference frames and each generated frame. Motion intensity is evaluated using SSIM as introduced in Sec. 3.4. We measure the average CLIP similarity [32] between adjacent frames to evaluate the frame consistency following previous works [8, 43].

4.2 Ablation Studies

In this section, we thoroughly analyze each of our proposed modules to substantiate their effectiveness. We first analyze how to add content guidance with the reference image, which is an essential part of our framework. Following that, we delve into the specifics of our newly introduced motion intensity guidance and prompt adapter.

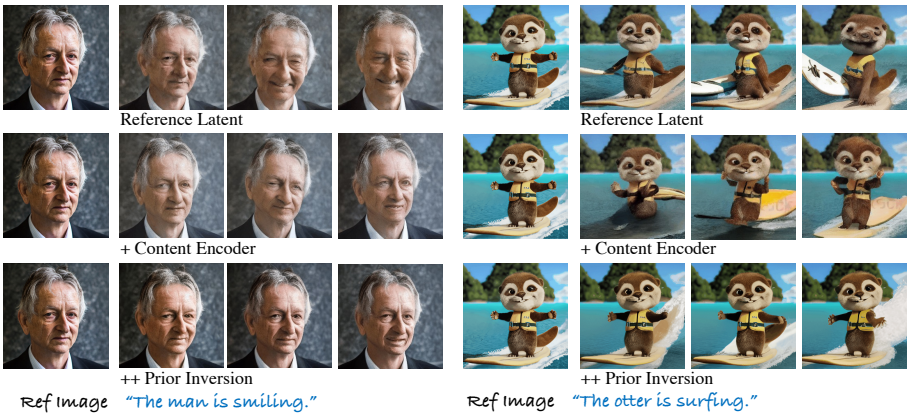

Fig. 4: Ablations for the image content guidance. Only concatenating the reference latent with the model input meets challenges in preserving the identity. The content encoder and prior inversion gradually enhance the performance.

Table 1: Quatitative analysis for image content guidance on WebVID [1]. We assess the ID preservation ability, the motion intensity, and the frame consistency. Each module brings improvements in ID-preserving and frame consistency. The motion intensity decreases because of the suppression of collapse and distortions.

Method	ID Preservation	Motion Intensity	Frame Consistency
Reference Latent	64.1	**47.8**	91.7
+ Content Encoder	77.0	29.5	93.2
++ Prior Inversion	**81.2**	24.4	**95.2**

Image content guidance. As introduced in Sec. 3.2, we concatenate the reference latent with the input as the pixel-wise guidance and use a content encoder to provide the holistic identity information. Besides, the prior inversion further assists the generation of details. In Fig. 4, we illustrate the step-by-step integration of these elements. In row 1, the reference latent could only keep the identity for the starting frames as the contents are similar to the reference image. After adding the content encoder in row 2, the identity for the subsequent frames could be better preserved but the generation quality for the details is not satisfactory. With the inclusion of prior inversion, the overall quality sees further improvement. The quantitative results in Tab. 1 consistently confirm the effectiveness of each module. The motion intensity decreases because of fewer collapses and distortions. These three strategies serve as the core of our strong baseline for real image animation.

Motion intensity guidance. As introduced in Sec. 3.4, we parameterize the motion intensity as a coefficient, and use it to indicate the motion speed and ranges. We carry out ablation studies in Fig. 5. The absence of motion intensity guidance often leads to static or erratic video outputs, as depicted in the first row. However, with the introduction of intensity guidance, the subsequent rows

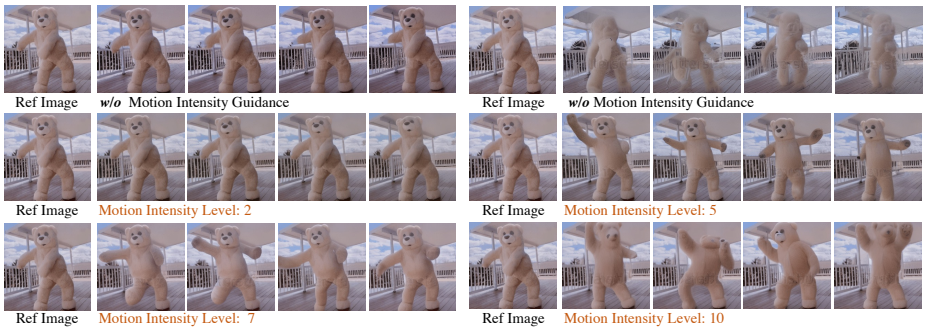

Fig. 5: Illustrations of motion intensity guidance. The prompt is **"The bear is dancing".** Without intensity guidance, the generated video tends to either keep still or quickly become blurry. With the option to set varying intensity levels, users can finely control the motion range and speed. It should be noted that excessively high-intensity levels might induce motion blur, as observed in the last case.

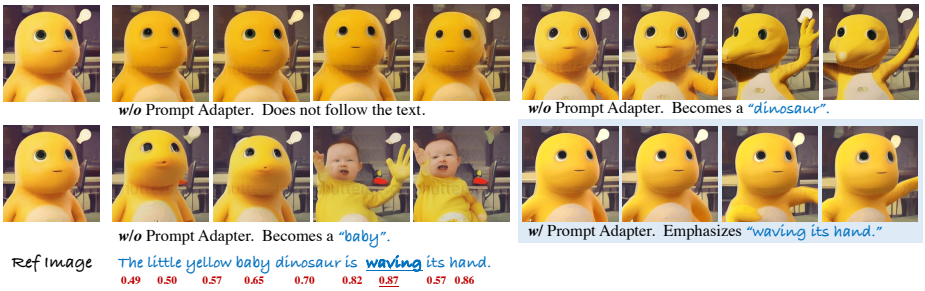

Fig. 6: Ablation for prompt adapter. Without the prompt adapter, the model tends to either disregard the text entirely or fixate on content-related descriptions like "baby dinosaur". When re-weighting is applied, content descriptions are suppressed while motion-related details like "waving its hand" gain emphasis. The predicted weights of the prompt adapter are marked at the bottom.

display varying motion levels, allowing for the production of high-quality videos with different motion ranges. Notably, lower levels like level 2 generate almost static videos, while higher levels like 10 occasionally produce overly vigorous motions. Users could directly use the default value (level 5) or tailor the intensity according to specific preferences. The motion intensity guidance eases the learning of motion, significantly improves the motion quality, and thus sets a basis for aligning the motion with the text instructions.

Prompt adapter. In Fig. 6, we demonstrate the efficacy of prompt adapter. In the given examples, the content description "baby dinosaur" would conflict with the desired content depicted by the reference image. In the first three rows, without the assistance of prompt adapter, the frozen Stabel Diffusion tends to synthesize the content through its understanding of the text. Thus, the produced

Table 2: Quatitative analysis for novel components on WebVID [1]. We evaluate the ability of ID preservation, motion intensity, and frame consistency. The results show that our newly designed modules bring consistent improvements.

Method	ID-Preservation	Motion Intensity	Frame Consistency
Baseline (*w/* Image Content Guidance)	75.1	15.3	93.1
Baseline + Prompt Adapter	77.5	18.8	94.8
Baseline + Motion Intensity Guidance	79.1	21.4	93.9
LivePhoto-Full	**81.2**	**24.7**	**95.2**

Table 3: Qualitative comparisons with previous works on MSR-VTT [47]. `LivePhoto` demonstrates superior results compared with previous methods. We leverage an extended version of AnimateDiff [41] supporting image-to-video.

Method	FVD (↓)	ID-Preservation	Motion Intensity	Frame Consistency
AnimateDiff* [41]	687	65.4	55.4	81.6
VideoComposer	356	73.2	25.8	90.5
LivePhoto	**289**	**80.2**	26.7	**93.1**

video tends to ignore the text and follow the reference image as in row 1. In other cases, it has risks of becoming a "baby" (row 2) or a "dinosaur" (row 3). As visualized in the bottom of Fig. 6, the prompt adapter elevates emphasis on motion descriptions like "waving its hand". Prompt adapter enables our model to faithfully follow text-based instructions for motion details while upholding image-consistent content with the reference image.

Quantitative results. The numerical results are listed in Tab. 2. As the motion intensity guidance significantly increases the motion quality, reducing the rate of collapse and distortion, it brings consistent gains for all three metrics. The prompt adapter contributes to a more precise prompt-following ability. It also leads to all-sided improvements.

4.3 Comparisons with Existing Alternatives

We compare `LivePhoto` with other works that support image animation with text control. VideoComposer [43] is a strong compositional generator covering various conditions including image and text. GEN-2 [34] and Pikalabs [33] are famous products that support image and text input. AnimateDiff-I2V [25] and Talesofai [41] are open-source projects claiming similar abilities. I2VGEN-XL [52], PIA [53], and DynamiCrafter [46] are concurrent works.

Qualitative analysis. In Fig. 7, we compare `LivePhoto` with VideoComposer [43], Pikalabs [33], and GEN-2 [34] with representative examples. The selected examples cover animals, humans, cartoons, and natural scenarios. To reduce the randomness, we ran each method 8 times to select the best result for more fair comparisons. VideoComposer demonstrates proficiency in creating videos with significant motion. However, as not specifically designed for photo animation, the identity-keeping ability is not satisfactory. The identities of the reference images are lost, especially for less commonly seen subjects. Additionally, it shows a

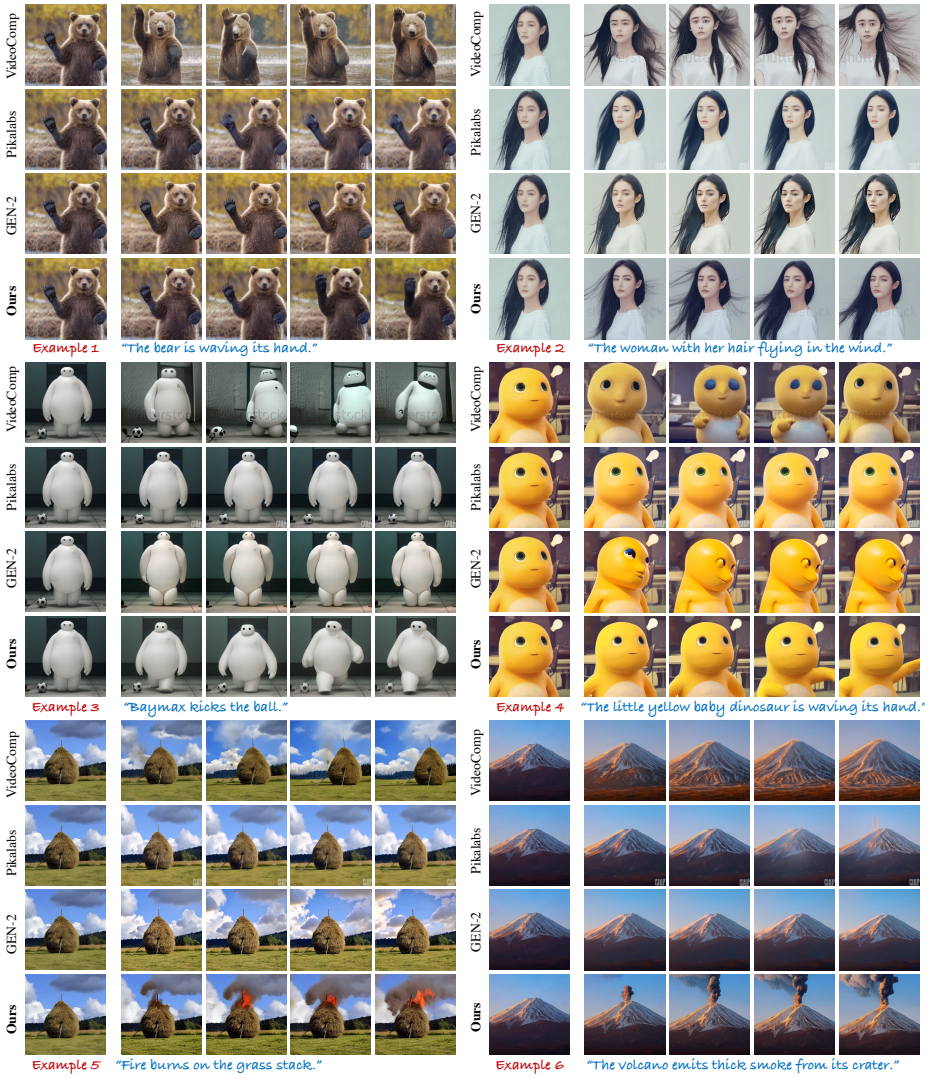

Fig. 7: Comparison results with other methods. We compare our `LivePhoto` with VideoComposer [43], Pikalabs [33], and GEN-2 [34]. We select representative cases covering animal, human, cartoon, and natural scenery. To ensure a fair evaluation, we executed each method 8 times, presenting the most optimal outcomes for comparison. In each example, the reference image is displayed on the left, accompanied by the text prompt indicated at the bottom. `LivePhoto` demonstrates a superior prompt-following ability while finely preserving the identity.

lack of adherence to the provided text instructions. Pikalabs [33] and GEN-2 [34] produce high-quality videos. However, as a trade-off, the generated videos own limited motion ranges. Although they support text as supplementary, the

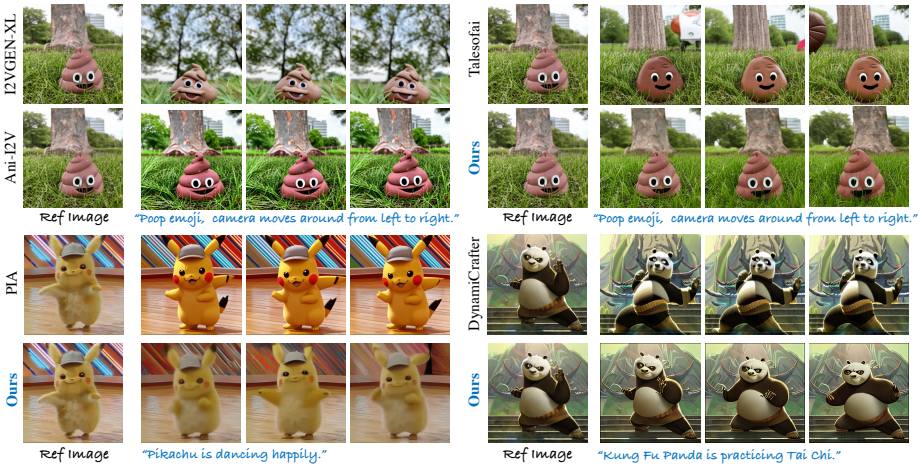

Fig. 8: Comparisons with concurrent works or open-source projects. I2VGEN-XL [52], AnimateDiff-I2V [25], and Talesofai [41] also support image animation. I2VGEN-XL generates "relevant" content with the reference image. The produced videos of AnimateDiff-I2V rarely move. Talesofai struggles for real photos. PIA [53] and DynamiCrafter [46] also struggle for keeping the consistent identities.

text descriptions seldom work. The motions are generally estimated from the content of the reference image. In contrast, `LivePhoto` adeptly preserves the identity of the reference image and generates consistent motions with the text instructions. It performs admirably across various domains, encompassing animals, humans, cartoon characters, and natural sceneries. It not only animates specific actions (examples 1-4) but also conjures new effects from thin air (examples 5-6).

We also compare `LivePhoto` with the open-sourced project and concurrent works in Fig. 8. I2VGEN-XL [52], Talsofai [41], PIA [53], and DynamiCrafter [46] suffer from the limited ID-preservation ability. AnimateDiff-I2V [25] keeps the image identity but exhibits very limited motion intensity.

User studies. The qualitative metrics have limitations in thoroughly evaluating the model, especially for the motion quality and prompt-following ability.Thus, we carry out user studies. We ask the annotators to rate the generated videos from 4 perspectives: Text consistency measures whether the motion follows the text descriptions. Motion quality assesses the reasonableness of generated motion, encompassing aspects such as speed and deformation. Image consistency evaluates the identity-keeping ability of the reference image. Content quality considers the general quality of videos like the smoothness, the resolution, *etc.*

We construct a benchmark with five tracks: humans, animals, cartoon characters, still objects, and natural sceneries. We collect 10 reference images per track and manually write 2 prompts per image. Considering the variations that commonly exist in video generation, each method is required to predict 8 results. Thus, we get 800 samples for each method. We ask 10 annotators to rate the predictions according to the aforementioned four perspectives. We compare

Table 4: Results of user study. We let annotators rate from four perspectives: Text consistency ($\mathbf{C}_{\text{text}}$) measures the adherence to the text prompt in directing motion. Motion quality ($\mathbf{Q}_{\text{mot}}$) evaluates appropriateness of motions. Image consistency ($\mathbf{C}_{\text{image}}$) evaluates the capability to maintain the identity of the reference image. Content quality ($\mathbf{Q}_{\text{cont}}$) focuses on the inter-frame coherence and resolutions.

	$\mathbf{C}_{\text{text}}$ (↑)	$\mathbf{Q}_{\text{mot}}$ (↑)	$\mathbf{C}_{\text{image}}$ (↑)	$\mathbf{Q}_{\text{cont}}$ (↑)
VideoComposr [43]	2.9	3.1	2.5	3.1
Pikalabs [33]	2.5	2.9	**3.8**	4.3
GEN-2 [34]	2.4	3.0	3.6	**4.6**
Baseline (*w/* Image Content Guidance)	2.0	2.5	2.9	3.5
+ Motion Intensity Guidance	3.2	3.3	3.1	3.7
++ Prompt Adapter	**3.9**	**3.5**	3.3	3.8

`LivePhoto` with VideoComposer [43], GEN-2 [34], and Pikalabs [33]. To compare GEN-2 and Pika, we leverage an upscaled version of LivePhoto as introduced in Sec. 4.1 with high-resolution outputs.

Results are reported in Tab. 4. Compared with VideoComposer [43], our I2V-Baseline shows better ID-preserving ability ($\mathbf{C}_{\text{image}}$) and inferior prompt-following ability ($\mathbf{C}_{\text{text}}$) as a trade-off. From this baseline, the motion intensity guidance and prompt adapter bring steady improvements for the prompt-following ability and generation quality. At the same time, with improvements in motion, the collapse and distortion cases decrease. Thus the image-related quality ($\mathbf{C}_{\text{image}}$, $\mathbf{Q}_{\text{content}}$) also improves.

GEN-2 and Pika are commercial products that investigate more training data and larger models. Compared with them, `LivePhoto` shows significantly better text consistency and motion quality. GEN-2 and Pika show better ID-preservation ability ($\mathbf{C}_{\text{image}}$), however, their generated video seldom moves as a trade-off. We admit that GEN-2 and Pikalabs own superior smoothness and resolution. We infer that they might collect much better training data and training with higher resolutions. However, as an academic method, `LivePhoto` shows distinguishing advantages over mature products in certain aspects. We have reasons to believe its potential for future applications.

5 Conclusion

We introduce `LivePhoto`, a novel framework for photo animation with text control. We propose a strong baseline that gathers the image content guidance from the given image and utilizes motion intensity as a supplementary to better capture the desired motions. Besides, we propose a prompt adapter to accentuate the motion descriptions. The whole pipeline illustrates impressive performance.
Limitation and potential effects. `LivePhoto` is implemented on SD-1.5 and trained limited data (WebVID [1]). We believe that with more training data and stronger models like SD-XL [31] or even transformer-based larger diffusion models, the overall performance could be further improved significantly.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 62201484), HKU Startup Fund, and HKU Seed Fund for Basic Research.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV (2021) 8, 9, 11, 14
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023) 1, 2, 4, 5
3. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: ICCV (2023) 1
4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv:2310.00426 (2023) 1
5. Chen, T.S., Lin, C.H., Tseng, H.Y., Lin, T.Y., Yang, M.H.: Motion-conditioned diffusion model for controllable video synthesis. arXiv:2304.14404 (2023) 2, 3
6. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv:2307.09481 (2023) 1, 6, 8
7. Cheng, C.C., Chen, H.Y., Chiu, W.C.: Time flies: Animating a still image with time-lapse video as reference. In: CVPR (2020) 3
8. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: ICCV (2023) 1, 8
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv:2208.01618 (2022) 8
10. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv:2307.04725 (2023) 1, 2, 4, 5, 8
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020) 3, 5
12. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022) 3
13. Holynski, A., Curless, B.L., Seitz, S.M., Szeliski, R.: Animating pictures with eulerian motion fields. In: CVPR (2021) 3
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv:2106.09685 (2021) 4
15. Hu, Y., Luo, C., Chen, Z.: Make it move: controllable image-to-video generation with text descriptions. In: CVPR (2022) 2, 3
16. Jhou, W.C., Cheng, W.H.: Animating still landscape photographs through cloud motion creation. TMM (2015) 3
17. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. arXiv:2304.06025 (2023) 2, 3
18. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023) 1

19. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv:2303.13439 (2023) 1, 6
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013) 5
21. Li, Z., Tucker, R., Snavely, N., Holynski, A.: Generative image dynamics. arXiv:2309.07906 (2023) 3
22. Liew, J.H., Yan, H., Zhang, J., Xu, Z., Feng, J.: Magicedit: High-fidelity and temporally coherent video editing. arXiv:2308.14749 (2023) 1
23. Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones: Concept neurons in diffusion models for customized generation. arXiv:2303.05125 (2023) 1
24. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones 2: Customizable image synthesis with multiple subjects. arXiv:2305.19327 (2023) 1
25. Luan, T.: Animatediff-i2v. https://github.com/ykk648/AnimateDiff-I2V (2023) 2, 6, 11, 13
26. Mahapatra, A., Kulkarni, K.: Controllable animation of fluid elements in still images. In: CVPR (2022) 3
27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv:2108.01073 (2021) 6
28. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv:2302.08453 (2023) 1
29. Okabe, M., Anjyo, K., Igarashi, T., Seidel, H.P.: Animating pictures of fluid using video examples. In: Computer Graphics Forum. Wiley Online Library (2009) 3
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv:2304.07193 (2023) 6
31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952 (2023) 14
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 5, 8
33. reseachers, P.: Pikalabs: An innovative text-to-video platform. https://www.pika.art/ (202310) 2, 11, 12, 14
34. reseachers, R.: Gen-2: The next step forward for generative ai. https://research.runwayml.com/gen2 (202310) 2, 11, 12, 14
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 1, 4, 5, 8
36. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023) 1, 8
37. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022) 1
38. Shalev, Y., Wolf, L.: Image animation with perturbed masks. In: CVPR (2022) 3
39. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. NeurIPS (2019) 3

40. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv:2209.14792 (2022) 1, 3, 5
41. talesofai: Animatediff talesofai. https://github.com/talesofai/AnimateDiff (2023) 2, 6, 11, 13
42. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv:2307.00040 (2023) 3
43. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. NeurIPS (2023) 1, 3, 8, 11, 12, 14
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004) 6, 7
45. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV (2023) 1, 2, 4, 6
46. Xing, J., Xia, M., Zhang, Y., Chen, H., Wang, X., Wong, T.T., Shan, Y.: Dynamicrafter: Animating open-domain images with video diffusion priors. arXiv:2310.12190 (2023) 8, 11, 13
47. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR. pp. 5288–5296 (2016) 8, 11
48. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. NeurIPS (2023) 1
49. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. arXiv:2303.12346 (2023) 1, 3
50. Zhang, J., Yan, H., Xu, Z., Feng, J., Liew, J.H.: Magicavatar: Multimodal avatar generation and animation. arXiv:2308.14748 (2023) 2, 3
51. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv:2302.05543 (2023) 1
52. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv:2311.04145 (2023) 2, 5, 11, 13
53. Zhang, Y., Xing, Z., Zeng, Y., Fang, Y., Chen, K.: Pia: Your personalized image animator via plug-and-play modules in text-to-image models. CVPR (2023) 11, 13
54. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: CVPR (2022) 3
55. Zhao, R., Wu, T., Guo, G.: Sparse to dense motion transfer for face image animation. In: ICCV (2021) 3