Pix2Gif: Motion-Guided Diffusion for GIF Generation

Hitesh Kandala[1], Jianfeng Gao[1], and Jianwei Yang[1]

Microsoft Research
khitesh2000@gmail.com,{jfgao,jianwei.yang}@microsoft.com

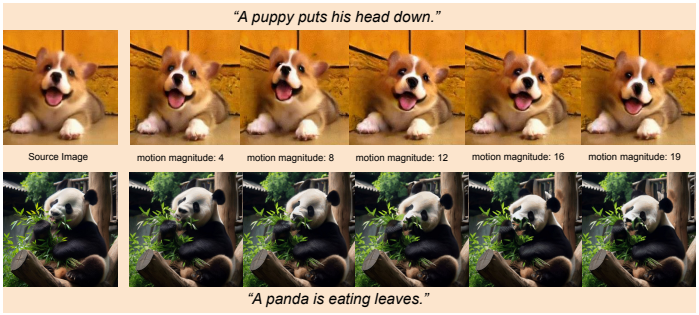


Fig. 1: Our model creates distinct frames based on the provided source image and caption, adjusting according to different levels of motion magnitude (optical flow magnitude) specified in the input conditions. It stands well for both high spatial quality and temporal consistency.

Abstract. We present *Pix2Gif*, a motion-guided diffusion model for image-to-GIF (video) generation. We tackle this problem differently by formulating the task as an image translation problem steered by text and motion magnitude prompts, as shown in Fig. 1. To ensure that the model adheres to motion guidance, we propose a new motion-guided warping module to spatially transform the features of the source image conditioned on the two types of prompts. Furthermore, we introduce a perceptual loss to ensure the transformed feature map remains within the same space as the target image, ensuring content consistency and coherence. In preparation for the model training, we meticulously curated data by extracting coherent image frames from the TGIF video-caption dataset, which provides rich information about the temporal changes of subjects. After pretraining, we apply our model in a zero-shot manner to a number of video datasets. Extensive qualitative and quantitative experiments demonstrate the effectiveness of our model – it not only captures the semantic prompt from text but also the spatial ones from motion guidance. We train all our models using a single node of 16×V100 GPUs.

Keywords: Diffusion · image-to-video generation · image-to-image translation

1 Introduction

Visual content generation has been significantly advanced by the huge progress of diffusion models [21, 22, 35, 51]. Recently, the development of latent diffusion models (LDMs) [45] has led us to a new quality level of generated images. It has inspired a lot of works for customized and controllable image generation [32, 46, 65, 69], and fine-grained image editing [6, 20, 28, 36].

In this work, we focus on converting a single image to an animated Graphics Interchange Format (GIF), which is valuable for design yet under-explored. Despite the absence of image-to-GIF generation models, diffusion-based video generation has emerged as a hot topic recently. Compared with text-to-image generation, however, text-to-video generation requires not only high quality for individual frames but also visual consistency and temporal coherence across frames. To achieve this goal, existing works expand the LDMs to video diffusion models (VDMs) by either inflating the 2D CNNs in LDMs to 3D ones [23] or introducing an additional temporal attention layer to bridge the diffusion for each frame [5, 15, 19, 50, 59]. In addition to text prompts, a few recent works also explored the way of using images or other prompts to make the video generation model more customizable and controllable [11, 38]. However, due to the high cost of VDMs to generate a sequence of video frames in one run, most (if not all) of these works require a compromise of reducing the resolution of generated frames (64×64 typically), and the usage of extra super-resolution diffusion models for upscaling [31, 47]. Moreover, since these methods use the temporal attention layers to model the cross-frame dependency implicitly, it is quite hard for them to preserve good controllability of the frame-to-frame temporal dynamics in a fine-grained manner.

Given that animated GIF usually contains less number of frames and requires more specializations, we take a different strategy and formulate the image-to-GIF generation as an image translation process. To decouple the generation of visual contents and temporal dynamics, we further introduce a motion flow magnitude as extra guidance in addition to image and text prompts. Unlike the aforementioned works, our model takes one or more history frames as the condition and produces only one future frame at once. This brings some unique advantages: (*i*) *simplicity* - our model can be purely built on top of LDMs and trained end-to-end with high resolution, without any cascaded diffusion processes for upscaling. (*ii*) *controllability* - we could inject detailed and different text and motion prompts at each time step for generating a frame, which gains much better controllability of the model. Our work is inspired by a line of canonical works for future frame prediction [40, 48, 56]. However, due to the lack of a powerful image-generation engine, these works fail to produce high-quality results and can only be applied to specific video domains [16, 30, 52]. Moreover, they cannot support other types of prompts or conditions than the history frames. To address this problem,

we exploit a modern diffusion-based pipeline. More specifically, we follow text-conditioned image editing approaches (*e.g.*, InstructPix2Pix [6]), and propose a new temporal image editing to produce future frames given history frames. To train the model, we curate a new training dataset based on TGIF [33] by extracting frames and calculating the magnitude of optical flow between them. We then selected an appropriate range of the optical flow magnitude and sampled frame pairs from each GIF in a manner that ensures diversity. In the end, we train our diffusion model called *Pix2Gif*, which can generate high-quality animated GIF consisting of multiple frames, given a single image and text and motion magnitude prompts. In summary, our main contributions are:

- We are the first to explore an image-to-image translation formula for generating animated GIFs from an image, guided by a text prompt and motion magnitude.
- We propose a flow-based warping module with a perceptual loss in the diffusion process that takes motion magnitude as input and controls the temporal dynamics and consistency between future frames and the initial ones.
- We curate a new dataset, comprised of 78,692 short GIF clips for training, and 10,546 for evaluation. The new dataset covers a variety of visual domains.
- Quantitative and qualitative results demonstrate the effectiveness of our proposed method for generating visually consistent coherent GIFs from a single image, and it can be generalized to a wide range of visual domains.

2 Prior Work

Image and video generation has been a long-standing problem in the community. It can be tackled by different approaches, which can be categorized into four groups: generative adversarial networks (GANs) [17,26,27,44], transformer-based autoregressive decoding [10,12,41,42,60,66], masked image modeling [7,8,55,67]. Most of the recent works exploited diffusion models for image generation given their high-quality outputs and huge open-source supports [45, 47]. Recently, a number of works have extended the text-to-image generation model into image translation or editing models [6, 20, 28, 36, 69] or video generation models [5, 15, 23,50,57]. Below we provide a brief overview of the related diffusion-based image and video generation methods.

Image-to-Image Translation. Diffusion-based image-to-image generation has drawn increasing attention. Different from text-to-image generation, it takes an image as input and edits its contents following the text instructions while keeping the irrelated parts unchanged. SDEdit [36] and ILVR [9] are two pioneering works that impose reference image conditions to an existing latent diffusion model for controllable image generation. Later on, to conduct local edits, the authors in [2] proposed blended latent diffusion to steer the diffusion process with a user-specified mask, where the pixels out of the mask remain the same

as the input image while the region inside is edited following the textual description. Instead of manipulating the image space, Prompt2Prompt [20] proposed to edit the image by manipulating the textual context (*e.g.*, swapping or adding words.) to which the latent diffusion model cross-attends. However, this method requires forwarding a text-to-image generation process to obtain the cross-attention maps, and thus cannot be applied to real images. Imagic [28] proposed to blend the embeddings of a real image with the textual context embedding so that the generated image obeys both the image and text conditions.

All the aforementioned works leverage a frozen latent diffusion model and control the generation with modified text or image prompts. To enable arbitrary image editing, InstructPix2Pix [6] proposed to finetune the LDM to follow user instructions that precisely convey the user intents, *e.g.*, "change the cat to dog". The model is trained by a synthetic dataset consisting of triplets $\langle image_{src}, instruction, image_{tgt} \rangle$. The resulting model could allow both realistic and generated images and support arbitrary language instructions. Some other works also exploit a similar way to train the model to follow instructions [18,68]. To further enhance the language understanding, MGIE [13] exploited a large multimodal model to produce a more comprehensive textual context for the instructed image editing.

In this work, we employ the image-to-image translation pipeline and are the first to formulate a GIF generation as an image translation problem. Given a reference image, the goal is to generate a realistic *future* frame following a textual instruction. Therefore, the focus is on how to perform temporal rather than spatial editing on a source image. When the process rolls out, it gradually gives a sequence of frames.

Conditioned Video Generation. Speaking of the high-level goal, our work resembles conditioned video generation. For video generation, a conventional way is inflating the 2D U-Net used in LDM to 3D U-Net [71] by replacing the 2D convolution layers with 3D ones. Likewise, a similar strategy is taken in [5,15,19,50,59], but with a slight difference in that they use interleaved spatial and temporal attention layers in the U-Net. Due to the high cost of generating a sequence of video frames in one shot, the output videos usually have a resolution as low as 64×64. To attain high-resolution videos, these methods need to use one or more super-resolution diffusion models [31, 47] to upscale the resolution by 4 or 8 times. To accelerate the training, a pre-trained text-to-image LDM is usually used to initialize these models. Adding spatial-temporal modules is also a commonly used strategy for autoregressive models [24, 55, 61, 62]. Similarly, both [62] and [24] exploit a pretrained autoregressive image generation model as the starting point. In [55], however, the authors introduce and pretrained a new video encoder, which is then used to train a masked video decoder.

Besides text-to-video generation, using images as the condition for video generation draws increasing attention. On one hand, once a text-to-video generation is trained, it can be further finetuned for image-to-video generation [4]. In [11], the authors introduce additional structural conditions (*e.g.*, depth maps) for

more controllable video generation. Alternatively in [38], a latent flow diffusion model is introduced for image-conditioned video generation by explicitly generating a sequence of optimal flows and masks as the guidance. On the other hand, a few concurrent works to ours directly approach image-to-video generation on top of video diffusion models [58, 63, 70]. All these works share a similar spirit to text-to-video generation models but add additional images as the reference.

Our method uses a diffusion model but differs from all the aforementioned methods in that we reformulate video generation as a frame-to-frame translation problem based on the history frames. As [6] suggests image-to-image translation can maintain a decent visual consistency. In addition, we also introduce a motion flow magnitude as another condition to explicitly control the temporal dynamics.

Future Frame Prediction. Future frame prediction or forecasting [40, 48, 56] has been a long-standing problem before the prevalence of diffusion models. It has been used as an anomaly detection approach by comparing the observed frame and the predicted ones [3, 34] and video representation learning for various downstream tasks [14]. For these problems, a recurrent network such as LSTM [53], ConvLSTM [39, 56] or 3D-CNN [1] is usually used as the model architecture, and GAN [17] or Variational Autoencoder (VAE) [29] is used as the learning objectives. With the emergence of VQ-VAE [43], the authors in [25] exploited axial transformer blocks to chain the encoder-decoder for autoregressive next-frame prediction. In [57], the authors proposed masked conditional video diffusion to unify different tasks of video prediction, generation and interpolation. Nevertheless, all of these models are trained on domain-specific video datasets such as MovingMNIST [30], CATER [16] and UCF-101 [52], *etc*, far from being a generic video generation model.

Our work takes inspiration from future frame prediction methods but proposes a simpler yet effective strategy by formulating it as an image-to-image translation problem. Furthermore, our model simultaneously takes image, text and motion magnitude as the guidance for better controllability. To attain a model as general as possible, we curate a new training dataset covering a wide range of domains. Without any further dataset-specific finetuning, our model achieves plausible video generation results as shown in Fig. 1.

3 Method

Our goal is to generate GIFs, given an initial frame, a motion description, and a measure of optical flow. We frame this as an image-to-image translation problem based on latent diffusion. We first explain how we created our training dataset in Sec. 3.1, then outline our model's principles and training strategy in Sec. 3.2. Next, we delve into the specifics of our proposed model, explaining its various components in Sec. 3.3. Finally, we focus on the loss functions used to train our model in Sec. 3.4.

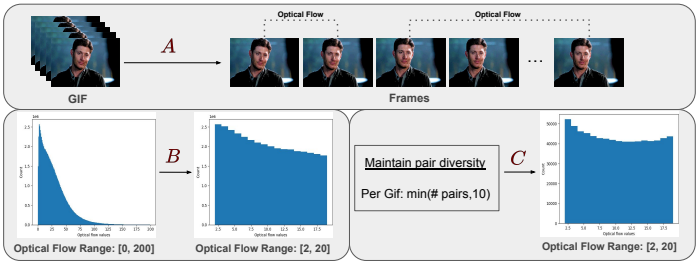


Fig. 2: The three step process of curating the TGIF dataset. Starting from extracting frames A to restricting the range of optical flow B and then maintaining the diversity of pairs C.

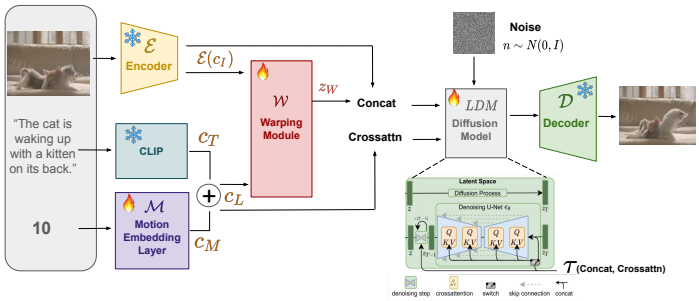


Fig. 3: *Pix2Gif* model pipeline. We propose an end-to-end network where the inputs are encoded by $\mathcal{E}$, CLIP and $\mathcal{M}$ to output $\mathcal{E}(c_L)$,c_T and c_M respectively, which then goes into $\mathcal{W}$ to form the conditioning input for LDM.

3.1 Dataset

We used the Tumblr GIF (TGIF) dataset [33], which predominantly consists of human-centric animated GIFs described by captions. The dataset features a range of GIFs with varied movements over 1-3 seconds.

The curation process, as shown in Fig. 2, involved extracting frames from all GIFs and calculating the optical flow between all possible frame pairs. The number of extracted frames from each GIF varied, with an average of about 41 frames. The optical flow histogram calculated between all frames ranged from [0, 200]. We selected the [2, 20] range, capturing small yet significant motion and excluding pairs with drastic changes. Despite the restricted range, we still had a significant number of training pairs. To avoid model overfitting and maintain diversity, we randomly selected a minimum of 10 pairs or the number of pairs within the restricted range from each GIF. This approach resulted in a final dataset with nearly equal representation of all values within the selected range. The final dataset contained 783,184 training pairs and 105,041 validation pairs. Each data point consisted of a pair of frames from the same GIF, the corresponding caption, and the calculated optical flow between the frames.

3.2 Preliminary: Instructed Image Editing

Our model is fundamentally grounded in the latent diffusion models (LDMs) for image generation and editing [6, 45]. More specifically, we build upon InstructPix2Pix [6] by framing our objective in the context of an instructed image-to-image translation task. Given an image x, the forward diffusion procedure introduces noise to the encoded latent z, thereby producing a noisy latent vector z_t. This process is carried out over T timesteps, with each timestep $t \in \{1, ..., T\}$ seeing an increment in the noise level until it culminates into a random noise n. A network e_θ is trained by minimizing the following latent diffusion objective to predict noise existing in the noisy latent z_t, considering factors image conditioning c_I and textual instruction c_T:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T)||_2^2\right] \tag{1}$$

where $\mathcal{E}$ is the VQ-VAE encoder that transforms the images from pixel space to discrete latent space. To facilitate image conditioning, z_t and $\mathcal{E}(c_I)$ are concatenated and then fed into a convolutional layer. The model is trained for conditional and unconditional denoising, given the image and caption condition individually or collectively.

3.3 Our Model: Pix2Gif

We build our model similar to InstructPix2Pix and frame our objective in the context of a text-instructed and motion-guided temporal editing problem. Compared with the original InstructPix2Pix pipeline, the main innovation is the newly introduced motion-based warping module. The overall model pipeline is shown in Fig. 3.

Our model takes three inputs: an image, a text instruction, and a motion magnitude. These inputs are fed into the model through two pathways - once through the diffusion model directly, and again through the warping module, which will be discussed in Sec. 3.3 and Sec. 3.3. When passed through the caption, we add the phrase "The optical flow is _." to the original caption. The flow input is then appended at the end in a word format rather than a numerical one, as the CLIP model tends to assign higher similarity scores to word forms than to numerical representations of numbers for the same image. Finally, our model is trained by minimizing the following loss function:

$$L'_{LDM} = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,c_M,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T, c_M)||_2^2\right] \tag{2}$$

where c_M is the motion condition. The altered caption is processed via the pre-trained CLIP model to yield c_T, while the output of $\mathcal{M}$ gives us c_M. These two conditions are then added linearly, serving as the conditioning input for both the Warping Module $\mathcal{W}$ (discussed in Sec. 3.3) and the Latent Diffusion Model LDM (referenced in Sec. 3.2).

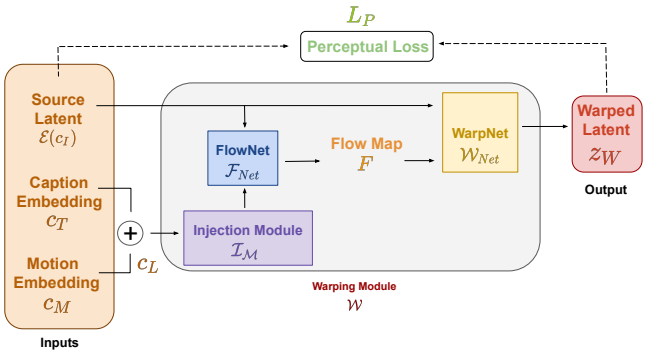


Fig. 4: Deep dive into the Warping Module $\mathcal{W}$. It comprises of three units: $\mathcal{I}_M$, $\mathcal{F}_{Net}$ and $\mathcal{W}_{Net}$.

Motion Embedding Layer In traditional conditional diffusion models like [6], text prompts are often enough for image generation. We initially passed the motion input indirectly through the prompt. However, this divided the model's attention on a single caption token, which is problematic when the main caption stays the same but the motion input changes. To let the model focus on the motion input, we included an embedding layer that converts the motion input into an integer and selects an embedding vector. This vector is then duplicated and concatenated with itself to generate c_M, which when combined with the caption embedding c_T, provides the conditioning input $c_L = c_T + c_M$ for both warping module $\mathcal{W}$ and LDM.

Warping Module One of the main components of Pix2Gif is the Warping Module $\mathcal{W}$. As illustrated in Fig. 4, it technically comprises two networks: the FlowNet ($\mathcal{F}_{Net}$) and the WarpNet ($\mathcal{W}_{Net}$). Ordinarily, the computation of optical flow involves two images. However, in this case, we initially have only one image - the source image - and that too in the latent domain. Thus, our goal is to learn the optical flow utilizing just one latent image. This is achieved via $\mathcal{F}_{Net}$, conditioned on c_L, which guides it to generate a flow feature map in the intended direction with the hint of text and motion prompts. This condition is processed by the Injection Module ($\mathcal{I}_{\mathcal{M}}$), a compact encoder designed to make c_L compatible for concatenation with one of the intermediate feature maps near the end of the network. This configuration enables $\mathcal{F}_{Net}$ to independently learn high-level features, which are then guided in the desired direction with the introduction of c_L. The architecture of $\mathcal{F}_{Net}$ resembles that of UNet, producing an output with a fixed channel of 2 to capture changes in the horizontal and vertical components. This optical flow feature map $F = \mathcal{F}_{Net}(\mathcal{E}(c_I), \mathcal{I}_{\mathcal{M}}(c_L))$, along with the source latent ($\mathcal{E}(c_I)$), is then processed through $\mathcal{W}_{Net}$ to yield a Fischer map $z_W = \mathcal{W}_{Net}(\mathcal{E}(c_I), F)$. This transformation is learned more efficiently and abstractly in the latent space than in the pixel space.

Table 1: Quantitative comparison with state-of-the-art image-text-to-video generation models for the zero-shot setting.

Method	UCF-101 [52]			MSR-VTT [64]		
	FVD ↓	CLIPSim ↑	PIC ↑	FVD ↓	CLIPSim ↑	PIC ↑
I2VGen-XL [70]	563.12	**0.2865**	0.6329	278.62	0.2272	0.6018
DynamiCrafter [63]	527.06	0.2796	0.6307	271.63	**0.2602**	0.6135
Pix2Gif (Ours)	**285.02**	0.2815	**0.8763**	**168.69**	0.2573	**0.8521**

3.4 Losses

Our model incorporates two different types of losses. The first type is the standard L2 loss Eq. (2), which is utilized by the stable diffusion model and talked about in Sec. 3.2

The second loss type in our model is the perceptual loss, which is calculated by comparing the latent features of the image condition $\mathcal{E}(c_I)$ and the warped image z_W. This is implemented using a pre-trained VGG network [49], modified to accommodate 4 channels instead of the standard 3. The modification involves averaging the weights from the first three channels to initialize the fourth. The perceptual loss, L_p, is defined as:

$$L_p(\mathcal{E}(c_I), z_W) = \sum_k \lambda_k ||\phi_k(\mathcal{E}(c_I)) - \phi_k(z_W)||^2 \tag{3}$$

Here, $\phi_k(.)$ be the feature map of the k-th layer of the VGG network, $||.||$ denotes the Frobenius norm, and λ_k is a weighting factor. This loss ensures that the warped image maintains high-level features like edges, textures, and object types, making the images more perceptually and semantically similar and preserving the overall structure of the source image.

In conclusion, the total loss function, denoted as L_T, for our objective is computed by a weighted sum of the two individual losses.

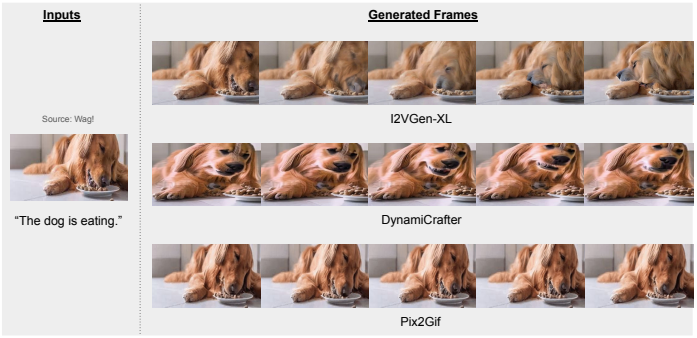
$$L_T = L'_{LDM} + \lambda_P L_P \tag{4}$$

Here, λ_P is the weighting factor for perceptual loss. These two losses together provide a holistic framework to train our model by ensuring pixel-level accuracy, preservation of high-level features, and smooth motion transitions.

4 Experiments

4.1 Setup

Datasets We utilize the Tumblr GIF (TGIF) dataset for our training and validation purposes as discussed in Sec. 3.1. We evaluate our model on two datasets: MSR-VTT [64] and UCF-101 [52], following the common practice. For these datasets, we follow the sampling strategy as outlined in [63].



(a)

(b)

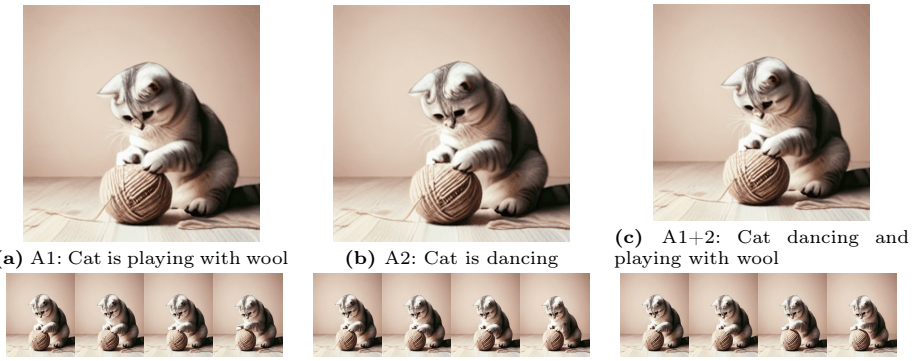
Fig. 5: Comparison studies with other image-text to video models. Given a source image and a caption, frames are extracted from the generated 16-frame video at 256x256 resolution.

Implementations Our model is initialized with the exponential moving average (EMA) weights of the Stable Diffusion v1.5 checkpoint[1] and the improved ft-MSE autoencoder weights[2]. We trained the model at 256x256 resolution for 7 epochs on a single node of 16 V100 GPUs for 25k steps. We used the AdamW optimizer with a learning rate of 10^{-4}. We set the weighting factor for perceptual loss (λ_P) as 10^{-2}.

Metrics We report Frechet Video Distance (FVD) [54], CLIP Similarity (CLIP-Sim) which is the average similarity calculated for all the generated frames with the input caption and Perceptual Input Conformity (PIC) as described in [63] for all methods. For comparison, we assess the zero-shot generation performance on I2VGen-XL [70] and DynamiCrafter [63].

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5/blob/main/v1-5-pruned-emaonly.ckpt

[2] https://huggingface.co/stabilityai/sd-vae-ft-mse-original/blob/main/vae-ft-mse-840000-ema-pruned.ckpt



(a) A1: Cat is playing with wool **(b)** A2: Cat is dancing **(c)** A1+2: Cat dancing and playing with wool

Fig. 6: *Pix2Gif* showing composition capabilities for different types of motions. [GIFs best viewed in Adobe Acrobat Reader]

Table 2: Ablation study comparing image translation methods with a focus on motion coherency at varying cfg_img values.

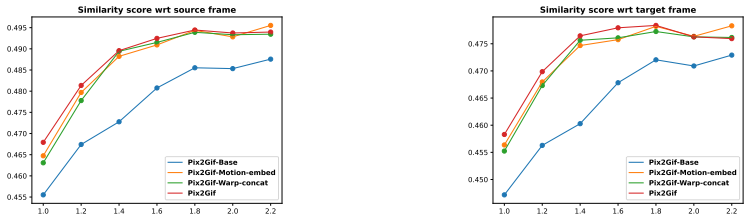
Method / cfg_img	1.4		1.6		1.8		2.0	
	L2 ↓	**PCC** ↑	**L2** ↓	**PCC** ↑	**L2** ↓	**PCC** ↑	**L2** ↓	**PCC** ↑
InstructPix2Pix [6]	23.429	-0.229	25.492	-0.028	27.037	-0.423	27.530	0.139
Pix2Gif-Base	7.580	0.989	5.188	0.987	5.595	0.992	7.029	0.991
Pix2Gif	**1.746**	**0.995**	**1.972**	**0.995**	**2.944**	**0.997**	**4.076**	**0.997**

4.2 Results

Comparisons with previous works Fig. 5 and Tab. 1 provide a qualitative and quantitative comparison of three image-text to video models: I2VGen-XL [70], DynamiCrafter [63], and our *Pix2Gif*.

In Fig. 5a, the I2VGen-XL model misshapes the dog's face and generates it sideways in a nonsensical manner. DynamiCrafter appears to disregard the input parameters, as the initial frame differs significantly in position, color, and texture. It is also challenging to discern whether the dog is eating or merely moving its mouth. Our model, *Pix2Gif*, accurately retains all the dog's details and successfully depicts it eating from a plate. In Fig. 5b, we assess the models' capabilities by generating a video from a relatively dark image. Once again, I2VGen-XL starts strong, producing some impressive frames, but these soon turn into highly stylized and improbable images. DynamiCrafter appears to misinterpret the input image, generating something significantly different, although it seems to adhere to the caption. Conversely, *Pix2Gif* comprehends the inputs effectively and produces corresponding motion while preserving the overall integrity of the source image.

Quantitatively, *Pix2Gif* excels in both the FVD and PIC metrics shown in Tab. 1, which aligns with our observations of the frames generated in Fig. 5. These frames effectively preserve the structure and closely adhere to the input



(a) source frame score **(b)** target frame score

Fig. 7: Ablation study between the earlier variants of our model by comparing average similarity score for 100 samples.

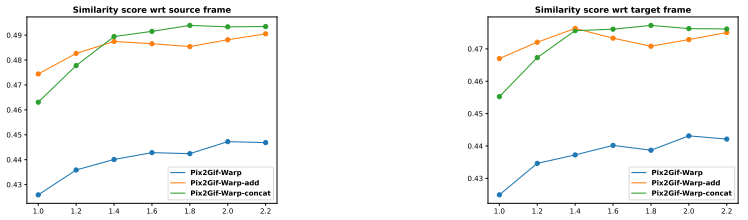
prompts (source image and caption). However, *Pix2Gif* does not perform as well in the CLIPSim metric, despite accurately following the caption. The other two models as seen in Fig. 5 do follow the caption, but they fail to adhere to the input image and produce plausible temporal transitions. This is partially attributed to the inherent model design in these two methods. Both methods attempt to generate a full sequence of frames at once using the 3D diffusion network, which inevitably compounds the spatial and temporal dimensions. Moreover, the results indicate that they function more as text-to-video models than image-text-to-video models, especially DynamiCrafter. This discrepancy also raises questions about the effectiveness of the CLIPSim metric for evaluating image-text-to-video models and calls for more sophisticated metrics for evaluating video generation.

Compositionality of actions Fig. 6 illustrates an intriguing emerging capability of *Pix2Gif*: the ability to combine actions. In Fig. 6a, we see a cat playing with wool, with only the cat's paws and the wool moving. In Fig. 6b, we instruct the cat to dance, resulting in the cat moving its body but the wool remaining still. Finally, in Fig. 6c, we provide a caption that blends the actions from Fig. 6a and Fig. 6b. The result is a scene where the cat is both moving the wool and its body. This demonstrates *Pix2Gif*'s ability to comprehend the caption and its associated motion, and to convert that understanding into a GIF. Such compositional capability significantly increases user controllability, a crucial aspect for practical applications.

4.3 Ablations

We design a few variants of *Pix2Gif* for our ablation studies:

- *Pix2Gif-Base*: We train InstructPix2Pix with our data, and append the text prompt with "The optical flow is _.".
- *Pix2Gif-Motion-embed*: The motion embedding layer is added to encode the motion magnitude and combined with textual embedding.
- *Pix2Gif-Warp*: We further add the warping module into the model but differently only use the warped feature for the LDM.



(a) source frame score **(b)** target frame score

Fig. 8: Ablation study on ways to input z_W to LDM by comparing average similarity score for 100 samples.

- *Pix2Gif-Warp-add*: Different from *Pix2Gif-Warp*, we instead add the warped feature and source image feature as input to the LDM.
- *Pix2Gif-Warp-concat*: Instead of adding in *Pix2Gif-Warp-add*, we concatenate the warped feature and source image feature as the input, but do not include the perceptual loss.

For comparative studies with our model's variants, we generate an 8-frame video. We use the X-CLIP model [37] to extract features from our generated video, and CLIP to extract features from the source, target, and generated frames. The optimal range of cfg_img for best results is considered to be [1.6, 2.2]. Throughout this discussion, we evaluate our model's performance using four metrics, which we believe effectively measure the different aspects of generating motion through the image translation framework used in *Pix2Gif*.

Motion Coherency Our task is framed as an image translation problem with motion magnitude as a guide. We assess the motion quality or temporal coherence in the GIFs using L2 loss and Pearson Correlation Coefficient (PCC) for InstructPix2Pix, *Pix2Gif-Base*, and *Pix2Gif*. The L2 loss evaluates the match between the motion values of the generated frames and the actual inputs, while the PCC checks if they follow the same trend. These metrics are calculated between the input motion magnitude values and the optical flow values, derived from the source image and the generated frames. Our model exhibits the highest correlation and the lowest L2 loss across all cfg_img values as seen in Tab. 2, proving its efficacy and controllability in generating GIFs with specific motions. The *Pix2Gif-Base* outperforms the original InstructPix2Pix, emphasizing the importance of our new dataset.

Image-Video Similarity Score To evaluate the semantic properties of the video produced by our model, we created two similarity scores: a source frame score and a target frame score. The source frame score quantifies how well the video retains the primary attributes of the source frame, essentially measuring the accuracy of the source image portrayal throughout the video. The target frame score indicates the precision of the scene or subject development from

the source frame in the video. It also underscores the model's ability to handle uncertainty and potential changes, as the target frame represents a possible state that the video might reach.

We began our experiments with *Pix2Gif-Base* and then enhanced it incrementally, first with $\mathcal{M}$ (*Pix2Gif-Motion-embed*), then by adding $\mathcal{W}$ (*Pix2Gif-Warp*). Here, we conducted an ablation study shown in Fig. 8 to understand the best way to feed z_W into LDM. Our experiments included feeding only z_W (*Pix2Gif-Warp*), adding it with the $\mathcal{E}(c_I)$ (*Pix2Gif-Warp-add*), and concatenating both (*Pix2Gif-Warp-concat*). Both *Pix2Gif-Warp-add* and *Pix2Gif-Warp-concat* had higher average similarity scores than *Pix2Gif-Warp*, and *Pix2Gif-Warp-concat* performed the best. This can be attributed to the fact that in the addition process, $\mathcal{E}(c_I)$ loses its unique characteristics, which are required by the diffusion model for effective unconditional denoising. Therefore, to achieve the best results, we combined $\mathcal{E}(c_I)$ and z_W before inputting them into the concat attention layer of the LDM. Now finally we integrate L_P (*Pix2Gif*) and compare them in Fig. 7. As expected, *Pix2Gif-Base* was outperformed by the other versions. *Pix2Gif* generated more coherent and controlled motion, albeit often with limited extent. All models began to converge outside the optimal range, producing similar frames. The *Pix2Gif-Motion-embed* model creates a significant amount of one-directional motion, which can sometimes be nonsensical, and hence the addition of $\mathcal{W}$ helps to mitigate this issue.

5 Limitations and Future Work

The current *Pix2Gif* model is our initial attempt to generate videos by treating it as an image translation task. However, this method has some limitations that prevent us from generating high-quality and long GIFs or videos. Firstly, the model generates images with a resolution of 256x256 pixels. If these images are used to generate subsequent frames, the quality of the frames deteriorates further. Secondly, due to limitations in computational power, we are only able to use a small portion of a larger, curated dataset for training our model. Our primary objective now is to improve the quality of the generated frames, as this could significantly enhance the effectiveness of this method.

6 Conclusion

In this work, we proposed *Pix2Gif*, an image-to-GIF (video) generation model based on an image-to-image translation paradigm. To ensure temporal coherence across frames, we proposed a motion-guided warping module that learns to spatially warp the source image feature into the target one while maintaining visual consistency via a perceptual loss. Starting from TGIF, we curated a new dataset specifically used for training our model. The experimental results demonstrated the effectiveness of our model to generate GIFs with better temporal coherence compared with current state-of-the-art methods. Interestingly, the model also exhibits better controllability and some emerging action compositionality.

References

1. Aigner, S., Körner, M.: Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans (2018)
2. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics **42**(4), 1–11 (Jul 2023). https://doi.org/10.1145/3592450, http://dx.doi.org/10.1145/3592450
3. Baradaran, M., Bergevin, R.: Future video prediction from a single frame for video anomaly detection (2023)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023)
5. 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 (2023)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
7. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., Li, Y., Krishnan, D.: Muse: Text-to-image generation via masked generative transformers (2023)
8. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer (2022)
9. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models (2021)
10. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems **34**, 19822–19835 (2021)
11. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
12. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis (2021)
13. Fu, T.J., Hu, W., Du, X., Wang, W.Y., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models (2023)
14. Fujitake, M., Sugimoto, A.: Video representation learning through prediction for online object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 530–539 (2022)
15. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models (2023)
16. Girdhar, R., Ramanan, D.: Cater: A diagnostic dataset for compositional actions and temporal reasoning (2020)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
18. Gu, Y., Yang, J., Usuyama, N., Li, C., Zhang, S., Lungren, M.P., Gao, J., Poon, H.: Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. arXiv preprint arXiv:2310.10765 (2023)

19. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation (2023)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
22. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
23. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models (2022)
24. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers (2022)
25. Hu, Y., Luo, C., Chen, Z.: Make it move: Controllable image-to-video generation with text descriptions (2022)
26. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
27. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
28. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
29. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022)
30. Lee, J., Lee, J., Lee, S., Yoon, S.: Mutual suppression network for video prediction using disentangled features (2019)
31. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
32. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation (2023)
33. Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., Luo, J.: Tgif: A new dataset and benchmark on animated gif description (2016)
34. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection – a new baseline (2018)
35. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps (2022)
36. 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 (2022)
37. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition (2022)
38. Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models (2023)
39. Oh, J., Guo, X., Lee, H., Lewis, R., Singh, S.: Action-conditional video prediction using deep networks in atari games (2015)
40. Oliu, M., Selva, J., Escalera, S.: Folded recurrent neural networks for future video prediction (2018)
41. van den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks (2016)
42. van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., Kavukcuoglu, K.: Conditional image generation with pixelcnn decoders (2016)

43. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018)
44. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)
46. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2023)
47. 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. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
48. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., kin Wong, W., chun Woo, W.: Convolutional lstm network: A machine learning approach for precipitation nowcasting (2015)
49. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
50. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-a-video: Text-to-video generation without text-video data (2022)
51. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022)
52. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)
53. Srivastava, N., Mansimov, E., Salakhutdinov, R.: Unsupervised learning of video representations using lstms (2016)
54. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
55. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description (2022)
56. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction (2018)
57. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation (2022)
58. Wang, C., Gu, J., Hu, P., Xu, S., Xu, H., Liang, X.: Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance. arXiv preprint arXiv:2312.03018 (2023)
59. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)
60. Weissenborn, D., Täckström, O., Uszkoreit, J.: Scaling autoregressive video models (2020)
61. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions (2021)
62. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation (2021)
63. 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 preprint arXiv:2310.12190 (2023)

64. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
65. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., Wang, L.: Reco: Region-controlled text-to-image generation (2022)
66. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
67. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023)
68. Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., Ross, C., Polyak, A., Howes, R., Sharma, V., Xu, P., Tamoyan, H., Ashual, O., Singer, U., Li, S.W., Zhang, S., James, R., Ghosh, G., Taigman, Y., Fazel-Zarandi, M., Celikyilmaz, A., Zettlemoyer, L., Aghajanyan, A.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning (2023)
69. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
70. 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 preprint arXiv:2311.04145 (2023)
71. Özgün Çiçek, Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation (2016)