Video Interpolation with Diffusion Models

Siddhant Jain*
Google Research
siddhantjain@google.com

Daniel Watson*
Google DeepMind
watsondaniel@google.com

Eric Tabellion
Google Research
etabellion@google.com

Aleksander Hołyński
Google Research
holynski@google.com

Ben Poole
Google DeepMind
pooleb@google.com

Janne Kontkanen
Google Research
jkontkanen@google.com

Abstract

We present VIDIM, a generative model for video interpolation, which creates short videos given a start and end frame. In order to achieve high fidelity and generate motions unseen in the input data, VIDIM uses cascaded diffusion models to first generate the target video at low resolution, and then generate the high-resolution video conditioned on the low-resolution generated video. We compare VIDIM to previous state-of-the-art methods on video interpolation, and demonstrate how such works fail in most settings where the underlying motion is complex, nonlinear, or ambiguous while VIDIM can easily handle such cases. We additionally demonstrate how classifier-free guidance on the start and end frame and conditioning the super-resolution model on the original high-resolution frames without additional parameters unlocks high-fidelity results. VIDIM is fast to sample from as it jointly denoises all the frames to be generated, requires less than a billion parameters per diffusion model to produce compelling results, and still enjoys scalability and improved quality at larger parameter counts. Please see our project page at vidim-interpolation.github.io.

1. Introduction

Diffusion Models [15, 49, 50] have recently exploded in popularity for generative modeling of images and other forms of continuous data such as audio [5] and video [18]. Compared to previous methods for generative modeling such as Generative Adversarial Networks (GANs) [12], diffusion models enjoy significantly more training stability due to the fact that they optimize the evidence lower bound (ELBO) [24], as opposed to having the complex dynamics of two models in a zero-sum game like GANs, and do not suffer from posterior collapse like variational autoencoders (VAEs) due to the limited form of the approximate posterior. This ease of training has led to significant advances in generative modeling, such as compelling results in text-to-image [38, 40, 44] and text-to-video [17, 55] generation, image-conditioned generation tasks [43], and even 3D novel view synthesis [27, 36, 58].

In this paper, we explore the specific task of video interpolation with diffusion models. Video interpolation refers to the problem of generating intermediate frames between two consecutive frames of video. Video interpolation techniques have been used for many desirable applications, e.g., generating slow motion videos from existing videos, video frame rate up-sampling (*e.g.* 30 fps $\rightarrow$ 60 fps) or interpolating between near-duplicate photographs.

Numerous methods have been proposed in prior work [11], but even the state-of-the-art [20, 26, 39] fails to generate plausible interpolations when the start and end frame become increasingly distinct, as these methods rely on linear or unambiguous motion. Moreover, while existing video diffusion models can be used for the video interpolation task [17, 18], as we carefully study in this work, quantitative and qualitative results improve significantly by explicitly training generative models that are conditioned on the start and end frames, as generative models have the key advantage of producing samples, as opposed to predicting the mean. As we will show, this is also strongly supported by ratings from human observers.

In this work, we show that diffusion based generative models can overcome the limitations of prior state-of-the-art models for video interpolation. We summarize our main contributions below:

- We develop a cascaded video interpolation diffusion model, which we dub **VIDIM**, capable of generating high-quality videos in between two input frames.
- We carefully ablate some of the design choices of VIDIM, including parameter sharing to process condi-

*Equal contribution.

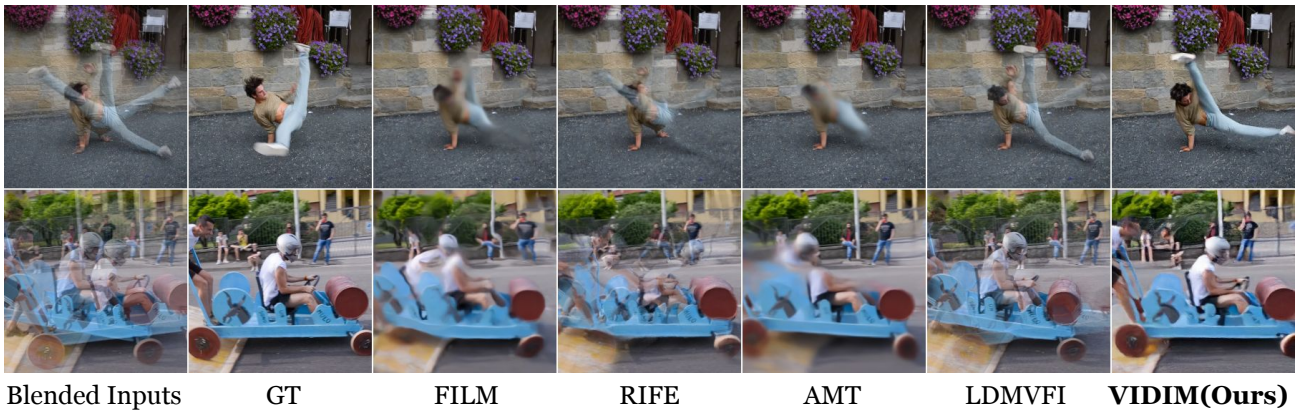

Figure 1. Frame interpolation for very large and ambiguous motion. The middle frame of an interpolated video with FILM[39], RIFE [20], LDMVFI [9] and AMT[26] shows large blurry artifacts. VIDIM, however, is able to recover a plausible output frame. Note that due to the ambiguity of the problem, VIDIM's output is not always similar to the ground truth (especially clear in the top example), but corresponds to a different choice of motion. See the Supplementary Website for video outputs.

tioning frames and the use of classifier-free guidance [14], demonstrating their importance to achieve good results.

- We propose two curated difficult datasets targeted for generative frame interpolation: Davis-7 and UCF101-7, based on widely used Davis [35] and UCF101 [51] datasets.
- We show that VIDIM generally achieves better results compared to prior state-of-the-art in these difficult interpolation problems across generative modeling metrics. We show by user study that VIDIM is almost always preferred over the baselines in qualitative evaluation.

2. Background and Related Work

Video frame interpolation (VFI) is a classic computer vision problem with a sizable body of existing work. VFI is closely related to optical flow computation, which is an equally deeply studied problem. Instead of attempting a comprehensive list of works in these areas we discuss the recent state-of-the-art that is most relevant for our work. For a recent survey on this topic, see [11].

Most recent video frame interpolation architectures contain a feature extractor (e.g. decoder) correspondence estimation and image warping (e.g. optical flow) and frame synthesis (e.g. decoder). Most works also agree that optical flow is best learned for the frame interpolation task specifically [8, 25, 39, 60] or fine tuned for it [21, 32, 33]. Some methods use backward warping (gather) [20, 25, 26, 33, 39] while others use forward warping (splatting/scatter) [32].

Recent works employ hybrid CNN and transformer architectures [29, 61] and all-pairs dense feature matching [26] inspired by the state-of-the-art optical flow method [53]. Regardless of the details, the progress of video frame interpolators have been driven by benchmarks [2, 4, 35, 51, 60] that are usually prepared to assume linear or mostly unambiguous motion, by either having explicit linearity constraints [60] or just using samples that are not very far apart in time. It is rare to go beyond the non-linear motion assumption, although a few methods do assume a quadratic motion model [28, 59]. Some recent works specialize on large motions [39, 47], but yet only to an extent where the motion is mostly unambiguous and thus can be solved with a non-generative model. In our work we show that when the input images are much further than 1/30s apart, the problem becomes highly ambiguous and best addressed as a conditional generative problem.

LDMVFI [9] and MCVD [56] are some recent diffusion based frame interpolatation methods. Different from LDMVFI we model in pixel space and generate the entire video at once which is key for consistent motion. We also focus on explicitly training for video frame interpolation where as MCVD studies a range of video generative modeling tasks. While it is difficult to design truly fair comparisons as existing video models are often too large and source code is not available [17, 48], the conditioning mechanism can still be studied. Ho et al. [18] propose the use of imputation (and additionally with a mechanism resembling classifier-free guidance) to adapt video models to be conditioned on input frames. We train a super-resolution model adopting this strategy, and in section 4.4 demonstrate that this approach performs worse despite being having the exact same hyperparameters otherwise.

3. Methodology

We now present the technical details behind VIDIM, our cascaded diffusion model for video interpolation. Prior

work has shown that diffusion models do not achieve good sample quality for high-resolution generation with a single model without revising several hyperparameters and architecture details [19]. We in fact tried training a base VIDIM model to generate a 9x128x128[1] video following the changes proposed by Hoogeboom et al. [19], but we found early on that these models did a very poor job at modeling high-frequency details at this resolution. We thus followed the cascaded model strategy of Ho et al. [16], i.e., training separate base and super-resolution models. While there is additional overhead to maintaining multiple diffusion models, this still avoids several complexities of latent diffusion models [40], such as finding an optimal encoder-decoder model, and having to address temporal inconsistency in the decoder with other training or fine-tuning procedures [3]. We train two video diffusion models: first we train a **base model** that is conditioned on 2 64x64 frames and generates 7 64x64 in-between frames. Then we train a **super-resolution model** conditioned on 2 256x256 frames and 7 64x64 frames that generates the 7 corresponding 256x256 frames. We intentionally chose an odd number of frames to allow evaluating the middle frame, similarly to prior work on video interpolation.

3.1. Model architecture

VIDIM is inspired by Imagen Video [17], where the UNet architecture [41] is adapted for video generation by sharing all convolution and self-attention blocks over frames, and feature maps are only allowed to mix over frames with the addition of temporal attention blocks where the query-key-value sequence lengths are the number of frames. Simple positional encodings (differing over frames) for video timestamps normalized to $[0, 1]$ are summed to the usual noise level embeddings (identical for all noisy frames). We propagate these embeddings to each UNet block using FiLM [34], similarly to Nichol and Dhariwal [31]. We do the same for the super-resolution model to condition on the high-resolution start and end frames. The super-resolution model only differs from the base model in that (1) it concatenates each (naively upsampled) low-resolution conditioning frame to the noisy high-resolution frames along the *channel* axis, and (2) it downsamples *before* the first convolutional residual block, following Saharia et al. [44] to reduce memory usage. For more stable and efficient training, we additionally use attention blocks following Dehghani et al. [10], which employ query-key normalization and an MLP block that runs in parallel to the attention block.

Parameter-free frame conditioning. One key innovation we highlight is introducing frame-conditioning without any additional parameters: in both the base and super-resolution models, we condition on the start and end frames simply by feeding these two additional frames to the entire UNet (i.e., concatenating along the *frame* axis). Because the feature maps for each frame additionally depends on the noise levels, we simply set fake noise levels for the conditioning frames as the minimum noise level (maximum log-signal-to-noise-ratio). This adds two new sequence elements to the temporal attention layer, which lets information from the conditioning frames propagate to the rest of the network without any additional parameters. This contrasts other diffusion architectures [27, 63], where the usual choice is additional cross-attention layers which doesn't scale to condition on more frames and make the parameter count dependent on the number of frames. Other work has found that concatenating extra frames along the *channel* axis following, e.g., Saharia et al. [43], leads to worse sample quality when the generated and conditioning frames are not perfectly aligned [58].

Guidance on conditioning frames. In our results, we show that classifier-free guidance (CFG) [14] on the conditioning frames is essential to achieve the best sample quality. Similarly to our parameter-free frame conditioning strategy, we would like masked conditioning frames for CFG to play naturally with parameter sharing across frames. To achieve this, instead of zeroing-out the conditioning frames, we replace them with isotropic Gaussian noise and set their corresponding noise levels to the maximum value. It would also be confounding if we zero-out the timestamps for the conditioning frames, so we instead replace them with a learned null token.

3.2. Diffusion modeling choices

We now provide a brief overview of the training objective formulation we utilize for VIDIM. We begin with the formulation of Kingma et al. [22], where we use the simpler continuous-time objective (as opposed to that of Ho et al. [15]) for learning a data distribution $p(\boldsymbol{x}|\boldsymbol{c})$, where $\boldsymbol{c}$ are the start and end frames and $\boldsymbol{x}$ are the middle frames. We define a forward process at every possible log signal-to-noise-ratio λ (a.k.a. "log-SNR") in the usual manner via

$$q(\boldsymbol{z}_\lambda|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}_\lambda; \alpha_\lambda \boldsymbol{x}, \sigma_\lambda^2 \boldsymbol{I}\right) \tag{1}$$

where $\alpha_\lambda = \sqrt{\text{sigmoid}(\lambda)}$ and $\sigma_\lambda = \sqrt{\text{sigmoid}(-\lambda)}$. Our training objective is then

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x},\boldsymbol{c}\sim p(\boldsymbol{x},\boldsymbol{c}) \\ t\sim U(0,1) \\ \boldsymbol{z}_{\lambda_t}\sim q(\boldsymbol{z}_{\lambda_t}|\boldsymbol{x})}} e^{\frac{\lambda_t}{2}} \left\|\alpha_{\lambda_t}\boldsymbol{z}_{\lambda_t} - \sigma_{\lambda_t}\boldsymbol{v}_\theta(\boldsymbol{z}_{\lambda_t}|\boldsymbol{c}) - \boldsymbol{x}\right\|_1 \tag{2}$$

Note that, following Saharia et al. [43], we use the L1 loss as we found early on that it helps produce better high-frequency details in samples compared to the standard L2

[1] Not that large resolution by video interpolation standards

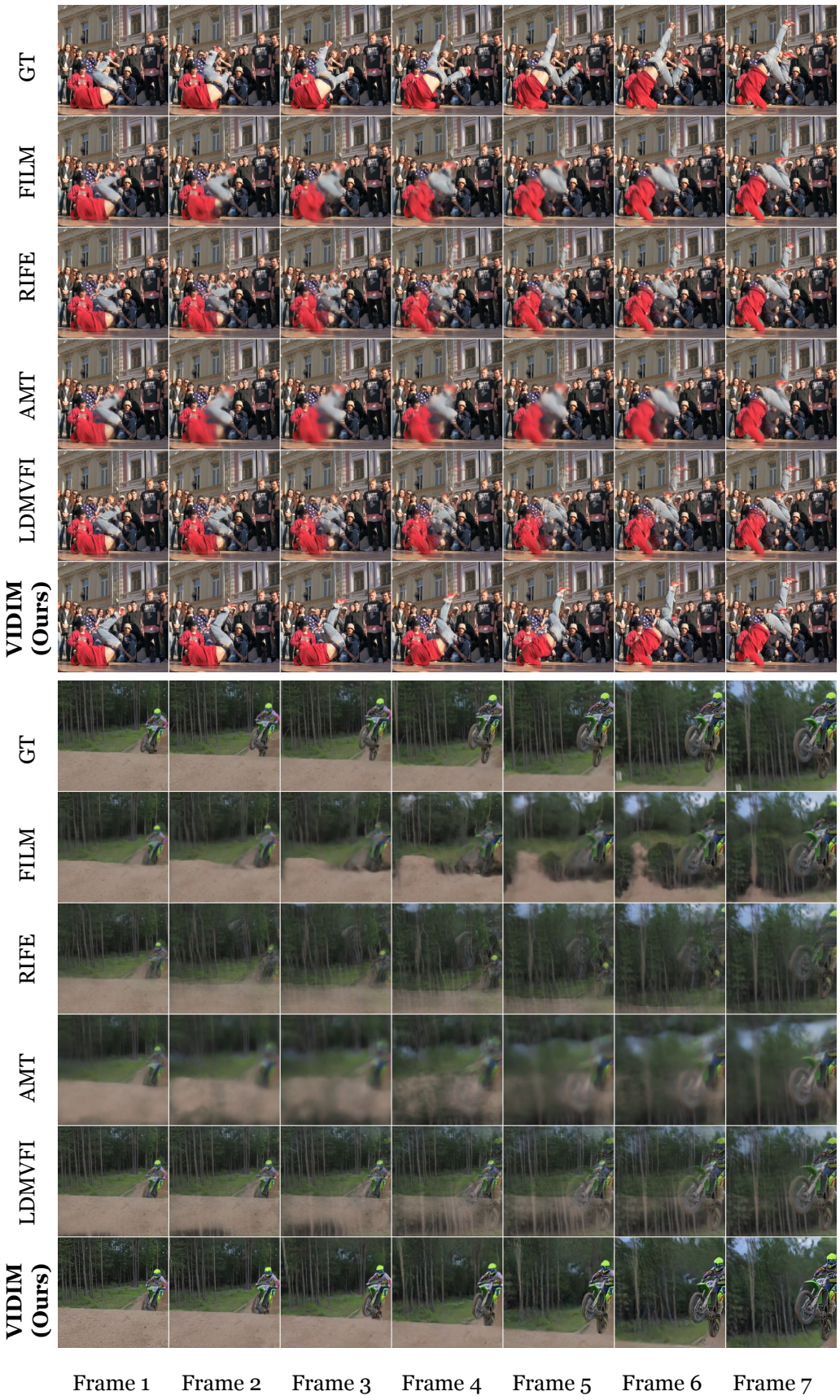

Figure 2. Two examples from DAVIS-9 dataset, showing the predicted in-between frames. **Top**: The break-dancer example demonstrates highly ambiguous motion. Our method can produce plausible video with sharp details whereas the baselines [20, 26, 39] trained with regression objective resort into predict blurry images. **Bottom**: On a very large motion with significant perspective change on the dirt bike, the baselines fail to reconstruct sharp results, where as our method produces sharp results with plausible motion.

	Davis-7 (mid-frame)				UCF101-7 (mid-frame)			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
AMT [26]	**20.12**	**0.4853**	0.2865	69.34	**25.16**	**0.7903**	0.1691	63.92
RIFE [20]	19.54	0.4546	0.2954	57.68	25.73	0.7769	**0.1564**	**42.33**
FILM [39]	19.75	0.4718	0.3048	68.88	24.96	0.7869	0.162	54.98
LDMVFI [9]	19.07	0.4175	**0.2765**	56.28	24.53	0.7712	**0.1564**	42.96
VIDIM (ours)	18.73	0.4221	0.2986	**53.38**	22.88	0.688	0.1768	53.71

Table 1. Comparison between different video interpolation baselines and VIDIM on reconstruction and generative metrics, evaluating only the middle frame out of all 7 generated frames. VIDIM samples were obtained from our best cascade with guidance weight 2.0. Note that under this setting, it does not make sense to report FVD scores.

	Davis-7					UCF101-7				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
AMT [26]	**21.09**	**0.5443**	**0.254**	34.65	234.5	**26.06**	**0.8139**	0.1442	31.6	344.5
RIFE [20]	20.48	0.5112	0.258	23.98	240.04	25.73	0.804	0.1359	18.72	323.8
FILM [39]	20.71	0.5282	0.2707	30.16	214.8	25.9	0.8118	0.1373	26.06	328.2
LDMVFI [9]	19.98	0.4794	0.2764	**22.1**	245.02	25.57	0.8006	**0.1356**	**18.09**	316.3
VIDIM (ours)	19.62	0.4709	0.2578	28.06	**199.32**	24.07	0.7817	0.1495	34.48	**278**

Table 2. Comparison between different video interpolation baselines and VIDIM on reconstruction and generative metrics, evaluating all 7 generated frames. VIDIM samples were obtained from our best cascade with guidance weight 2.0. Note these numbers (especially FID scores) are not comparable to those in 1 as the number of samples differs (here we use 7x as many images per set).

loss. We use a cosine log-SNR schedule λ_t following Kingma et al. [22], with maximum log-SNR of 20 at $t = 0$ and minimum log-SNR of -20 at $t = 1$. Note how this objective is equivalent to the re-weighted ELBO objective from Ho et al. [15] in the sense that the loss is a norm between the predicted and actual noise (hence the $e^{\frac{\lambda_t}{2}}$ factor), but using the "v-parametrization" from Salimans and Ho [46] rather than predicting the added noise directly. This is well-known to improve training stability of diffusion models.

4. Experiments

We now present our main experiments and results. All base models were trained to generate 7 64x64 frames in between the start and end 64x64 frames, and the super-resolution models condition on the original 256x256 start and end frames to upsample the 7 input frames at resolution 64x64.

In order to leverage a large-scale video dataset for these tasks, we train all VIDIM models on a mixture of the publicly available WebVid dataset [1] and other internal video datasets. In order to handle cases with large and more difficult motion, our input pipelines take bursts of 32 contiguous frames from the original videos and evenly subsample 9 frames so some frames are skipped. To additionally reduce the number of cuts and other examples that are undesirable for video interpolation, we follow the motion bracketing procedure employed by FILM [39].

4.1. Training and architecture hyperparameters

We trained all VIDIM models with the Adam optimizer [23] with a learning rate of 5e-4 (with linear warm-up for the first 10,000 steps) and $\beta_1 = .9, \beta_2 = .999$, gradient clipping at norm 1, and maintaining an EMA of the model parameters with decay rate .9999 following Ho et al. [15]. To make our ablation studies fair, *all* base models were trained for 500k steps and all super-resolution models were trained for 200k steps. All super-resolution models were trained with noise conditioning augmentation [16] on the low-resolution frames, where we re-use the noise schedule and add noise to these frames with $t \in U(0, 0.5)$ for each training example.

We additionally study different parameter counts for our VIDIM models in Tab. 3. In these experiments, we scale up the number of parameters exclusively by changing the hidden size (a.k.a. number of channels) in the *last* UNet resolution (16x16 for both the base and super-resolution models). All other UNet resolutions have the same hyperparameters across experiments: the first resolution always has 128 channels and 2 subblocks, each subblock having a convolutional block and a temporal attention block with one attention head. Middle resolutions always have 256 channels and 4 subblocks, with each subblock's temporal attention block having two attention heads. The last 16x16 block has 8 subblocks, and additionally includes a self-attention block shared over frames before each temporal attention block. We avoid self-attention at other resolutions as it is too computationally expensive. We use dropout [52] at all UNet resolutions with resolutions up to 64x64 but not beyond, following Hoogeboom et al. [19].

4.2. Comparisons with prior work

We evaluate VIDIM on several reconstruction and generative metrics, and compare against prior methods on video interpolation by running these models ourselves on the

same benchmarks. Specifically, we create subsets of the Davis [35] and UCF101 [51] datasets of 400 videos with examples that contain large and ambiguous motion and consisting of 9 frames per video. It should be emphasized that our numbers cannot be directly to prior work also due to the fact that diffusion models operate at a fixed resolution. Thus, in order to ensure that we are accurately representing all the prior work we consider, we run all the baselines ourselves on said benchmarks. We refer the reader to our Supplementary Material and website where we release our curated evaluation datasets (dubbed Davis-7 and UCF101-7) for future work and the general public, as well as playable video samples.

For each dataset, we report the following reconstruction-based metrics: peak-signal-to-noise-ratio (PSNR), structural similarity (SSIM) [57], and LPIPS [62]. We additionally report more popular metrics for generative models (specifically, FID [13] and FVD [54]), which, unlike reconstruction metrics, do not penalize *plausible* extrapolations that differ from the ground truth. It is well-known that generative models should not be expected to achieve the best scores in reconstruction-based metrics [42, 58]. In fact, it has been consistently shown in prior work that blurrier images tend to score higher in reconstruction metrics despite being rated as worse by human observers [6, 7, 30, 45]. We nevertheless report said reconstruction metrics for both completeness and utility to compare different configurations of the same model, which we do find provides useful signal for development.

We additionally create a second set of results where we only evaluate the middle frame, similarly to prior work on video interpolation, as we find that otherwise the evaluation is skewed towards better frames: most frames are too close to the start and end frames and most methods do a good job on these, while it is clear that the middle frame is usually the most difficult. For these sets of results, we must exclude FVD as the I3D network requires a video with 9 frames as input. We show results of evaluating all 7 generated frames in Tab. 2, and results of evaluating only the middle frame as in Tab. 1. All samples were obtained using the ancestral sampler proposed by Ho et al. [15] with 256 denoising steps per diffusion model.

Our quantitative evaluation demonstrates the superiority of VIDIM compared to other baselines in most generative metrics. While there is one exception, namely, that RIFE [20] achieves a better FID score than our cascaded model when evaluated over all 7 frames, this is not the case when considering only middle frame. The baseline methods will not have a difficult job with frames that are very close to the input frames, skewing the evaluation, while they clearly perform worse in the more difficult frames. This is also qualitatively apparent in our samples in Fig. 2 and in our samples in the Supplementary Material and website, especially in the cases with the largest amounts of motion. It is also noteworthy that VIDIM is always superior in FVD scores by a significant margin, showing that the aforementioned baselines produce much less natural-looking videos. This is an important quantity to consider, as FID does not consider any aspects of temporal consistency and only considers each frame individually. Notably, the RIFE baseline that achieves the best FID scores when considering all output frames, actually has the worst FVD scores and temporal consistency.

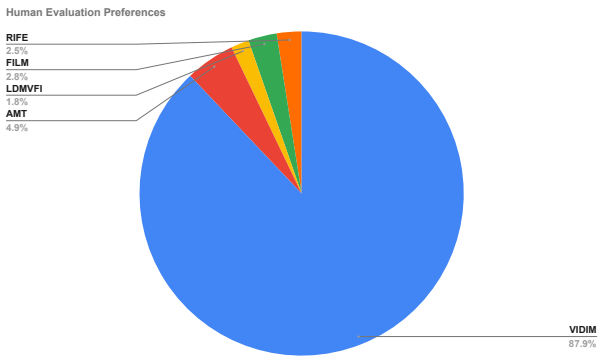

Figure 3. Human evaluation results on Davis-7, showing how often VIDIM and each baseline was preferred by human raters.

4.3. Human evaluation

To evaluate our method qualitatively, we additionally conducted a user-study where participants were shown video quadruplets playing side-by-side, each generated using the frame interpolation methods AMT, RIFE, FILM, LDMVFI and VIDIM (ours), from the same input frame pair. Users were shown up to 400 video quadruplet examples, in random order, and were asked to choose which of the four videos looks most realistic. The order used to layout the videos side-by-side in each example screen was also randomized. We evaluate on the Davis benchmark we used for all other evaluations in the paper, which contains both very challenging interpolation examples with ambiguous motion and difficult dis-occlusions, and very easy examples showing very little motion. The study involved 31 participants, and resulted in an aggregate of 1334 video quintuplet ratings. Each of the 400 examples was rated at least by one participant. Results in Fig. 3 show that VIDIM samples are very strongly preferred by human observers. Moreover, if we normalize by the number of times each individual example was rated, results change by at most by 0.3%, i.e., different participants tend to make the same choices.

4.4. Ablation study on start+end frame conditioning

We now show the importance of explicitly training both the base and super-resolution models to be conditional on

Figure 4. Sample comparison between our VIDIM medium super-resolution model (top) and an identically trained baseline minus high-resolution frame conditioning.

the input frames. To create a fair comparison, we train a super-resolution diffusion model that generates all 9 high-resolution frames of a video at training time, otherwise with all hyperparameters kept identical (including the number of training steps), and generate conditional samples via imputation: at every denoising step, we replace the start and end frame of the predicted $\boldsymbol{x}$ with the conditioning frames, add noise again and repeat. While this has been explored in prior work [18], our key hypothesis is that without the frame conditioning we propose, the training task becomes significantly more difficult. To further strengthen this baseline, we additionally try *reconstruction guidance* following Ho et al. [18], which has been shown to improve sample quality by creating a similar effect to classifier-free guidance. Specifically, reconstruction guidance amounts to using the following prediction for $\boldsymbol{x}$ at every denoising step:

$$\hat{\boldsymbol{x}}_{\text{guided}} = \hat{\boldsymbol{x}}_{\text{inpaint}} - (w-1)\frac{\alpha_t}{2}\nabla_{\boldsymbol{z}_t}\left\|\boldsymbol{c} - \hat{\boldsymbol{c}}\right\|_2^2 \qquad (3)$$

where w is the guidance weight (so $w = 1$ corresponds to standard inpainting, hence the subtraction by 1, but higher values of w entail more guidance), and $\hat{\boldsymbol{c}}$ are the *predicted* start and end frames that in standard inpainting we simply throw away.

Importantly, we note that for this to be a fair comparison, we only evaluate the baseline on the 7 middle frames, i.e. we explicitly discard the first and last frame. This is of paramount importance to not evaluate on any ground truth samples and to make the FID scores comparable (so they use the same number of samples). Results are included in Fig. 5. The baseline model achieves an FID score of 60.11 when disabling inpainting, i.w., when it has no access to any original high-resolution frames.

As we hypothesized, we find that conditioning on the

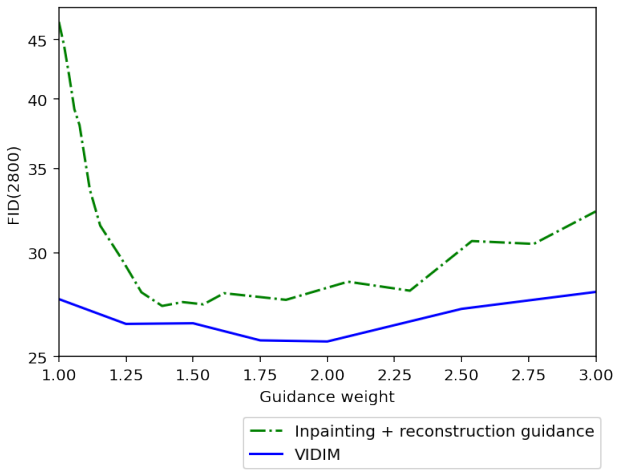

Figure 5. FID scores comparison between VIDIM and an inpainting baseline model at different guidance and reconstruction guidance weights, respectively. Note that the *reconstruction* guidance weights (x-axis) for the baseline are re-scaled via $f(w) = (w-1)/13 + 1$ to more easily compare scores at the optimal region to VIDIM; the true range for the baseline guidance weights is from 1 to 27. The baseline model achieves an FID score of 60.11.

	Davis-7					UCF101-7				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Base(lg)	24.06	0.6987	0.094	21.15	116.42	23.04	0.6967	0.0848	24.39	194.6
Base(md)	22.89	0.6529	0.1108	22.59	116.48	23.04	0.6942	0.0857	25.34	198.0
SSR(lg)	27.76	0.7825	0.1468	23.94	132.68	31.49	0.8216	0.1554	33.84	168.368
SSR(md)	28.22	0.7976	0.1309	22.11	128.6	31.48	0.8288	0.1377	32.29	145.2
Base(lg) + SSR(lg)	19.62	**0.4709**	**0.2578**	28.06	199.32	**24.07**	**0.7187**	**0.1495**	**34.48**	278
Base(lg) + SSR(md)	**20.11**	0.4632	0.3042	38.78	**196.64**	22.53	0.6775	0.2485	40.44	**263.31**
Base(md) + SSR(lg)	19.49	0.4481	0.269	**26.58**	217.14	24.1	0.7168	0.1507	35.88	280.8
Base(md) + SSR(md)	19.52	0.4327	0.3167	39.17	219.05	22.97	0.675	0.2513	42.24	279.83

Table 3. Comparison between different model sizes to illustrate the scalability of VIDIM. We compare the base and spatial super-resolution (SSR) separately and as a cascade. For each case, we consider two variants, a large model *lg* and a medium model *md*. Note that isolated models should only be compared to each other, as for the base model the ground truth is at a lower resolution, and the super-resolution models in isolation are conditioned on ground truth low-resolution frames.

high-resolution frames makes a significant difference in the quality of the results. Having access to sharp, small text helps the network preserve its legibility across the frames it generates. Even larger text, facial features, texture details, etc. can be botched without access to the high-resolution input frames. We provide a qualitative example in 4. Even without CFG, VIDIM models explicitly trained with the conditioning start and end frames achieve much better FID scores than the reconstruction guidance baselines. Interestingly, the range of "good" reconstruction guidance weights is quite different to CFG weights. Qualitatively, we find that it is key to use *some* amount of CFG, but at CFG weights of around 4.0 and above, we begin noticing significant color artifacts.

4.5. Scalability of VIDIM

Finally, we also study the effect of scaling up the number of parameters of VIDIM models. As briefly mentioned in Sec. 3.1, we only change the hidden size (a.k.a. number of channels) of the *last* UNet resolution to maximize memory savings. Because higher resolutions will have more activations *per-parameter*, increasing the width of these layers is detrimental to peak accelerator memory usage as activations must be stored for backpropagation at training time. We thus increase the hidden size of the base model only at the 16x16 resolution from 1024 to 1792 and the number of attention heads from 8 to 14, making the parameter count change from 441M to 1.6B. For our super-resolution model, the hidden size at the 16x16 resolution was increased from 1024 to 1536 and the number of attention heads from 8 to 12, making the parameter count change from 644M to 1.01B. In order to avoid memory padding in the accelerator, the number of attention heads is always set such that the per-head hidden dimension is 128. With the use of ZeRO sharding [37], both our medium and large models can be trained on accelerators with as little as 16GB of memory per chip at batch size 1 per chip, and there is enough remaining memory to maintain a ZeRO-sharded copy of the gradients to use microbatching and train on larger batch sizes with the same amount of accelerators. All our "medium" models are trained with a batch size of 256, and our "large" models were trained with twice the data, i.e., at batch size 512, for the same number of training steps as their corresponding medium-sized models. Results are included in Tab. 3.

Our quantitative results show the ability of VIDIM to favorably scale with more parameters and training: a larger base model is essential to not produce severe artifacts that will be amplified, and a larger super-resolution model is essential for sharpness in regions with the most amount of motion. Surprisingly, comparing the super-resolution models in isolation, the medium model achieves better quantitative results; however, using the large super-resolution is the most essential component to achieve low FID scores when sampling from the full cascade. Qualitatively, we see a clear difference in the samples; the large super-resolution model samples look sharper and have noticeably less artifacts. We hypothesize that the large super-resolution model has more capacity to hallucinate missing details and might be more robust than its medium counterpart to artifacts from the base model.

5. Discussion and Future Work

Through qualitative, quantitative and human evaluation, we show that our simple VIDIM models and architectures are capable state-of-the-art video interpolation, especially for large and ambiguous motion. Key components for good sample quality include explicit frame conditioning at training time and the use of classifier-free guidance. Still, several directions should be explored further. VIDIM-like models could be used for frame expansion, video restoration, among other tasks. Additionally, key problems remain to make these models maximally useful, including generating videos at arbitrary aspect ratios, and further improving the quality of super-resolution models. We also expect our architectures, framework, and demonstration of the effectiveness of image-guided conditioning will be generally useful for the community in other video generation tasks such as extrapolation, text-guided generation, and others.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 5

[2] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *ICCV*, 2007. 2

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625, 2012. 2

[5] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 1

[6] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018. 6

[7] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 5439–5448, 2017. 6

[8] Duolikun Danier, Fan Zhang, and David Bull. Stmfnet: A spatio-temporal multi-flow network for frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3521–3531, 2022. 2

[9] Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. *arXiv preprint arXiv:2303.09508*, 2023. 2, 5

[10] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 3

[11] Jiong Dong, Kaoru Ota, and Mianxiong Dong. Video frame interpolation: A comprehensive survey. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(2s), 2023. 1, 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

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

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 5, 6

[16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23 (1):2249–2281, 2022. 3, 5

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23 (1):2249–2281, 2022. 1, 2, 3

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 1, 2, 7

[19] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 3, 5

[20] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 4, 5, 6

[21] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik G. Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. *CoRR*, abs/1712.00080, 2017. 2

[22] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3, 5

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[25] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[26] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4, 5

[27] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 3

[28] Yihao Liu, Liangbin Xie, Siyao Li, Wenxiu Sun, Yu Qiao, and Chao Dong. Enhanced quadratic video interpolation. *CoRR*, abs/2009.04642, 2020. 2

[29] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer, 2022. 2

[30] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 6

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[32] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. *CoRR*, abs/2003.05534, 2020. 2

[33] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *International Conference on Computer Vision*, 2021. 2

[34] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[35] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 6

[36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[37] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 8

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[39] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 4, 5

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[42] Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv preprint arXiv:2302.07864*, 2023. 6

[43] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1, 3

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3

[45] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 6

[46] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 5

[47] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: extreme video frame interpolation. *CoRR*, abs/2103.16206, 2021. 2

[48] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[49] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[50] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

[51] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 6

[52] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014. 5

[53] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 2

[54] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[55] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 1

[56] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 2

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

[58] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 1, 3, 6

[59] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[60] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019. 2

[61] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation, 2023. 2

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[63] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 3