北京師範大學
BEIJING NORMAL UNIVERSITY

《科研写作与表达》平时作业

23/06/2025

目录

1 Homework3

1.1 Time-Adaptive Video Frame Interpolation Based on Residual Diffusion[1]

Q1. What is the problem?

Interpolation of video frames in the context of traditional hand-made animation, which exhibits particularly large variations compared to natural videos.

Q2. What have others done?

We provide extensive comparisons with respect to state-of-the-art models and show that our model outperforms these models on animation videos.

Q3. What' s the gap?

Existing diffusion-based VFI methods do not explicitly handle the interpolation time, nor do they provide uncertainty estimates to anticipate where the model may be wrong.

Q4. What have you done?

In this work, we propose a new diffusion-based method for video frame interpolation, in the context of traditional hand-made animation.

Q5. What do you contribute?

1. We explicitly handle the interpolation time in our model, which we also re-estimate during the training process, to cope with the particularly large variations observed in the animation domain.
2. We adapt and generalize a diffusion scheme called ResShift, recently proposed in the super-resolution community to VFI, which allows us to perform a very low number of diffusion steps.
3. We leverage the stochastic nature of the diffusion process to provide a pixel-wise estimate of the uncertainty on the interpolated frame.

1.2 ResShift: Efficient Diffusion Model for Image Super-Resolution by Residual Shifting[2]

Q1. What' s the problem?

Diffusion-based image super-resolution methods are mainly limited by low inference speed due to the requirements of hundreds or even thousands of sampling steps.

Q2. What have others done?

Existing acceleration sampling techniques inevitably sacrifice performance to some extent, leading to over-blurry SR results.

Q3. What' s the gap?

Post-acceleration is still needed during inference, causing performance deterioration.

Q4. What have you done?

To address this issue, we propose a novel and efficient diffusion model for SR that significantly reduces the number of diffusion steps, thereby eliminating the need for post-acceleration during inference.

Q5. What do you contribute?

1. We construct a Markov chain that transfers between the high-resolution image and the low-resolution image by shifting the residual between them, substantially improving transition efficiency.
2. We develop an elaborate noise schedule to flexibly control the shifting speed and the noise strength during the diffusion process.
3. Extensive experiments demonstrate superior or comparable performance to current state-of-the-art methods on both synthetic and real-world datasets, even with only 15 sampling steps.

1.3 A Simple Baseline for Video Restoration With Grouped Spatial-Temporal Shift[3]

Q1. What' s the problem?

Video restoration, which aims to restore clear frames from degraded videos, has numerous important applications.

Q2. What have others done?

Existing deep learning methods often rely on complicated network architectures, such as optical flow estimation, deformable convolution, and cross-frame self-attention layers.

Q3. What' s the gap?

These architectures result in high computational costs.

Q4. What have you done?

In this study, we propose a simple yet effective framework for video restoration based on grouped spatial-temporal shift.

Q5. What do you contribute?

1. By introducing grouped spatial shift, we attain expansive effective receptive fields for multi-frame aggregation.
2. Combined with basic 2D convolution, our framework can effectively aggregate inter-frame information.
3. Extensive experiments demonstrate that our framework outperforms the previous state-of-the-art method while using less than a quarter of its computational cost on both video deblurring and video denoising tasks.

1.4 Motion-Aware Latent Diffusion Models for Video Frame Interpolation[4]

Q1. What' s the problem?

With the advancement of AIGC, video frame interpolation has become a crucial component in existing video generation frameworks, attracting widespread research interest.

Q2. What have others done?

For the VFI task, the motion estimation between neighboring frames plays a crucial role in avoiding motion ambiguity.

Q3. What' s the gap?

Existing VFI methods always struggle to accurately predict motion information, leading to blurred and visually incoherent interpolated frames.

Q4. What have you done?

In this paper, we propose a novel diffusion framework, Motion-Aware latent Diffusion models, specifically designed for the VFI task.

Q5. What do you contribute?

1. We incorporate motion priors between the conditional neighboring frames and the target interpolated frame predicted throughout the diffusion sampling procedure.
2. MADIFF progressively refines intermediate outcomes, culminating in visually smooth and realistic results.
3. Extensive experiments on benchmark datasets demonstrate that our method achieves state-of-the-art performance, significantly outperforming existing approaches under challenging scenarios with dynamic textures and complex motion.

1.5 LDMVFI: Video Frame Interpolation With Latent Diffusion Models[5]

Q1. What' s the problem?

Existing works on video frame interpolation mostly employ deep neural networks trained by minimizing L1, L2, or deep feature space distance, which are poor indicators of perceptual VFI quality.

Q2. What have others done?

Recent works have shown that these metrics are poor indicators of perceptual quality.

Q3. What' s the gap?

There is a lack of perceptually-oriented VFI methods.

Q4. What have you done?

In this work, we propose latent diffusion model-based VFI, LDMVFI, approaching VFI as a conditional generation problem.

Q5. What do you contribute?

1. LDMVFI is the first effort to address VFI using latent diffusion models.
2. We rigorously benchmark our method on common test sets and conduct a user study.
3. Our experiments and user study indicate favorable perceptual quality compared to the state of the art, even in the high-resolution regime.

2 Homework4

2.1 Video Generation Documentation

To position our contributions within the rapidly evolving field of video diffusion, we conduct a systematic review of representative models, conditioning mechanisms, application scenarios, and dataset resources. Tables 1 and 2, along with Figures 1–3, summarize and categorize the key developments, revealing trends and open challenges.

2.1.1 Model Architectures and Core Techniques

Table 1 lists over thirty prominent methods, organized by input/output modalities (e.g., image-to-video, text-to-video), maximum supported resolution, and principal algorithmic components such as latent-space diffusion (L), factorized attention (FA), temporal upsampling (T), and autoregressive sampling (AR). Early works like VDM relied

表 1: Overview of mainstream video diffusion models and their core characteristics.

Model	Application	Max. Resolution	Methodology	Shots
VDM	T V L	128×128×64	FA ↑S ↑T AR	
Make-a-Video	T I V	768×768×76	FA ↑S ↑T	
ImagenVideo	T	1280×768×128	FA ↑S ↑T	
MagicVideo	T I V	1024×1024×61	P L FA ↑S ↑T	
VideoLDM	T V	2048×1280×90000	P L FA ↑S ↑T AR	
Text2Video-Zero	T V	512×512×8+	P L	
AnimateDiff	T	512×512×16	P L FA	
MCDiff	I	256×256×10	L AR	
SEINE	I	512×320×16	P L AR	
Nuwa-XL	T L	NaN×NaN×1024	P L FA ↑T	
LVDM	T L	256×256×1024	P L FA ↑T AR	
FDM	V L	128×128×15000	P FA ↑T AR	
VDT	I V	256×256×30	L FA ↑T	
Gen-L-Video	T V L	512×512×hundreds	P L 3D	
MovieFactory	T L	3072×1280×NaN	P L FA ↑S	
GLOBER	T L	256×256×128	P L 3D FA	
VideoFusion	T L	128×128×512	P ↑S AR	
GAIA-1	T I L	288×512×minutes	FA ↑T AR	
Soundini	A	256×256×NaN	P	
AADiff	I A	512×512×150	P L	
Generative Disco	A	512×512×NaN	P L	
Composable Diffusion	T I V A	512×512×16	P L FA	
Diffused Heads	A	128×128×8-9s	AR	
(Audio Heads)	A	1024×1024×NaN	P ↑S AR	
Laughing Matters	A	128×128×50	FA AR	
Dreamix	I V	1280×768×128	P FA ↑S ↑T	
Tune-A-Video	V	512×512×100	P L FA AR	
FateZero	V	512×512×100	P L FA AR	
Video-P2P	V	512×512×100	P L FA AR	
Pix2Video	V	512×512×NaN	P L FA AR	
Runway Gen-2	T I V	448×256×8	P L FA	
Make-Your-Video	V	256×256×64	P L FA AR	
Follow Your Pose	V	512×512×100	P L FA AR	

T: txt2vid, I: img2vid, V: vid2vid, A: aud2vid, L: long vid

P: pre-trained model, L: latent space, 3D: full 3D attn./conv., FA: factorized attn./conv.,

↑S: spatial upsampling, ↑T: temporal upsampling, AR: auto-regressive

on full 3D attention at modest resolutions (128×128×64), whereas more recent systems —VideoLDM and GAIA-1—push toward ultra–high-resolution and long-duration generation by combining efficient attention factorizations with latent-space representations. This progression highlights a clear community trend: hybrid paradigms that balance spatial fidelity, temporal coherence, and computational efficiency have become the de facto standard.

2.1.2 Taxonomy of Video Generation Methods

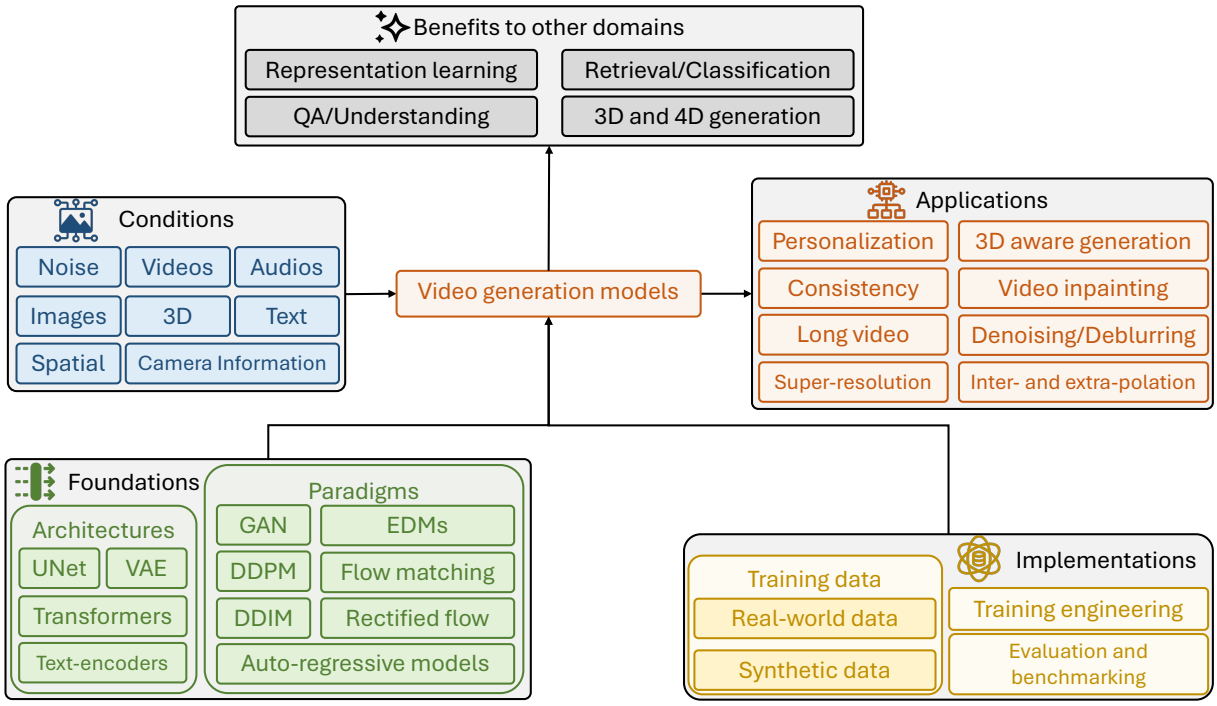


图 1: Overall framework of video generation methods.

Building on the model-level insights, Figure 1 (Overall framework of video generation methods) introduces a four-module taxonomy:

- **Conditions:** Input signals ranging from pure noise to multi-modal cues (images, video frames, audio, text, 3D point clouds).
- **Foundations:** Backbone architectures such as UNet, VAE, Transformers, and text encoders that encode and process those inputs.
- **Paradigms:** Training objectives and sampling strategies, including GANs, DDPM/EDM diffusion processes, flow-matching, and autoregressive models.
- **Applications:** Downstream tasks like personalization, consistency-aware editing, long-form synthesis, super-resolution, and video inpainting.

This modular perspective clarifies how innovations in one component (for instance, a novel attention mechanism within the Foundation) can cascade through the pipeline to enable new application scenarios.

2.1.3 Application Scenarios

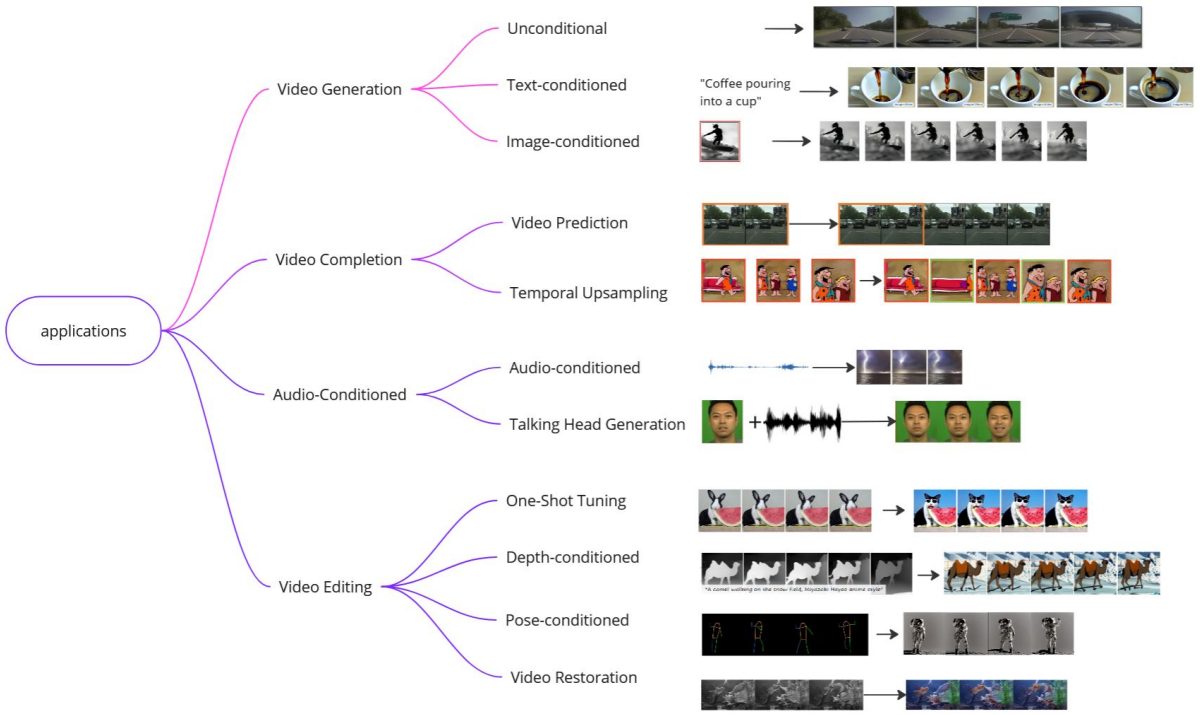


图 2: Classification of main application scenarios for video diffusion models

While the high-level framework outlines the building blocks, concrete research often targets specific applications. Figure 2 (Classification of main application scenarios for video diffusion models) partitions the field into five principal categories:

1. *Video Generation:* Unconditional, text-conditioned, and image-conditioned synthesis.
2. *Video Completion:* Frame prediction and temporal upsampling to fill missing segments.
3. *Audio-Conditioned Generation:* Talking-head and speech-driven video synthesis.
4. *Video Editing:* One-shot fine-tuning, depth- or pose-guided transformations.
5. *Video Restoration:* Denoising, deblurring, and super-resolution.

Mapping representative works onto these categories reveals that text-to-video and frame interpolation techniques have matured considerably, while high-quality, long-duration audio-driven generation and robust one-shot editing remain active research frontiers.

2.1.4 Image-based Conditioning Mechanisms

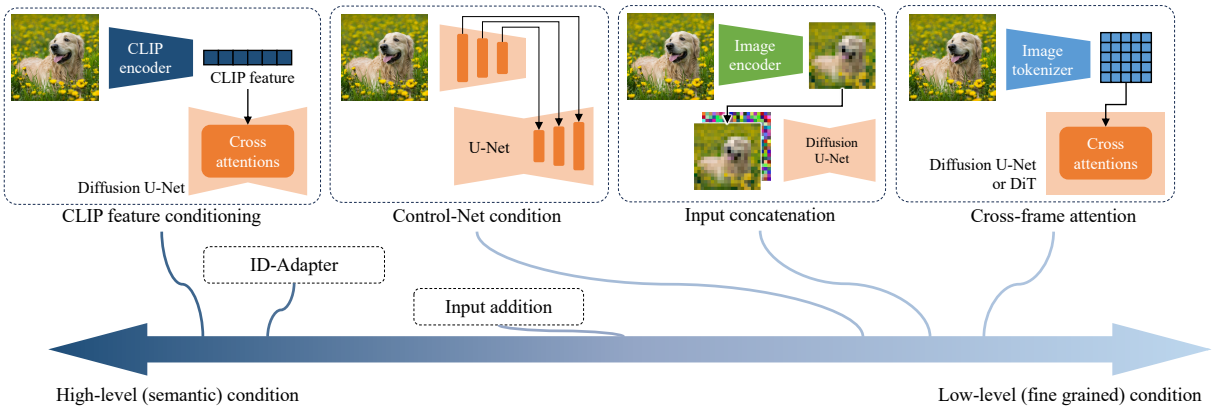


图 3: Methods for controlling video generation using image-based conditions. We categorize these methods along a spectrum ranging from high-level semantic conditions to low-level conditions. Additionally, we illustrate four classic approaches commonly used to condition input images.

Conditioning mechanisms critically influence output quality when using reference images. Figure 3 (Classification of image-based conditioning methods for video generation) arranges four canonical strategies along a spectrum from high-level semantic control to low-level fine-grained guidance:

- **CLIP-feature conditioning:** Injects semantic cues via cross-attention using pretrained vision–language embeddings.
- **Control-Net injection:** Integrates adapter modules within the UNet backbone for feature-space control.
- **Input concatenation:** Appends conditioning image channels to the UNet input for early-fusion guidance.
- **Cross-frame attention:** Propagates spatial details across adjacent frames to reinforce temporal coherence.

This categorization underscores the trade-off between flexibility of high-level control and the preservation of fine visual detail.

2.1.5 Dataset Resources

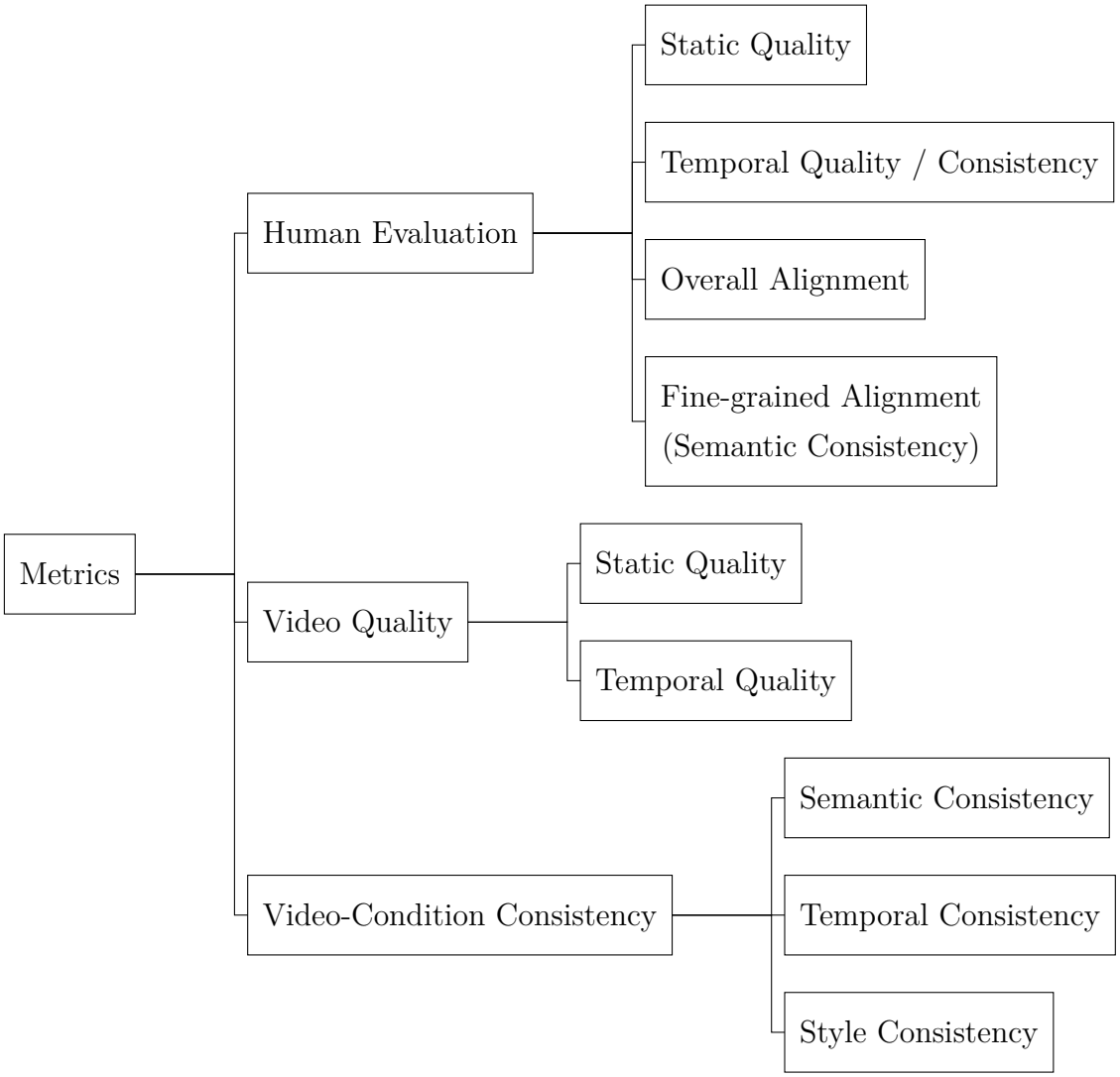
Finally, robust training and meaningful evaluation hinge on appropriate datasets. Table 2 (Summary of commonly used datasets for video generation) collates over twenty

Datasets	Modalities	Vision	Text	Duration	Resolution	Domains	# of samples	Sources	License
Academia datasets									
UCF-101 [6]	V, A	Real	-	AVG. 7 s	320x240	-	V (13,320)	YouTube	CC-BY
Kinetics-400 [7]	V, A	Real	-	AVG. 10 s	-	-	V (306,245)	YouTube	-
BSCV [8]	V	Real	-	-	480P	-	V (28 k)	-	-
VFHQ [9]	V	Real	-	-	700x700	-	V (16,827)	FFHQ, VoxCeleb1	-
DMLab-40k [10]	V	Synthetic	-	-	-	3D Rendered	V (40 k)	-	-
Habitat [10]	V	Synthetic	-	-	-	3D Rendered	V (200 k)	-	-
Minecraft [10]	V	Synthetic	-	-	-	3D Rendered	V (200 k)	-	-
DEVIL [11]	V	Real	-	-	-	Camera Infos	V (1,250)	Flickr	MIT
Inter4K-1k [12]	V	Real	-	AVG. 5 s	4K	-	V (1 k)	-	-
MSR-VTT [13]	V, T, A	Real	Real	AVG. 15 s	320x240	-	T/V (10 k)	Youtube	-
DiDeMo [14]	V, T, A	Real	Real	AVG. 7 s	-	-	T/V (27 k)	Flickr	BSD 2-Clause
LSMDC [15]	V, T, A	Real	Real	AVG. 5 s	1080p	-	T/V (118 k)	-	-
VATEX [16]	V, T, A	Real	Real	AVG. 10 s	-	-	T/V (41 k)	Youtube	CC-BY-4.0
YouCook2 [17]	V, T, A	Real	Real	AVG. 5 mins	-	-	T/V (14 k)	Youtube	-
How2 [18]	V, T, A	Real	Real	AVG. 90 s	-	-	T/V (79 k)	Youtube	Creative Commons BY-SA 4.0
ActivityNet Caption [19]	V, T, A	Real	Real	AVG. 120 s	-	-	T/V (100 k)	Youtube	-
VideoCC3M [20]	V, T, A	Real	Real	AVG. 10 s	-	-	T/V (10.3 M)	-	-
WebVid10M [21]	V, T	Real	Real	AVG. 18 s	360p	-	T/V (10.7 M)	-	AGPL-3.0
WTS70M [22]	V, T	Real	Real	AVG. 10 s	-	-	T/V (70 M)	-	-
HowTo100M [23]	V, T, A	Real	Auto Captioning	AVG. 4 s	240p	-	T/V (136 M)	Youtube	Apache License 2.0
HD-VILA-100M [24]	V, T	Real	Auto Captioning	AVG. 13 s	720p	-	T/V (100 M)	-	See license
YT-Temporal-180M [25]	V, T, A	Real	Real	-	-	-	T/V (180 M)	Youtube	-
ACAV100M [26]	V, T, A	Real	Real	AVG. 10 s	-	-	T/V (100 M)	-	MIT License
Vript-400k [27]	V, T, A	Real	Real	AVG. 11 s	720p	-	T/V (420 k)	-	-
VidProdM [28]	V, T	Synthetic	Real	AVG. 2 s	-	-	T (1.67 M), V (6.69 M)	-	CC-BY-NC 4.0
FETV [29]	V, T	Real	Real	-	-	-	T (619), V (541)	MSR-VTT and WebVid	CC-BY-NC 4.0
InternVid [30]	V, T	Real + Synthetic	Auto Captioning	AVG. 12 s	720p	-	T/V (234 M)	-	CC BY-NC-SA 4.0
AIGCBench [31]	I, V, T	Real	Real	-	-	-	T/I (2,928), T/V (1,000)	-	Apache License 2.0
AVE [32]	V, T	Real	Real	AVG. 4 s	-	-	V (196k)	-	-
Panda-70M [33]	V, T	Real	Auto Captioning	AVG. 9 s	720p	-	T/V (70 M)	-	See license
HD-VG-130M [34]	V, T	Real	Auto Captioning	-	720p	-	T/V (130 M)	-	See license
MiraData-77k [35]	V, T	Real	Auto Captioning	AVG. 72 s	720p	Camera Infos	V (330 k)	-	GPL-v3
LAION-AESTHETICS 6.5+	I, T	Synthetic	Real	-	-	-	T/I (625 k)	-	-
Animate bench	I, T	Synthetic	Real	-	-	-	T/I (105)	-	Apache License 2.0
DrawBench [36]	I, T	Real	Real	-	-	-	I/T (200)	-	-
PartiPrompts [37]	I, T	Real	Real	-	-	-	T (1.6 k)	-	Apache License 2.0
Commercial datasets									
Midjourney-v5-1.7M	I, T	Synthetic	Real	-	-	-	T/I (1.7 M)	-	Apache License 2.0
Midjourney-Kaggle-Clean	I, T	Synthetic	Real	-	-	-	T/I (250 k)	-	cc0-1.0
Unsplash-lite	I, T	Synthetic	Real	-	-	-	T/I (250 k)	-	See license
Mixkit	V, T	Real	Real	AVG. 18 s	720p	-	T/V (1,234)	-	Commercial and non-commercial
Pixabay	I, V, T	Real	Real	AVG. 25 s	720p	-	T/V (31,616)	-	Commercial and non-commercial
Pexels-400k	V, T	Real	Real	-	720p	-	T/V (400,476)	-	MIT

表 2: Summary of commonly used datasets for video generation. We also include image datasets as they are usually used in training. “I”, “V”, “T”, and “A” represent image, video, text, and audio. Other commercial datasets include those released by Pond5, Adobe Stock, Shutterstock, Getty, Coverr, Videvo, Depositphotos, Storyblocks, Dissolve, Freepik, Vimeo, and Envato.

academic and commercial sources, detailing each dataset ’ s modalities (image, video, text, audio), average duration, resolution, sample count, licensing terms, and domain focus. The surveyed datasets span controlled synthetic simulations to large-scale web-scraped video corpora, providing a spectrum of complexity that guides model selection and benchmarking. Insights from this overview point to the need for richer, more diverse benchmarks to further drive innovation in the field.

2.1.6 Evaluation Metrics for Video Diffusion Models



Effective evaluation of video diffusion models demands a multi-faceted metric suite that captures alignment with conditioning inputs, low-level visual fidelity, and human perceptual judgment. Figure above presents our proposed taxonomy, which divides metrics into three primary categories:

- **Video-Condition Consistency** measures how faithfully the generated video follows the given conditioning signal. It comprises:
  - *Semantic Consistency*: the degree to which high-level content (objects, scene layout, actions) matches the reference input.
  - *Temporal Consistency*: the smoothness of frame-to-frame transitions, often quantified by temporal LPIPS or frame-difference statistics.
  - *Style Consistency*: preservation of visual style attributes (color palette, texture) across all frames.
- **Video Quality** evaluates the intrinsic visual and motion fidelity of the output:

  - *Static Quality*: per-frame image metrics such as PSNR, SSIM, and LPIPS.
  - *Temporal Quality*: sequence-level measures like FVD or T-PSNR that capture motion accuracy and temporal artifacts.

- **Human Evaluation** provides subjective assessments that often correlate better with perceived realism:
  - *Static Quality*: human ratings of individual frame sharpness and realism.
  - *Temporal Quality / Consistency*: perceived smoothness and coherence of motion.
  - *Overall Alignment*: holistic scoring of how well the video matches the intended concept or condition.
  - *Fine-grained Alignment (Semantic Consistency)*: detailed judgments on specific semantic attributes, such as object identity, pose accuracy, or action correctness.

By combining automated consistency and quality measures with targeted human studies, this taxonomy ensures a comprehensive assessment of video diffusion performance.

2.2 Excerpts of "Related Work" sections

2.2.1 Time-Adaptive Video Frame Interpolation Based on Residual Diffusion[1]

> "For a few years now, deep learning–based methods have formed the bulk of the state of the art in video frame interpolation (VFI), typically using convolutional encoder–decoder architectures to predict intermediate frames. Deep Voxel Flow (DVF) interpolates pixels by learning a 3D voxel flow layer across space and time, while BiPN jointly predicts forward and backward frames via a bidirectional encoder–decoder. PhaseNet decomposes frames into steerable-pyramid phase and amplitude coefficients and learns a levelwise decoder to reconstruct the middle image. Other works model motion and occlusion by linearly combining bidirectional optical flows and refining them with soft visibility maps, and DAIN uses depth-aware flow projection to handle occlusions by sampling closer objects more densely. Separately, separable convolution methods estimate pixelwise convolution kernels (soft-splatting) to capture motion and re-sampling in a unified framework, later extended with adaptive deformable kernels. More recently, generative approaches such as FIGAN

and latent diffusion–based LDMVFI treat interpolation as a conditional image generation problem. In this work, we adapt the ResShift diffusion scheme from super-resolution to VFI, explicitly re-estimating interpolation time during training to handle large temporal variations in animation and leveraging the stochastic diffusion process to quantify per-pixel uncertainty."

2.2.2 ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting[2]

"Diffusion-based image super-resolution methods typically inherit DDPM' s long Markov chains, requiring hundreds or thousands of sampling steps and resulting in slow inference. Common acceleration techniques (e.g., DDIM) compress sampling steps but often produce over-smooth images. To address this, ResShift constructs a Markov chain that transfers between the high-resolution and low-resolution images by iteratively shifting their residual, dramatically improving transition efficiency. An elaborate noise schedule flexibly controls shift speed and noise strength. Extensive experiments show that ResShift achieves state-of-the-art performance with as few as 15 sampling steps on both synthetic and real datasets, eliminating the need for post-acceleration and its associated performance degradation."

2.2.3 A Simple Baseline for Video Restoration With Grouped Spatial-Temporal Shift[3]

"Video restoration hinges on effectively aggregating inter-frame information, yet many deep-learning solutions employ complex alignment modules—optical flow estimation, deformable convolution, or cross-frame self attention—that incur high computation and can fail under large displacement, noise, or blur. To simplify this, we propose grouped spatial-temporal shift blocks, which implicitly capture correspondence by shifting feature groups across time and space, followed by simple 2D convolution fusion. This lightweight approach achieves large effective receptive fields for multi-frame aggregation and, when stacked, models long-term dependencies without costly flow or attention modules. Empirical results on deblurring and denoising benchmarks demonstrate that Group Shift-Net outperforms prior methods while using less than one-quarter of their FLOPs."

2.2.4 Motion-Aware Latent Diffusion Models for Video Frame Interpolation[4]

"Existing VFI methods struggle with accurately predicting inter-frame motion, leading to blurred or inconsistent interpolations, particularly in dynamic scenes. Early deep models rely on 3D convolutions or RNNs to fuse adjacent frames, while GAN-based approaches optimize perceptual realism at the risk of texture artifacts. Recent works have applied diffusion models to VFI as conditional image generation, yet they neglect explicit motion modeling between the target frame and its neighbors. We introduce MADIFF, a motion-aware latent diffusion framework that injects motion priors—extracted via an image-to-event (I2E) generator—into both the VQ-MAGAN decoder and the reverse diffusion network. Furthermore, our MA-SAMPLING scheme uses coarse interpolations from previous steps to continually extract and refine motion hints, yielding smooth, realistic frames and achieving state-of-the-art performance under complex motion."

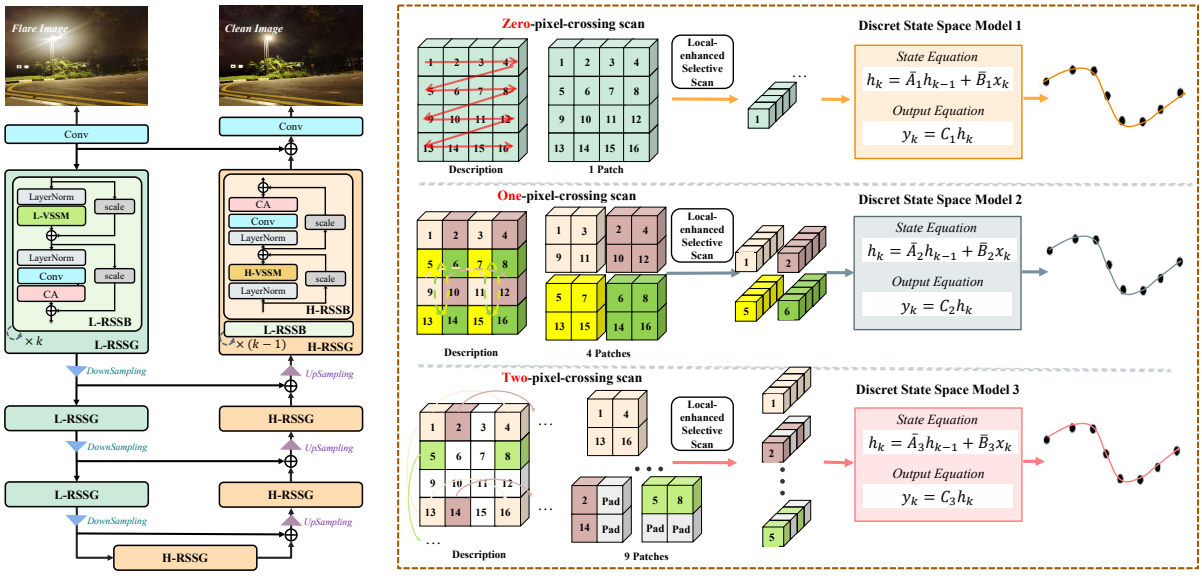
2.2.5 LDMVFI: Video Frame Interpolation with Latent Diffusion Models[5]

"Conventional VFI approaches—whether flow-based (optical flow estimation and warping) or kernel-based (adaptive convolution)—are optimized with L1/L2 or VGG feature losses that correlate poorly with perceptual quality, often failing on dynamic textures. Denoising diffusion probabilistic models (DDPMs) have recently excelled at realistic image synthesis but remained unexplored for VFI. LDMVFI formulates interpolation as conditional latent diffusion: an autoencoder (VQ-FIGAN) projects frames into a compact latent space enriched by cross-attention among neighboring features, and a denoising U-Net performs reverse diffusion to generate intermediate latents. This paradigm shift yields superior perceptual fidelity, as confirmed by quantitative metrics and a user study, demonstrating that latent diffusion can serve as a new, perceptually-oriented VFI paradigm."

3 Homework5

In this project, I implemented an image processing framework named **Deflare-Mamba**. The framework is designed primarily to eliminate flares and ghosting in images and utilizes a multi-scale network architecture to enhance restoration quality. Below, the "Method Diagram" and the "Application" are presented.

3.1 Method Diagram



(a) Pipeline of DeflareMamba

(b) Hierarchical Selective Scan strategy in H-RSSB

图 4: Overview of our DeflareMamba framework. The network adopts a U-shaped architecture with Local-enhanced Residue State Space Groups in the encoding stage and Hierarchical Residue State Space Groups in the decoding stage.

3.2 Application

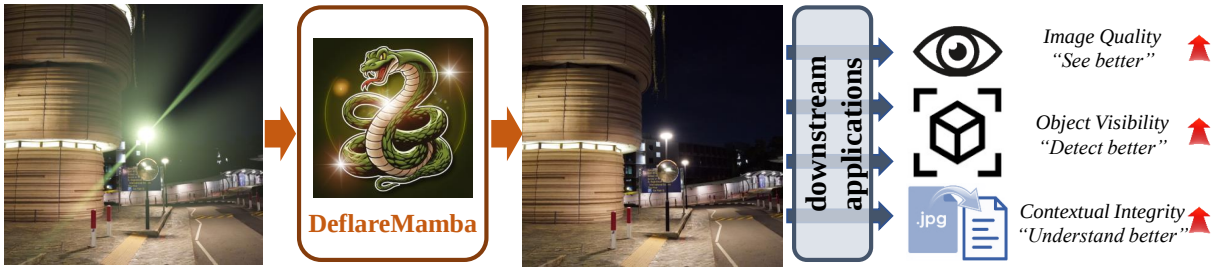


图 5: DeflareMamba effectively enhances the quality of image media, thereby enabling downstream applications that benefit from improved image quality, increased object visibility, and enhanced contextual integrity.

3.3 Experimental Results Comparison Using Table

表 3: Quantitative Comparison with State-of-the-art Methods

Metric	Network trained on Flare7K++						
	Input	U-Net[38]	HINet[39]	MPRNet*[40]	Restormer*[41]	Uformer[42]	DeflareMamba (C
PSNR↑	22.561	27.189	27.548	27.036	27.597	27.633	**27.778**
SSIM↑	0.857	0.894	0.892	0.893	0.897	0.894	**0.899**

表 4: Object Detection Performance (mAP)

Data	Faster R-CNN[43]	Deformable Det[44]	YOLO8x[45]
w/ flare	28.15	36.20	38.15
Uformer	29.13	37.35	39.04
DeflareMamba	**29.35**	**37.59**	**39.36**

表 5: Vision-Language Alignment Performance

Data	CLIP[46]	BLIP[47]			
	CS	CS	ITM	TR@1	TR@5
w/ flare	28.57	45.31	117.56	72.2	90.1
Uformer	29.14	46.37	127.79	73.1	91.0
DeflareMamba	**29.22**	**46.50**	**128.45**	**73.4**	**91.5**

3.4 formula expression

The following is a idea I recently developed in the field of video frame interpolation. I have extracted its core formulas and presented them in LaTeX format in order to achieve a high-quality completion of this section of the assignment.

3.4.1 Symbols and Assumptions

- **Known endpoints**: the starting frame x_0 and the ending frame x_T.
- **Optical flow fields**:
  - $F_{0\to T}$: the flow from x_0 to x_T;
  - $F_{T\to 0}$: the reverse flow.
- **Warp operators (linear)**:

$$W_f[x] = \mathrm{warp}\big(x, \tfrac{1}{T}\, F_{0\to T}\big), \quad W_b[x] = \mathrm{warp}\big(x, \tfrac{T-t}{T}\, F_{T\to 0}\big).$$

- **Mixing weight**: $\lambda_t \in [0,1]$, typically $\lambda_t = \frac{t}{T}$.
- **Noise covariance**: at forward step t we add isotropic Gaussian noise with $\Sigma_t = \beta_t I$.
- **Occlusion mask** $m \in \{0,1\}^{H\times W}$: a pixel-wise indicator of whether to use the forward or backward warp.

3.4.2 Forward Distribution $q(x_t \mid x_{t-1}, x_T)$

At step t, define the conditional mean

$$\mu_t(x_{t-1}, x_T) = (1-\lambda_t)\left[m \odot W_f[x_{t-1}]\right] + \lambda_t\left[(1-m)\odot W_b[x_T]\right],$$

then

$$\boxed{q(x_t \mid x_{t-1}, x_T) = \mathcal{N}\big(x_t;\, \mu_t(x_{t-1}, x_T),\, \beta_t I\big).}$$

- Here $m \odot W_f[x_{t-1}]$ means warp first, then element-wise mask.
- Likewise, $(1-m)\odot W_b[x_T]$ is defined analogously.
- **Bidirectional fusion**: each mean uses both the previous-frame warp and the final-frame warp.

3.4.3 Prior Distribution $q(x_{t-1} \mid x_0, x_T)$

At time $t-1$, given the endpoints, the prior is

$$\mu_{t-1} = (1-\lambda_{t-1})\,[\,m \odot W_f[x_0]\,] + \lambda_{t-1}\,[\,(1-m) \odot W_b[x_T]\,],$$

$$\boxed{q(x_{t-1} \mid x_0, x_T) = \mathcal{N}\big(x_{t-1};\, \mu_{t-1},\, \beta_{t-1} I\big).}$$

3.4.4 Joint Log-Density

Using

$$q(x_{t-1}, x_t \mid x_0, x_T) = q(x_t \mid x_{t-1}, x_T)\; q(x_{t-1} \mid x_0, x_T),$$

we get (up to constants independent of x_{t-1}):

$$\ln q \;\propto\; -\tfrac{1}{2}\,(x_t - \mu_t)^\top (\beta_t I)^{-1}(x_t - \mu_t) \;-\; \tfrac{1}{2}\,(x_{t-1} - \mu_{t-1})^\top (\beta_{t-1} I)^{-1}(x_{t-1} - \mu_{t-1}).$$

Expand μ_t as a linear function:

$$\mu_t(x_{t-1}, x_T) = A\,x_{t-1} + b, \quad A = (1-\lambda_t)\,M_f, \quad b = \lambda_t\,M_b[x_T],$$

where

$$M_f[x_{t-1}] = m \odot W_f[x_{t-1}], \quad M_b[x_T] = (1-m) \odot W_b[x_T].$$

Hence

$$x_t - \mu_t = x_t - A\,x_{t-1} - b.$$

3.4.5 Quadratic Form Expansion

Combine the two terms and collect quadratic and linear parts in x_{t-1}:
Likelihood Term:

$$-\tfrac{1}{2}\,(x_t - Ax_{t-1} - b)^\top \tfrac{1}{\beta_t} I\,(x_t - Ax_{t-1} - b) = -\tfrac{1}{2\beta_t}\big[x_{t-1}^\top A^\top A\,x_{t-1} - 2\,x_{t-1}^\top A^\top (x_t - b)\big] + \text{const.}$$

Prior Term:

$$-\tfrac{1}{2}\,(x_{t-1} - \mu_{t-1})^\top \tfrac{1}{\beta_{t-1}} I\,(x_{t-1} - \mu_{t-1}) = -\tfrac{1}{2\beta_{t-1}}\big[x_{t-1}^\top x_{t-1} - 2\,x_{t-1}^\top \mu_{t-1}\big] + \text{const.}$$

Summing gives

$$\ln q(x_{t-1} \mid x_t, \dots) \propto -\tfrac{1}{2}\,x_{t-1}^\top H\,x_{t-1} \;+\; x_{t-1}^\top h,$$

$$\boxed{H = \tfrac{1}{\beta_t}\,A^\top A + \tfrac{1}{\beta_{t-1}}\,I,} \qquad \boxed{h = \tfrac{1}{\beta_t}\,A^\top (x_t - b) + \tfrac{1}{\beta_{t-1}}\,\mu_{t-1}.}$$

3.4.6 Completing the Square for Posterior Parameters

By completing the square, the posterior is

$$q(x_{t-1} \mid x_t, x_0, x_T) = \mathcal{N}\big(x_{t-1};\, \tilde{\mu},\, \tilde{\Sigma}\big),$$

with

$$\boxed{\tilde{\Sigma} = H^{-1}, \qquad \tilde{\mu} = H^{-1}\, h.}$$

3.4.7 Element-wise Approximation and Physical Interpretation

Assume at the pixel level that $M_f^\top M_f \approx I$ and $M_f^\top M_b = 0$. Then:
*Posterior Covariance

$$\tilde{\Sigma} = \Big(\tfrac{(1-\lambda_t)^2}{\beta_t} I + \tfrac{1}{\beta_{t-1}} I\Big)^{-1} = \frac{\beta_t\, \beta_{t-1}}{(1-\lambda_t)^2\, \beta_{t-1} + \beta_t}\, I.$$

*Posterior Mean

$$\begin{aligned}\tilde{\mu} &= \tilde{\Sigma}\Big[\tfrac{1-\lambda_t}{\beta_t} M_f^\top (x_t - b) + \tfrac{1}{\beta_{t-1}}\, \mu_{t-1}\Big] \\ &\approx \underbrace{\tfrac{\beta_{t-1}}{(1-\lambda_t)^2\beta_{t-1}+\beta_t}}_{\rho}\, \mu_{t-1} + \underbrace{\tfrac{(1-\lambda_t)\,\beta_t}{(1-\lambda_t)^2\beta_{t-1}+\beta_t}}_{1-\rho}\, \underbrace{M_f^\top (x_t - b)}_{\substack{\text{observation}\\ \text{correction}}}^{\text{feedback}}.\end{aligned}$$

- μ_{t-1} is the prior interpolation from bidirectional endpoint warps.
- $M_f^\top (x_t - b)$ subtracts the backward term $b = \lambda_t M_b[x_T]$ from x_t and backward-warps along the forward flow for correction.
- ρ and $1-\rho$ are weights determined by the noise variances β_t, β_{t-1}.

3.4.8 Final Conclusion

$$\boxed{q(x_{t-1} \mid x_t, x_0, x_T) = \mathcal{N}\Big(x_{t-1};\, \tilde{\mu},\, \tilde{\Sigma}\Big),}$$

$$\tilde{\Sigma} = \frac{\beta_t\, \beta_{t-1}}{(1-\lambda_t)^2\, \beta_{t-1} + \beta_t}\, I, \quad \tilde{\mu} = \frac{\beta_{t-1}}{(1-\lambda_t)^2 \beta_{t-1} + \beta_t}\, \mu_{t-1} + \frac{(1-\lambda_t)\, \beta_t}{(1-\lambda_t)^2 \beta_{t-1} + \beta_t} M_f^\top \big(x_t - \lambda_t M_b[x_T]\big).$$

- **Bidirectional warp**: $\mu_{t-1} = (1-\lambda_{t-1})\, m \odot W_f[x_0] + \lambda_{t-1}\, (1-m) \odot W_b[x_T]$.
- **Observation correction**: $M_f^\top (x_t - \lambda_t M_b[x_T])$ adjusts the prior only in the forward-visible region.
- **Variance-based weights**: ensure the KL-minimization training objective yields the correct parameters.

Validation and Consistency

1. **During training:** $\tilde{\mu}$ is used as the regression target; the model learns parameters to match it.
2. **During sampling:** $\tilde{\mu}$ and $\tilde{\Sigma}$ guide the stepwise reverse sampling, ensuring endpoint alignment with dynamic feedback.
3. **Video frame interpolation:** endpoint conformity, motion coherence, and reasonable occlusion all follow from these consistent assumptions and derivations.

4 Homework6

4.1 Contrast

Time-adaptive Video Frame Interpolation based on Residual Diffusion

(1) "However, this approach did not perform as well as we expected."

(2) "However, this causes a decrease in the realism of the restored HR images. To compromise, the hyper-parameter p can be set to a relatively small value."

ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting

(1) "However, it is worth noting that we retained the sampling steps of 100 for IRSDE, consistent with its configuration in training, since we empirically found that accelerating the inference of IRSDE led to a severe performance drop.

4.2 Addition / Progression

ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting

(1) "Additionally, such a design can reduce the number of diffusion steps required for sampling, thereby improving inference efficiency."

(2) "Moreover, when κ is in the range of [1.0, 2.0], our method achieves the most realistic quality indicated by CLIPIQA and MUSIQ, which is more desirable in real applications."

Time-adaptive Video Frame Interpolation based on Residual Diffusion

(1) "Furthermore, in terms of MUSIQ, our approach achieves comparable performance with recent SotA methods."

4.3 Continuation

Time-adaptive Video Frame Interpolation based on Residual Diffusion

(1) "Following the aforementioned motivation, we propose an efficient diffusion model

involving a shorter Markov chain for transitioning between the HR image and its corresponding LR one."

(2) "Hence, in this work, we introduce the metric τ_{IFD}. Its logic is to generate a value that describes the amount of movement from $I_0 \to I_\tau$ and from $I_1 \to I_\tau$ in critical areas, i.e., areas with significant movement and change."

ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting

(1) "Therefore, we set $T = 15$ and $p = 0.3$, and yield our model named ResShift."

参考文献

[1] Victor Fonte Chavez, Claudia Esteves, and Jean-Bernard Hayet. Time-adaptive video frame interpolation based on residual diffusion. *arXiv preprint arXiv:2504.05402*, 2025.

[2] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36:13294–13307, 2023.

[3] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023.

[4] Zhilin Huang, Yijie Yu, Ling Yang, Chujun Qin, Bing Zheng, Xiawu Zheng, Zikun Zhou, Yaowei Wang, and Wenming Yang. Motion-aware latent diffusion models for video frame interpolation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1043–1052, 2024.

[5] Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1472–1480, 2024.

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

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[8] Tianyi Liu, Kejun Wu, Yi Wang, Wenyang Liu, Kim-Hui Yap, and Lap-Pui Chau. Bitstream-corrupted video recovery: A novel benchmark dataset and method. In *NeurIPS*, 2023.

[9] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *CVPR Workshops*, pages 656–665, 2022.

[10] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. In *ICML*, pages 39062–39098, 2023.

[11] Ryan Szeto and Jason J. Corso. The devil is in the details: A diagnostic evaluation benchmark for video inpainting. In *CVPR*, pages 21022–21031, 2022.

[12] Alexandros Stergiou and Ronald Poppe. Adapool: Exponential adaptive pooling for information-retaining downsampling. *IEEE Trans. Image Process.*, 32:251–266, 2023.

[13] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[14] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5804–5813, 2017.

[15] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, pages 3202–3212, 2015.

[16] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4580–4590, 2019.

[17] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

[18] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding, 2018.

[19] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[20] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13674, pages 407–426, 2022.

[21] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[22] Jonathan C. Stroud, Zhichao Lu, Chen Sun, Jia Deng, Rahul Sukthankar, Cordelia Schmid, and David A. Ross. Learning video representations from textual web supervision, 2021.

[23] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[24] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, pages 5026–5035, 2022.

[25] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, pages 23634–23651, 2021.

[26] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *ICCV*, pages 10254–10264, 2021.

[27] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In *NeurIPS*, 2024.

[28] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *NeurIPS*, 2024.

[29] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *NeurIPS*, 2023.

[30] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2023.

[31] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai, 2024.

[32] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13668, pages 201–218, 2022.

[33] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331, 2024.

[34] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap attention in spatiotemporal diffusions for text-to-video generation, 2024.

[35] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *NeurIPS*, 2024.

[36] 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. In *NeurIPS*, pages 36479–36494, 2022.

[37] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[38] 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.

[39] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2021.

[40] Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2704–2713, 2021.

[41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[42] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.

[43] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. 2021.

[45] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[47] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.