GenRec: Unifying Video Generation and Recognition with Diffusion Models

Zejia Weng[1,2], **Xitong Yang**[3], **Zhen Xing**[1,2], **Zuxuan Wu**[1,2†], **Yu-Gang Jiang**[1,2]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3] Department of Computer Science, University of Maryland

Abstract

Video diffusion models are able to generate high-quality videos by learning strong spatial-temporal priors on large-scale datasets. In this paper, we aim to investigate whether such priors derived from a generative process are suitable for video recognition, and eventually joint optimization of generation and recognition. Building upon Stable Video Diffusion, we introduce GenRec, the first unified framework trained with a random-frame conditioning process so as to learn generalized spatial-temporal representations. The resulting framework can naturally supports generation and recognition, and more importantly is robust even when visual inputs contain limited information. Extensive experiments demonstrate the efficacy of GenRec for both recognition and generation. In particular, GenRec achieves competitive recognition performance, offering 75.8% and 87.2% accuracy on SSV2 and K400, respectively. GenRec also performs the best on class-conditioned image-to-video generation, achieving 46.5 and 49.3 FVD scores on SSV2 and EK-100 datasets. Furthermore, GenRec demonstrates extraordinary robustness in scenarios that only limited frames can be observed. Code will be available at https://github.com/wengzejia1/GenRec.

1 Introduction

Diffusion models have achieved significant success in the field of image and video generation over the past few years. A variety of generative tasks have been revolutionized by using diffusion models trained on Internet-scale data, such as text-to-image generation [33, 30], image editing [23], and more recently, text-to-video generation [15, 2, 52] and text&image-to-video generation [56, 18, 21]. The excellent generative capabilities of diffusion models suggest that informative representation is learned during the generative training and strong visual priors are captured by the backbone models [9, 43, 6]. Therefore, recent work has explored leveraging the image diffusion models for image understanding tasks, including image recognition [9, 8], object detection [7, 57], segmentation [55] and correspondence mining [39]. However, the capability of *video diffusion* models to effectively capture spatial-temporal information is not fully understood, and their potential for downstream video understanding tasks remains under-explored.

In this paper, we study the potential of video diffusion models [28, 1, 27], particularly the unconditioned or image-conditioned models, for video understanding by addressing the three key problems: (a) Does the backbone model trained for video generation extract effective spatial-temporal representations for semantic video recognition? (b) Can we retain the video generation capability by jointly optimizing generation and recognition? (c) Will such a unified training framework further benefit video understanding, especially in noisy scenarios where only limited frames are available [3, 25].

[†]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3456

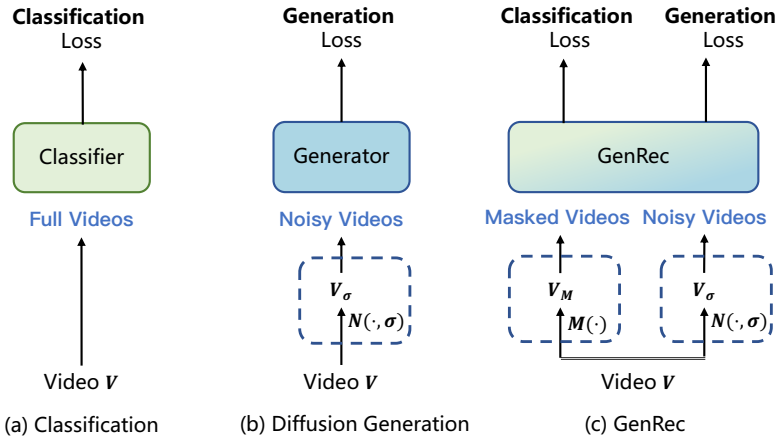

Figure 1: Comparison of classical pipelines for video classification and generation tasks with our proposed GenRec method. **(a) Classification**: Typical video classification focus on understanding complete videos. **(b) Diffusion Generation**: Diffusion models learn the noise reduction trajectory from videos with varying levels of noise. These two distinct training paradigms present challenges for task unification. To bridge this gap, we propose **(c) GenRec**: a learning framework that processes mask frames V_M using a masking function $M(\cdot)$ and noise videos V_σ with noise sampling $\mathcal{N}(\cdot, \sigma)$, aiming to simultaneously learn video understanding and content completion with the same partially observed visual content.

While conceptually appealing, unifying video generation and recognition into a diffusion framework is non-trivial. Prior work either views the diffusion models as frozen feature extractors [9, 55, 39], or deconstructs them for new tasks while sacrificing their original generation capability [8]. One major challenge comes from their distinct training and inference processes. Diffusion models are typically optimized using corrupted inputs, optionally augmented with a single conditioning frame, to achieve unconditioned or image-conditioned generation during inference [27, 1]. In contrast, video recognition models require access to multiple frames to reason about temporal relationships and expect clean inputs during inference [49, 51, 48]. Consequently, training a recognition model using corrupted videos and single-image conditions tends to suffer from inferior model optimization and a more significant training-inference gap.

To this end, we propose GenRec, a unified video diffusion model that enables joint optimization for video generation and recognition. Our model is built upon the open-source, image-conditioned Stable Video Diffusion model (SVD) [1], which encodes strong spatial-temporal priors by pretraining on large-scale image and video data. However, instead of conditioning on the same image across all video frames, we propose to condition on a random subset of frames while masking the remaining ones (see Figure 2). This simple random-frame conditioning process effectively bridges the gap between the learning processes of the two tasks. On the one hand, the generation capability of SVD is extended to handle arbitrary frame prediction, which provides more flexible and unambiguous video generation. On the other hand, conditioning on a random subset of frames allows the model to learn more discriminative and robust features for the recognition task. As shown in Figure 1, the model is jointly optimized using both generative supervision (*i.e.*, noise prediction) and classification supervision.

We conduct extensive experiments to evaluate the performance of GenRec for both recognition and generation. Without sacrificing the generation capabilities, GenRec demonstrates competitive video recognition performance, offering 75.8% and 87.2% accuracy on SSV2 and K400, respectively. Furthermore, GenRec demonstrates extraordinary robustness in scenarios that only limited frames can be observed. For example, when only the front half of the video can be observed, GenRec achieves the 57.7% accuracy, which corresponds to 76.6% of the accuracy (75.3%) when the entire video is visible, emonstrating a higher accuracy retention ratio than other methods. By leveraging the recognition model for classifier guidance [11], GenRec also achieves superior class-conditioned image-to-video generation results, with FVD scores of 46.5 and 49.3 on the SSV2 and EK-100 datasets, respectively.

2 Preliminary

Representing the data distribution as $p_{\text{data}}(\mathbf{z})$ with a standard deviation of σ_{data}, we can obtain a family of smoothed distributions $p(\mathbf{z};\sigma)$ by adding independent and identically distributed Gaussian noise with standard deviation σ. In the spirit of diffusion models, the generation process begins with a noise image $\mathbf{z}_N \sim N(0, \sigma_{\text{max}}^2\mathbf{I})$ and iteratively denoises it at decreasing noise levels $\sigma_N = \sigma_{\text{max}} > \sigma_{N-1} > \ldots > \sigma_0 = 0$. The final denoised result $\mathbf{z}_0$ is thus distributed according to the original data.

In the EDM [22] framework, the original $\mathbf{z}_0$ will be diffused as:

$$\mathbf{z}_\sigma = \mathbf{z}_0 + \sigma \cdot N(0, \mathbf{I}), \tag{1}$$

and the corresponding PF-ODE [34] follows:

$$d\mathbf{z}_\sigma = -\sigma \cdot \nabla_{\mathbf{z}} \log p_\sigma(\mathbf{z}_\sigma) d\sigma, \tag{2}$$

where $\nabla_{\mathbf{z}} \log p_t(\mathbf{z}_t)$ is the score function. Noise schedule $\sigma(t)$ is set as time step t. The training objective is to minimize the $L2$ loss with the denoiser network D_θ for different σ:

$$\mathbb{E}_{\mathbf{z}_0 \sim p_{data}} ||D_\theta(\mathbf{z}_\sigma) - \mathbf{z}_0||_2^2, \tag{3}$$

with the relation between D_θ and the score function $\nabla_{\mathbf{z}} \log p(\mathbf{z};\sigma)$ as follows:

$$\nabla_{\mathbf{z}} \log p(\mathbf{z};\sigma) = (D(\mathbf{z}_\sigma) - \mathbf{z})/\sigma^2. \tag{4}$$

SVD [31] utilizes the EDM framework to perform generative training on large-scale video datasets, resulting in a high-quality video generation model. An image-to-video generation model capable of forecasting future frames given the first frame has been released. Following SVD method, we also process videos in latent space. Given an input video $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$, a pretrained VAE encoder is used to project it into the latent space frame by frame, resulting in the latent representation $\mathbf{z}_0 \in \mathbb{R}^{T \times h \times w \times D}$. We then build GenRec based on SVD, inheriting its strong spatial-temporal priors as foundation for the subsequent generation and classification tasks.

3 GenRec

We now introduce GenRec, a simple yet efficient framework, that can not only generate temporally-coherent videos conditioned on an arbitrary number of provided frames but also is able to recognize actions and events with the help of encoded spatial-temporal priors. To this end, GenRec explores the strong spatial-temporal priors learned by a video diffusion model. In this work, we instantiate GenRec with the powerful open-source Stable Video Diffusion model (SVD) [1], which is pretrained on large-scale video datasets and is able to produce a photo-realistic video when provided a single frame. Then, for generation, GenRec follows the classical EDM framework to learn noise reduction trajectories. For recognition, on the other hand, GenRec operates on intermediate decoded features using a recognition head. Furthermore, to generate videos in a more free fashion, *i.e.* an arbitrary collection of frames used as condition, we design a latent masking strategy that "interpolates" masked frames. Such a strategy also benefits recognition by easing the training process. More importantly, by doing so GenRec supports a multitude of downstream tasks, particularly when limited visual information is provided.

3.1 Pipeline Overview

Latent diffusion and latent masking. During the diffusion process, the Gaussian noise with a certain noise level is added to the latent representation $\mathbf{z}_0$, creating a noisy latent representation $\tilde{\mathbf{z}}_i$ following Equation (1). Recall that while SVD contains powerful spatial-temporal priors, it can only perform generation when the first frame is provided. To allow a more "free" generation with an arbitrary number of frames as inputs, we design a latent masking strategy. More specifically, we apply a random mask $\mathbf{m}$ to the latent representation $\mathbf{z}_0$, producing a masked latent representation $\overline{\mathbf{z}_0}$. Such a strategy encourages the model to reconstruct the original video content from incomplete frames, which is in a similar spirit to MAE [19]. Note that when only the first latent is available, it degrades to the same as SVD; if all latents are masked out, this degrades to unconditional generation. Furthermore, doing so also benefits recognition tasks when limited visual clues are available. For example, in scenarios with limited bandwidth leading to reduced frame rates, the ability of video

https://doi.org/10.52202/079017-3456

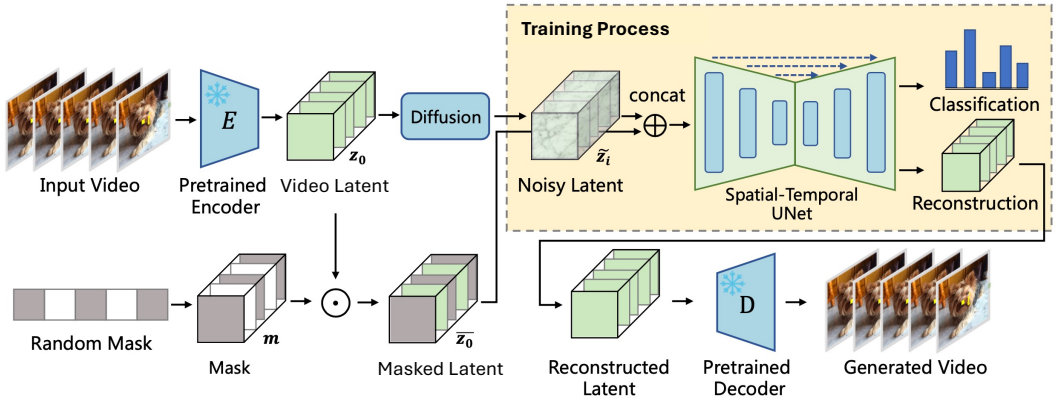

Figure 2: The pipeline of our proposed video processing method. The input video is first processed by a pretrained encoder E to produce a latent representation $\mathbf{z}_0$, then undergoes diffusion to generate a noisy latent $\tilde{\mathbf{z}}_t$. The random mask $\mathbf{m}$ is used to create the masked latent $\overline{\mathbf{z}_0}$. During training, the noisy latent is concatenated with the masked latent as condition and fed into a Spatial-Temporal UNet, resulting in both reconstruction and recognition outputs. The reconstructed latent can be decoded by the pretrained decoder D to produce the final generated video.

frame complementation enables the model to better predict and perceive complete video information. In practice, we simulate such conditions by randomly erasing half to all video conditions, retaining on average only about one-fourth of the original video information. This technique allows the model to effectively fill in the missing information, enhancing its ability to recognize and understand the video content despite the reduced data availability.

Unifying generation and understanding. To unify generation with masked latents, GenRec predicts pixel-level and semantic-level contents with the combination of the noisy latent $\tilde{\mathbf{z}}_i$ and the masked latent $\overline{\mathbf{z}_0}$ gained by the aforementioned latent diffusion and latent masking. The two latents are channel-wise concatenated $[\tilde{\mathbf{z}}_i, \overline{\mathbf{z}_0}]$ and are fed into a Spatial-Temporal UNet, together with features from observed frames, to learn spatial and temporal representations, following [1]. The weights of the UNet are initialized from [1] to obtain spatial and temporal priors, learned on large-scale video datasets.

For the generation task, the UNet aims to reconstruct the original latent representation from the combined noisy and masked inputs. Representing UNet as the mapping function F_θ, its goal is to predict clean latent, which, according to the EDM framework, takes the form of a representation mapping as follows:

$$D_\theta(\tilde{\mathbf{z}}_i; \overline{\mathbf{z}_0}, \sigma) = c_{skip}(\sigma)\tilde{\mathbf{z}}_i + c_{out}(\sigma)F_\theta([c_{in}(\sigma)\tilde{\mathbf{z}}_i, \overline{\mathbf{z}_0}]), \tag{5}$$

in which we set the same skip connection c_{skip}, scaling factor c_{out} and c_{in} as [1].

For the recognition task, we break down the UNet mapping function as $F = F_{tail} \cdot F_{head}$. And we consider $F_{head}([c_{in}(\sigma)\tilde{\mathbf{z}}, \overline{\mathbf{z}}]) \in \mathbb{R}^{T \times h' \times w' \times D'}$ as the compact video representation extracted from the intermediate layer of the UNet model, which is then fed into the classifier head ϕ_θ, consisting of an attentive pooler and a fully connected layer to predict video categories:

$$\hat{\mathbf{y}} = \phi(F_{head}([c_{in}(\sigma)\tilde{\mathbf{z}}_i, \overline{\mathbf{z}_0}])). \tag{6}$$

3.2 Optimization

We train GenRec with both generation and classification objectives, encouraging the model to learn high-quality video generation and accurate video understanding.

The generative loss uses a $L2$ loss to measure the difference between the original latent representation and the reconstructed output produced by the UNet, and is defined as:

$$L_G(\mathbf{z}_0, \tilde{\mathbf{z}}_i, \overline{\mathbf{z}_0}; \sigma) = \lambda(\sigma)\|D_\theta(\tilde{\mathbf{z}}_i; \overline{\mathbf{z}_0}, \sigma) - \mathbf{z}_0\|^2, \tag{7}$$

where $D_\theta(\tilde{\mathbf{z}}_i; \overline{\mathbf{z}_0}, \sigma)$ is the denoised output mentioned in Equation (5), and $\lambda(\sigma)$ is a weighting function based on the noise level σ referring to [1, 22]. While the classification loss uses a cross-entropy loss to measure the discrepancy between the true labels and the predicted labels, and is defined as:

$$L_D(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i \mathbf{y}_i \log(\hat{\mathbf{y}}_i), \tag{8}$$

where $\mathbf{y}$ denotes the ground truth labels, and $\hat{\mathbf{y}}$ represents the predicted labels referring to Equation Equation (6).

To balance the learning of generative and recognition tasks, we set a balancing weight γ to control the relative importance of each loss in the overall objective function. The total loss L is given by:

$$L = L_D + \gamma L_G, \tag{9}$$

3.3 Inference for Different Downstream Tasks

With the above training strategies, we now introduce how GenRec can flexibly support different types of generation and recognition tasks.

Video generation conditioned on frames. Once trained, GenRec is able to generate high-quality videos conditioned on an arbitrary number of given frames, thanks to the latent masking strategy. Particularly, following the EDM stochastic sampler framework and Equation (2), GenRec iteratively denoises the video conditioned on the masked latent $\overline{\mathbf{z}_0}$, as shown below:

$$\mathbf{z}_{i-1} = \hat{\mathbf{z}}_i + \epsilon_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}) = \hat{\mathbf{z}}_i + (t_{i-1} - \hat{t}_i)\frac{d\hat{\mathbf{z}}_i}{d\hat{t}_i} \tag{10}$$

$$= \hat{\mathbf{z}}_i + (t_{i-1} - \hat{t}_i)(-1)\frac{D_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}, \sigma) - \hat{\mathbf{z}}_i}{\hat{t}_i}, \tag{11}$$

where $\hat{\mathbf{z}}_i$ is derived from $\tilde{\mathbf{z}}_i$ adding a perturbation. With an iteratively denoising process, we can finally obtain the denoised video latent $\mathbf{z}_0$ which can be decoded as a complete video.

Video generation conditioned on classes. When the number of visible frames is extremely limited, the motion trajectory becomes unpredictable and thus it would be hard to make a reliable prediction of the future. To mitigate this issue, GenRec supports adding category information to guide video generation in the expected desired direction.

Formally, we simplify Equation (11) with Equation (4), and obtain:

$$\epsilon_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}) = (t_{i-1} - \hat{t}_i)(-1)\frac{D_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}, \sigma) - \hat{\mathbf{z}}_i}{\hat{t}_i} \tag{12}$$

$$= (-1)(t_{i-1} - \hat{t}_i)\hat{t}_i \nabla_{\hat{\mathbf{z}}_i} \log p_\theta(\hat{\mathbf{z}}_i). \tag{13}$$

We substitute the score function $\nabla_{\hat{\mathbf{z}}_i} \log p_\theta(\hat{\mathbf{z}}_i)$ with the conditional form $\nabla_{\hat{\mathbf{z}}_i} \log p_\theta(\hat{\mathbf{z}}_i|y)$, in which y denotes the conditional class. By applying Bayes' Theorem, the original score function can be replaced by $p(\hat{\mathbf{z}}_i)p(y|\hat{\mathbf{z}}_i)$, and we can get the conditional version of residual, denoted as $\epsilon^*_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0})$:

$$\epsilon^*_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}) = (-1)(t_{i-1} - \hat{t}_i)\hat{t}_i \nabla_{\hat{\mathbf{z}}_i} \log p(\hat{\mathbf{z}}_i)p(y|\hat{\mathbf{z}}_i) \tag{14}$$

$$= (-1)(t_{i-1} - \hat{t}_i)\hat{t}_i [\nabla_{\hat{\mathbf{z}}_i} \log p(\hat{\mathbf{z}}_i) + \nabla_{\hat{\mathbf{z}}_i} \log p(y|\hat{\mathbf{z}}_i)] \tag{15}$$

$$= \epsilon_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}) - (t_{i-1} - \hat{t}_i)\hat{t}_i \nabla_{\hat{\mathbf{z}}_i} \log p(y|\hat{\mathbf{z}}_i) \tag{16}$$

Considering the scaling factor of $\hat{\mathbf{z}}_i$: $c_{in}(\sigma) = \frac{1}{\sqrt{\sigma^2 + \sigma_{data}}}$ (following [22], and $\sigma_i = t_i$), that would pre-scale the input as $c(\mathbf{z}_i) = c_{in}(t_i) \cdot \mathbf{z}_i$ before model processing, the formulation can be further transferred as:

$$\epsilon^*_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}) = \epsilon_\theta(\hat{\mathbf{z}}_i; \overline{\mathbf{z}_0}) - \frac{(t_{i-1} - \hat{t}_i)\hat{t}_i}{\sqrt{\hat{t}_i^{\,2} + \sigma_{data}}} \nabla_{c(\hat{\mathbf{z}}_i)} \log p(y|c(\hat{\mathbf{z}}_i)) \tag{17}$$

Following [11], we sharpen the distribution of $p(y|\mathbf{z})$ by multiplying a scaling factor $s > 1$, shown as $s \cdot \nabla_{\mathbf{z}} \log p(y|\mathbf{z}) = \nabla_{\mathbf{z}} \log \frac{1}{Z} p(y|\mathbf{z})^s$ where Z is an arbitrary constant. Larger scaling value would

https://doi.org/10.52202/079017-3456

bring more attention to the target category. Here, $p(y|c(\hat{\mathbf{z}}_i))$ comes from the classification branch in GenRec. Finally, we can use the same EDM sampling procedure with the derived class information to generate samples.

Standard video recognition. Based on Equation (6), GenRec can do the classical video recognition by setting constant no-mask, and thus $\overline{\mathbf{z}_0}$ is replaced by $\mathbf{z}_0$ and the prediction follows:

$$\hat{\mathbf{y}} = \phi(F_{head}([c_{in}(\sigma)\tilde{\mathbf{z}}_i, \mathbf{z}_0])). \tag{18}$$

Video recognition with partially observed frames. Based on Equation (6), GenRec can be applied to video recognition with partially observed frames, *e.g.*, early action prediction that aims to predict future events based on the initial frames, sparse video recognition where videos are sparsely encoded and transmitted due to bandwidth limitations. By masking the invisible frames to get $\tilde{\mathbf{z}}_i$, and replacing the noisy latent with random noise $\sim$ obeying Gaussian distribution, GenRec can do the prediction for partially visible videos, following:

$$\hat{\mathbf{y}} = \phi(F_{head}([\sim, \overline{\mathbf{z}_0}])). \tag{19}$$

4 Experiments

4.1 Experimental Setup

Datasets. In our experiments, we use the following four datasets: Something-Something V2 (SSV2) [17], Kinetics-400 (K400) [24], UCF-101 [35] and Epic-Kitchen-100 (EK-100) [10]. SSV2 dataset is designed for fine-grained action recognition and it contains 174 action classes, 220,847 short video clips with an average duration of 4 seconds. K400 contains 400 action classes, 306,245 video clips with an average duration of 10 seconds. The UCF-101 dataset comprises 13,320 videos from 101 action categories and is widely utilized for human action recognition. The EK-100 dataset focuses on egocentric vision. It contains a total of 90,000 annotated action segments, encompassing 97 verb classes and 300 noun classes.

Evaluation protocols. GenRec performs both generation and recognition tasks. For generation, we use the Fréchet Video Distance (FVD) [41] metric to assess the quality of the generated videos. A lower FVD score indicates higher fidelity and realism. For recognition, we measure the top-1 accuracy that reflects the portion of correctly classified videos. We validate our model performance in formal video recognition, partial video recognition, class-conditioned image-to-video generation and frame completion with the above metrics.

Implementation details. We initially set the learning rate to 1.0×10^{-5} and set the total batch size as 32. Only generation loss will be retained for model adaptation on specific datasets. We train 200k steps on EK-100 and UCF, and 300k steps on SSV2 and K400, respectively. Subsequently, we finetune GenRec with both generation and recognition losses. The learning rate is set to 1.25×10^{-5} and decayed to 2.5×10^{-7} using a cosine decay scheduler. We warm up models with 5 epochs, during which the learning rate is initially set as 2.5×10^{-7} and linearly increases to the initial learning rate 1.25×10^{-5}. The loss balance ratio γ is set to 10, and the learning rate for the classifier head is ten times higher than the base learning rate. We drop out the conditions 10% of the time for supporting classifier-free guidance [20], and we finetune on K400 for 40 epochs and 30 epochs on other datasets. The training is executed on 8 A100s and each contains a batch of 8 samples. We sample 16 frames for each video.

4.2 Main Results

Comparison to state-of-the-art in video recognition and video generation. We compare with state-of-the-art methods in terms of their recognition accuracy and generation quality. The results are summarized in Table 1. The first two blocks of the table presents current advanced video recognition models, while the third block demonstrates the performance of the diffusion-based class-guided image-to-video generation.

As shown in the table, GenRec achieves optimal results or performs on par with the state-of-the-art approaches. In terms of video recognition, GenRec achieves 75.8% accuracy on SSV2 dataset,

Table 1: Performance of Video Recognition and Generation Methods. We evaluate on video recognition and class-conditioned image-to-video generation tasks. SEER† predicts 16 frames, while others predict 12 frames. Top-1 accuracy and FVD scores are reported. Baseline I adapts SVD to datasets with generative fine-tuning and then uses attentive-probing for classification. Baseline II fully finetunes SVD with classification supervision only in traditional classification framework.

			Classification Acc (↑)		**Generation FVD (↓)**	
Method	**Resolution**	**Param.**	**SSV2**	**K400**	**SSV2**	**EK-100**
w/o multi-modal align.						
VideoMAE-L [40]	224×224	305M	74.3	85.2	-	-
VideoMAE-H [40]	224×224	633M	-	86.6	-	-
OmniMAE-H [16]	224×224	650M	75.5	85.4	-	-
MVD-H [44]	224×224	633M	77.3	87.2	-	-
Hiera-L [32]	224×224	214M	75.1	87.3	-	-
Hiera-H [32]	224×224	673M	-	87.8	-	-
MaskFeat-L [47]	312×312	218M	75.0	86.4	-	-
w/ multi-modal align.						
InternVideo [45]	224×224	1.3B	77.2	91.1	-	-
InternVideo2 [46]	224×224	6B	77.4	92.1	-	-
OmniVec [36]	-	-	85.4	91.1	-	-
OmniVec-2 [37]	-	-	86.1	93.6	-	-
TATS [14]	128×128	-	-	-	428.1	920.0
MCVD [42]	256×256	> 3.5B	-	-	1407	4804
SimVP [13]	64×64	-	-	-	537.2	1991
VideoFusion [28]	256×256	1.8B	-	-	163.2	349.9
Tune-A-Video [50]	256×256	> 860M			291.4	365.0
SEER [18]	256×256	> 860M	-	-	112.9	271.4
SEER† [18]	256×256	> 860M	-	-	355.4	-
Baseline I	256×256	2.1B	63.7	82.0	50.3	53.6
Basline II	256×256	1.9B	75.9	86.6	-	-
GenRec	256×256	2.1B	75.8	87.2	46.5	49.3

surpassing the majority of current state-of-the-art methods. On K400, GenRec achieves 87.2% accuracy, which is on par with the performance of MVD-H (87.2%) and Hiera (87.3%, 87.8%), and surpasses other advanced methods. These results indicate the effectiveness of our approach in video recognition. In addition, GenRec shows a slight performance gap compared to the methods in the second block. It is important to note that these advanced methods benefit significantly from pretraining on large-scale multimodal alignment datasets, which provide extensive cross-modal supervision that enhances their ability to capture semantic relationships across video frames.

We further construct two strong baselines. Baseline I adapts SVD to the respective dataset through generative fine-tuning, followed by attentive-probing for classification, where the backbone is frozen and all frames are used as input. Baseline II involves fully fine-tuning the original SVD model with classification supervision only, ensuring that all frames are visible during training. Compared with them, GenRec performs on par or better. GenRec performs good in supporting not only classification but also generation, demonstrating its comprehensive capability in handling both tasks effectively.

In terms of video generation, we evaluate the model on class-conditioned image-to-video generation task following [18]. Comparing the FVD scores of SEER:112.9 and SEER†:355.4, it can be inferred that generating longer videos with 16 frames is more difficult than generating 12 frames. GenRec generates videos with 16 frames and achieves much lower FVD scores than the other methods, demonstrating the effectiveness of our approach in video generation.

It is worth highlighting that, current research always treats video recognition and generation tasks in a separate manner, and most of the advanced methods focus primarily on either recognition or generation tasks. For instance, SEER method excels in class-conditioned image-to-video generation, but lacks the ability to do video recognition. While current research on representation learning, shown as the first and second blocks in Table 1, lacks the ability to do video generation tasks. In contrast, GenRec not only unifies these tasks, but also achieves competitive results compared to the specialized methods.

https://doi.org/10.52202/079017-3456

Table 2: Early action prediction and limited interpolation problem on Something-Something V2 dataset, with one temporal crop. ρ denotes the visible ratio according to the whole video. Acc denotes to the top-1 accuracy. Ratio metric represent the percentage of maximum performance that the model can maintain at various frame rates. The 'w/o G' experiment refers to the results obtained by removing the generative supervision from our method.

		Early Action Prediction (ρ)					**Limited Inter. Frames**			
Method	**Metric**	0.1	0.3	0.5	0.7	1.0	2 fs	3 fs	4 fs	16 fs
TemPr [38]	Accuracy	20.5	28.6	41.2	47.1	66.3	-	-	-	-
	Retention	30.9%	43.1%	62.1%	71.3%	100%	-	-	-	-
MVD [44]	Accuracy	-	-	-	-	-	34.2	54.3	64.4	75.0
	Retention	-	-	-	-	-	45.6%	72.4%	85.9%	100%
MVD† [44]	Accuracy	26.9	39.8	55.6	70.2	75.0	53.6	65.0	68.8	75.0
	Retention	35.9%	53.1%	74.1%	93.2%	100%	71.5%	86.7%	91.7%	100%
w/o G	Accuracy	27.3 $\downarrow$1.6	39.6 $\downarrow$2.3	56.3 $\downarrow$1.4	71.6 $\downarrow$0.8	75.0 $\downarrow$0.3	53.5 $\downarrow$2.2	65.8 $\downarrow$1.5	69.8 $\downarrow$1.0	75.0 $\downarrow$0.3
GenRec	Accuracy	**28.9**	**41.9**	**57.7**	**72.4**	75.3	**55.7**	**67.3**	**70.8**	75.3
	Retention	**38.4%**	**55.6%**	**76.6%**	**96.1%**	100%	**74.0%**	**89.4%**	**94.0%**	100%

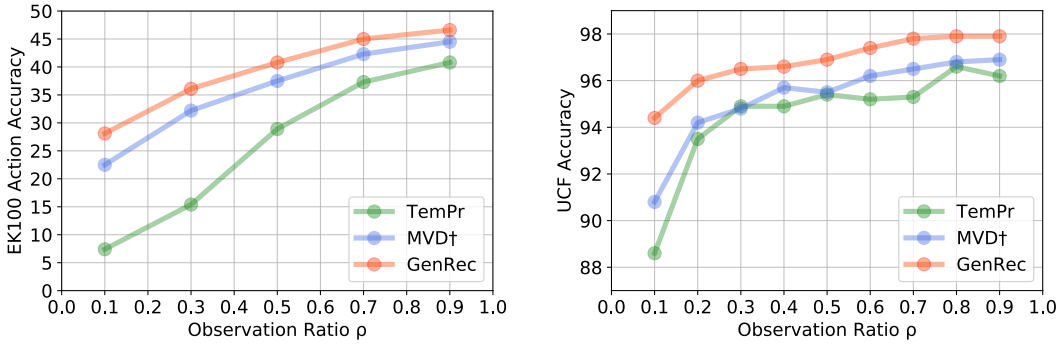

Figure 3: Early action prediction on EK-100 and UCF-100 datasets, with one temporal crop.

Comparison to state-of-the-art in video recognition with limited frames. GenRec supports video recognition when only partial frames can be observed. We evaluate this capability on the SSV2 and EK-100 datasets. Our evaluation includes two tasks: an early prediction task, where the model has access only to previous continuous frames following the setting of [38], and a recognition task where videos are sparsely sampled, and the model is expected to make correct predictions. For fair comparisons, we construct two strong baselines. We first apply MVD [44] to directly deal with the recognition task by constructing a dense video through nearest neighbor interpolation. We also construct another baseline similar to our training pipeline, where we apply frame dropout in the training process of MVD [44] for better fitting on task with partial frames, and is named as MVD†. In all settings, the number of fully observed frames is 16.

Table 2 shows the results under these settings, in which ρ denotes the visible ratio (there are a total of 16 frames). In the early action prediction task, GenRec achieves the highest accuracy and ratio metrics at all observation levels. Notably, GenRec and MVD† exhibit similar performance when all frames are observed, but as the number of observed frames decreases, GenRec demonstrates higher accuracy. GenRec also shows superior performance when videos are sparsely sampled, maintaining high accuracy even with fewer observed frames (e.g., 55.7% for 2 frames and 70.8% for 4 frames), indicating its robustness in handling sparse data. Moreover, we compute the ratio metric representing the percentage of maximum performance that the model can maintain at various frame rates, mitigating the unfairness caused by different backbone networks. In this scenario, GenRec still achieves the best performance.

We further investigate the contributions of generation supervision for recognition. As seen in Table 2, removing generation supervision results in noticeable performance degradation across various tasks, especially when the number of visible frames get less. For example, in the early prediction task, the accuracy decreases by 0.3% at $\rho = 1.0$ and by 2.3% at $\rho = 0.3$. These results suggest that generation supervision is essential for maintaining high performance, particularly when the model has to make

Table 3: Relation between generation and recognition.

Method	Metric	Early Frames				Limited Inter. Frames		
		2 fs	5 fs	8 fs	12fs	2 fs	3 fs	4 fs
GenRec	Acc ↑	28.9	41.9	57.7	72.4	55.7	67.3	70.8
	FVD↓	57.8	44.0	30.3	16.6	46.7	31.7	24.3

Table 4: Choice of UNet layers for feature extraction.

Up Index	1	2	3
SSV2	71.8	75.8	75.2

Table 5: Ablation study on different masking strategies.

Method	Expectation of masking ratio	SSV2	
		Acc (↑)	FVD (↓)
GenRec	75%(our choice)	75.8	46.5
	50%	+0.3	+0.5
	87.5%	-0.9	-0.3

predictions with limited visual information. By incorporating generation supervision, the model can better handle scenarios with incomplete data, improving robustness and accuracy.

We also evaluate the early action prediction on EK-100 and UCF-101. EK-100 is a temporally sensitive dataset similar to SSV2, demanding in terms of the model's temporal modeling capability, while UCF-101 demands more on appearance modeling. We conduct early prediction evaluation on them to further reveal the robustness of our GenRec. As shown in Figure 3, GenRec clearly outperforms TemPr and MVD†. In particular, the improvement becomes more significant as the number of observed frames decreases. More evaluation results can be seen in Appendix A.1.

These results collectively demonstrate that GenRec effectively handles missing video frames. The robustness and high accuracy of GenRec across different datasets and observation ratios highlight its potential for real-world applications where video data might be incomplete or sparsely sampled.

The relationships between generation and recognition . We further investigate the consistency between video generation and recognition, as shown in Table 3. We evaluate the performance of video recognition and generation with limited frames and find that the recognition accuracy not only depends on the number of visible frames but also significantly on the location of these frames. Interestingly, uniform sampling appears to facilitate video recognition better than dense sampling from the video prefix. Specifically, with the same number of frames, early prediction consistently shows lower accuracy compared to uniformly sampled frames (e.g., 28.9% vs. 55.7% with 2 frames) and worse FVD scores (e.g., 57.8 vs. 46.7 with 2 frames). When only three interpolated frames are visible, the 31.7 FVD score is comparable to that of an eight-frame prefix (30.3), while achieving much higher recognition accuracy. These results highlight the importance of complete state observation for action recognition and also suggest that video generation performance can potentially reflect task difficulty.

Choice of UNet layers. As described in Section 3, the UNet mapping function F is decoupled into $F_{tail} \cdot F_{head}$, where F_{head} serves as the feature extractor for video recognition. Our UNet model contains 4 main up-sampling blocks. We investigate which one is best suited for recognition. As shown in Table 4, using the second up-sampling block (Up Index 2) yields the best performance with an accuracy of 75.8%. The third block (Up Index 3) followed with 75.2%, while the first block (Up Index 1) has the lowest accuracy. As such, we choose the second block for feature extraction.

Explore the influence of the masking strategy. We also conduct an ablation study on the masking schemes using different expected masking ratios, as shown in Table 5. The results show that the FVD scores remain similar across different ratios, and a larger masking ratio might be beneficial for generation, as it closely resembles our class-conditioned frame prediction scenario with one or two given frames. However, an excessively large masking ratio (87.5%) negatively impacts action recognition accuracy, leading to a 0.9% decrease compared to our selected ratio.

Noise incorporation during inference for video recognition. In the inference stage for action recognition, GenRec applies a specific level of noise to video inputs before extracting visual features, as formulated in Equation 18. This added noise helps maintain consistency with the noisy training

https://doi.org/10.52202/079017-3456

Table 6: SSV2 action recognition accuracy with different random seeds.

Seed	0	1	2	3	4	5	AVG
SSV2 Acc (%)	75.83	75.82	75.83	75.83	75.86	75.84	75.835 ± 0.0125

process. To further understand the influence of noise randomness on classification accuracy, we conducted an experiment on the SSV2 dataset, using multiple random seeds to generate the noise, as presented in Table 6. The results demonstrate remarkable consistency in accuracy across different random seeds, with a standard deviation of only 0.0125%. This minimal variation highlights the model's robustness to noise fluctuations during inference, suggesting that the model's performance remains stable despite noise introduced by random sampling. If fully deterministic outcomes are desired, fixing the random seed for noise sampling will eliminate any remaining variability and guarantee consistent predictions across runs.

5 Related Work

Video diffusion models for generation. The great success of diffusion models in image generation has led to rapid advancements in video generation, including text-to-video generation [15, 2, 52, 54], image&text-to-video generation [56, 18, 21], and video editing [4, 5, 26, 29, 12, 53]. Many current works [52, 18] adapts the diffusion models from images to videos by incorporating temporal convolutions and attention mechanisms. One typical and excellent work, Stable Video Diffusion [1], follows the above description and has provided valuable foundations for generating high-quality, diverse, and temporally consistent videos. Different from the previous work, in our paper, we pursue not only the quality of generation, but also the unity of model generation capability and classification ability.

Diffusion models for visual understanding. Recently, researchers start to uncover the significance of diffusion models for discrimination tasks. A notable approach involves utilizing pretrained visual diffusion models for various downstream tasks, such as image segmentation [55] and visual content correspondence [39]. Additionally, some studies treat diffusion learning as a self-supervised method to acquire valuable feature representations [8]. However, most current works either use stable diffusion networks as pretrained backbones for downstream tasks or completely destroy their generative capabilities. Consequently, the potential benefits of integrating generation and classification abilities into a single model remain under-explored, which is the primary focus of our paper.

6 Conclusion

In this work, we presented GenRec, a unified video diffusion model that enables joint optimization for both video generation and recognition. GenRec exploits the significant temporal modeling power embedded in the diffusion model, allowing for mutual reinforcement between generation and recognition tasks. Extensive experiments were conducted to evaluate the performance of GenRec, demonstrate our approach contains strong generation and recognition capabilities at the same time in different kinds of scenarios, including normal or partial video recognition, video completion and class-conditioned image-to-video generation. Our findings highlight the potential of combining generation and classification tasks within a single unified model, providing valuable insights into the development of more sophisticated and versatile video analysis models. Future work will focus on further refining this integration and exploring its applications across various real-world scenarios.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (#62032006). The authors would like to thank Rui Wang, Rong Bao, Rui Tian for their help and suggestions.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[2] 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 *CVPR*, 2023.

[3] Yu Cao, Daniel Barrett, Andrei Barbu, Siddharth Narayanaswamy, Haonan Yu, Aaron Michaux, Yuewei Lin, Sven Dickinson, Jeffrey Mark Siskind, and Song Wang. Recognize human activities from partially observed videos. In *CVPR*, 2013.

[4] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023.

[5] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023.

[6] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. *arXiv preprint arXiv:2305.15241*, 2023.

[7] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, 2023.

[8] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024.

[9] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *NeurIPS*, 2024.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022.

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023.

[13] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *CVPR*, 2022.

[14] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022.

[15] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, 2023.

[16] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *CVPR*, 2023.

[17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

 https://doi.org/10.52202/079017-3456

[18] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

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

[21] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.

[22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[25] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, 2014.

[26] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023.

[27] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *The Twelfth International Conference on Learning Representations*, 2023.

[28] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023.

[29] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[32] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023.

[33] 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*, 2022.

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[35] 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.

[36] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. In *WACV*, 2024.

[37] Siddharth Srivastava and Gaurav Sharma. Omnivec2-a novel transformer based network for large scale multimodal and multitask learning. In *CVPR*, 2024.

[38] Alexandros Stergiou and Dima Damen. The wisdom of crowds: Temporal progressive attention for early action prediction. In *CVPR*, 2023.

[39] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023.

[40] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

[41] 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.

[42] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022.

[43] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023.

[44] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, 2023.

[45] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[46] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

[47] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.

[48] Zejia Weng, Zuxuan Wu, Hengduo Li, Jingjing Chen, and Yu-Gang Jiang. Hcms: Hierarchical and conditional modality selection for efficient video recognition. *ACM TOMM*, 2023.

[49] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *ICML*, 2023.

[50] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023.

[51] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data. *TPAMI*, 2024.

[52] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023.

[53] Zhen Xing, Qi Dai, Zihao Zhang, Hui Zhang, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023.

[54] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2023.

[55] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023.

 https://doi.org/10.52202/079017-3456

[56] Xi Ye and Guillaume-Alexandre Bilodeau. Stdiff: Spatio-temporal diffusion for continuous stochastic video prediction. In *AAAI*, 2024.

[57] Hui Zhang, Zheng Wang, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Denoising diffusion for anomaly detection. *arXiv preprint arXiv:2303.08730*, 2023.

A Appendix / supplemental material

A.1 Early Prediction on EK-100 and UCF-101

More detailed evaluation results of the video recognition with limited frames on EK-100 and UCF-101 can be seen here.

Table 7: Early action prediction on EK-100.

Method	Verb Obs. Ratio(ρ)					Noun Obs. Ratio(ρ)					Action Obs. Ratio(ρ)				
	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
TemPr [38]	21.4	34.6	54.2	63.8	67.0	22.8	32.3	43.4	49.2	53.5	7.4	15.4	28.9	37.3	40.8
MVD† [44]	49.3	60.0	64.7	68.8	71.3	35.7	44.7	49.1	53.4	55.2	22.5	32.2	37.5	42.3	44.5
GenRec	**55.8**	**63.6**	**67.9**	**71.7**	**73.1**	**40.1**	**47.8**	**52.0**	**55.3**	**56.7**	**28.1**	**36.1**	**40.8**	**45.0**	**46.6**

Table 8: Early action prediction on UCF dataset.

Method	Observation Ratio(ρ)								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TemPr [38]	88.6	93.5	94.9	94.9	95.4	95.2	95.3	96.6	96.2
MVD† [44]	90.8	94.2	94.8	95.7	95.5	96.2	96.5	96.8	96.9
GenRec	**94.4**	**96.0**	**96.5**	**96.6**	**96.9**	**97.4**	**97.8**	**97.9**	**97.9**

A.2 Training and Inference Time Cost Compared with Previous Classification Methods

We compare the training and inference time cost between MVD-H, Baseline II and GenRec, fairly on the same hardware resource: 4-nodes * 8 V100s, and same batch size. Baseline II refers to fully fine-tuning the original SVD model with classification supervision only.

As shown in the Table 9, GenRec and the Baseline II consumes more testing time than MVD-H. The difference primarily arises from the varying number of parameters. Our model is derived from a generative model, and the complexity of the generative task necessitates a larger number of parameters for effective learning. Compared with Baseline II, since GenRec requires additional decoder blocks for video generation training, the training time will get increased a little bit. As the additional up-sampling blocks will not be used when doing action recognition, GenRec shares the same testing time with Baseline II. It is worth noting that "†" in MVD-H is to highlight that MVD method would use repeated augmentation technique during training, but such augmentation will significantly increase the training time. Our approach does not need to use that augmentation. All the methods do the down-stream finetuning for 30 epochs.

Table 9: Comparison with recognition methods in terms of parameters, training time, and test time.

Method	Trainable Params	Total Params	Training Time	Training Epochs	Test Time (2×3 clips)
MVD-H†	633M	633M	3038 s/Epoch	30	441 s
Baseline II	1.3B	1.9B	2954 s/Epoch	30	1500 s
GenRec	1.5B	2.1B	3751 s/Epoch	30	1500 s

A.3 Ablation on Loss Balance Ratio

We conduct an ablation study on the loss balance ratio λ as shown in Equation 9. Results presented in Table 10 show that varying the loss ratio has a minor impact on action recognition accuracy. However, setting the ratio too low negatively affects the generation performance. In particular, when the ratio is set to zero, significant forgetting occurs, severely compromising the model's generation ability.

https://doi.org/10.52202/079017-3456

Table 10: Ablation study on the effect of different loss balance ratios on SSv2 accuracy and FVD. Higher SSV2 accuracy and lower FVD indicate better performance.

Balance Ratio λ	0	1	5	10	20
SSv2 Acc ↑	75.6	75.5	75.6	75.8	75.5
SSv2 FVD ↓	1579.2	52.0	47.4	46.5	46.9

A.4 Comparision with SVD Baseline

Comparing with the open-source Stable Video Diffusion directly is meaningful. We conduct frame prediction tests using SVD model on the SSV2 and Epic-Kitchen datasets, as shown in Table 11. Since the SVD model performs better at higher resolutions, we generated videos at 512x512 resolution and then downsampled them to 256x256 for FVD calculations. The results show that while SVD achieves competitive scores compared to previous state-of-the-art methods shown in Table 1, it is suboptimal compared to GenRec. This is likely due to SVD's design for general scenarios. Baseline I represents our enhanced SVD baseline, which fine-tunes SVD on the target datasets for better results and fairer comparison with GenRec.

Table 11: Comparison of SSv2 FVD and Epic FVD scores in frame prediction tests.

Method	SSv2 FVD	Epic FVD
SVD	**99.7**	**180.8**
Baseline I	50.3	53.6
GenRec	46.5	49.3

A.5 Impact of Classification Training on Generative Ability

To demonstrate classification and generative can coexist well in our training without negatively impacting each other, we construct the comparison with "Baseline I (SVD baseline)", which adapts SVD on the target dataset without classification supervision, and then trains another classifier head freeze the generation backbone, ensuring no influence from classification supervision. We conduct comparisons on SSv2: FP (frame prediction) and CFP (class-condition frame prediction), as shown in Table 12. By comparing the FP results, the similar scores between the two methods conclude that classification loss does not degrade the model's generative ability in our approach. Moreover, the CFP results indicate that our method, with its more accurate classification performance, can guide the model to achieve higher-quality video frame predictions.

Table 12: Comparison of SSv2 Acc and SSv2 FVD scores for Baseline I and GenRec methods in FP (frame prediction) and CFP (class-condition frame prediction) scenarios.

Method	SSv2 Acc ↑	SSv2 FVD ↓
Baseline I (FP)	-	55.5
GenRec (FP)	-	55.3
Baseline I (CFP)	63.7	50.3
GenRec (CFP)	75.8	46.5

A.6 Case Study for Class-Conditioned Image-to-Video Generation and Video Interpolation

We show the generated visualization of the GenRec. The model can support video generation given various numbers of frames, as well as category-guided generation. We show two of the most difficult generative scenarios, which are: (1) given the first frame and different action categories to guide the video generation, and (2) given the start and end frames, the model is expected to complement the video. We also compare our methods with SEER [18] with cases picked from its official website. The generation results can be seen in Figure 4, Figure 5 and Figure 6.

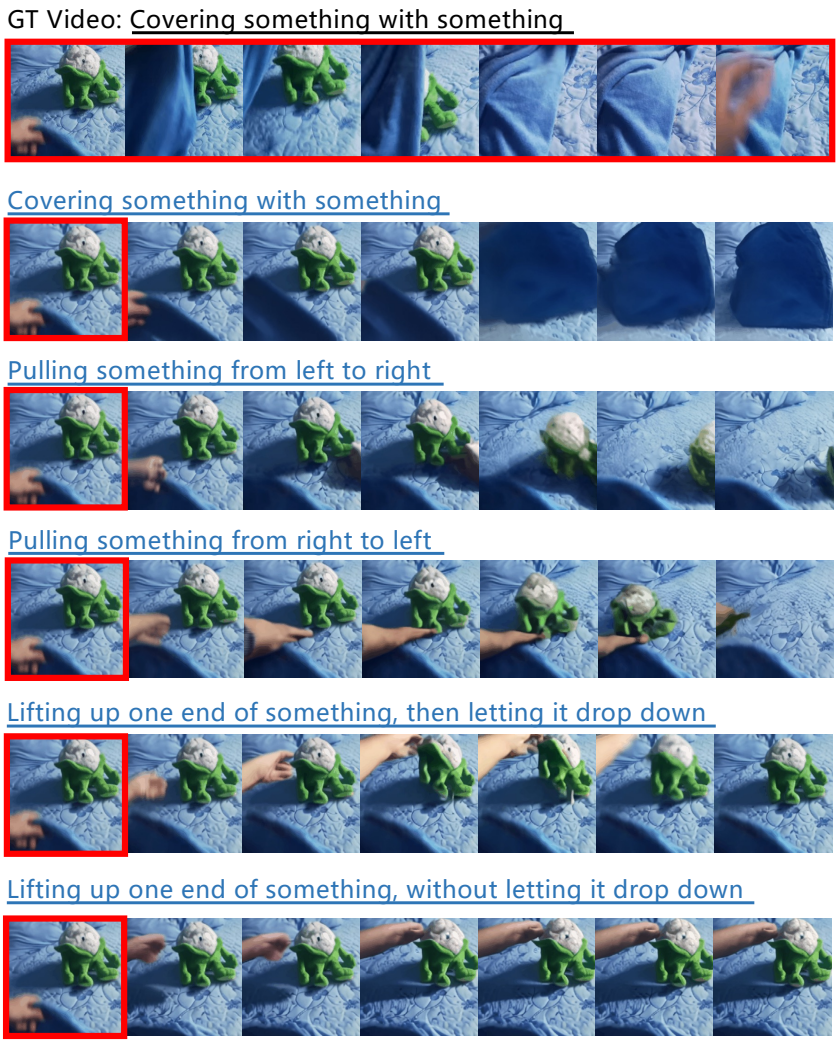

Figure 4: Video generation case study. We generate videos given the first frame together with the classifier guidance for various categories.

A.7 Limitations and Broader Impacts

Limitations and Future Work The objective of our paper is to unify the tasks of generation and recognition, achieving or even surpassing the state-of-the-art experimental performance across various tasks. However, our method is based on fine-tuning a pretrained video diffusion model, using more pretraining data and having a larger number of parameters compared to previous methods. This is an issue we need to address in the future, and exploring the distillation of a well-pretrained video diffusion model into a smaller model is a worthwhile future endeavor.

Broader Impacts The broader impact of the GenRec framework extends into various fields, enhancing capabilities in content creation, security, and accessibility. In the media industry, it allows for the automated generation of tailored, high-quality videos, reducing production costs and fostering creativity. For surveillance, its robustness in limited information scenarios improves monitoring effectiveness, particularly in challenging environments. Additionally, advancements of GenRec in video prediction can aid in developing assistive technologies, making digital content more accessible and interactive, particularly for individuals with visual impairments.

https://doi.org/10.52202/079017-3456

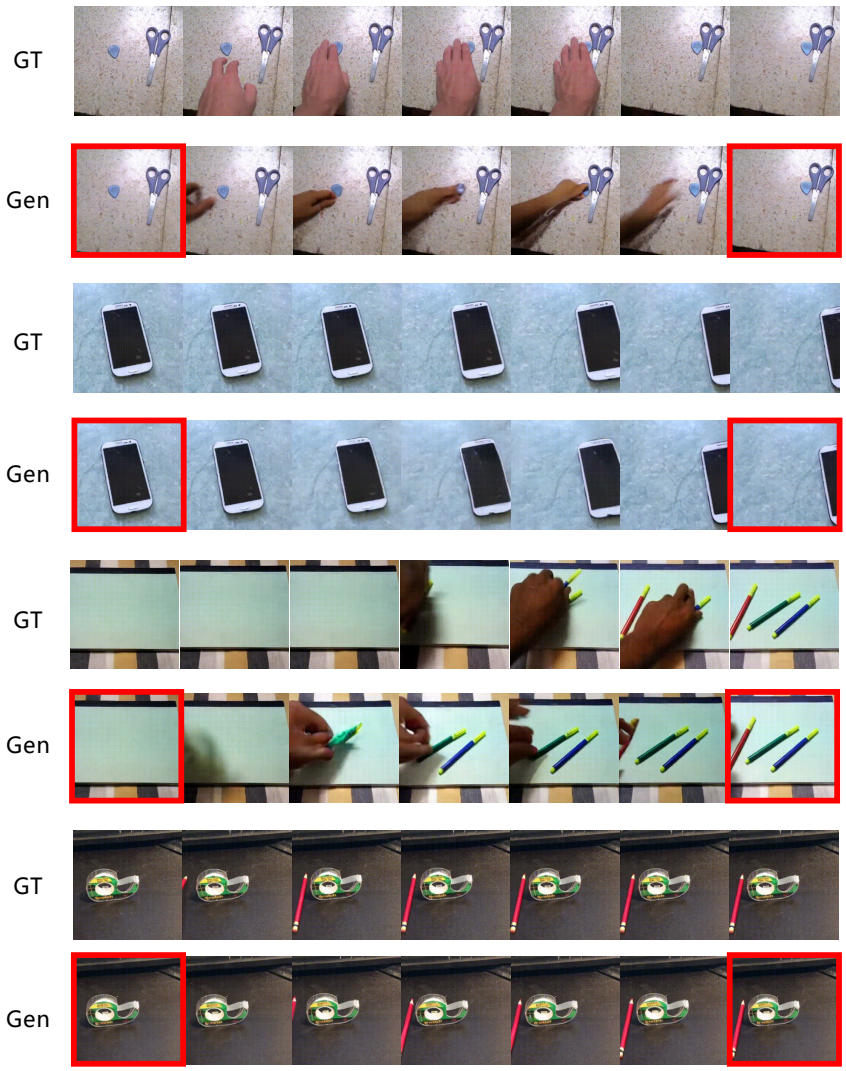

Figure 5: Video generation case study. We generate videos given the first frame and the last frame.

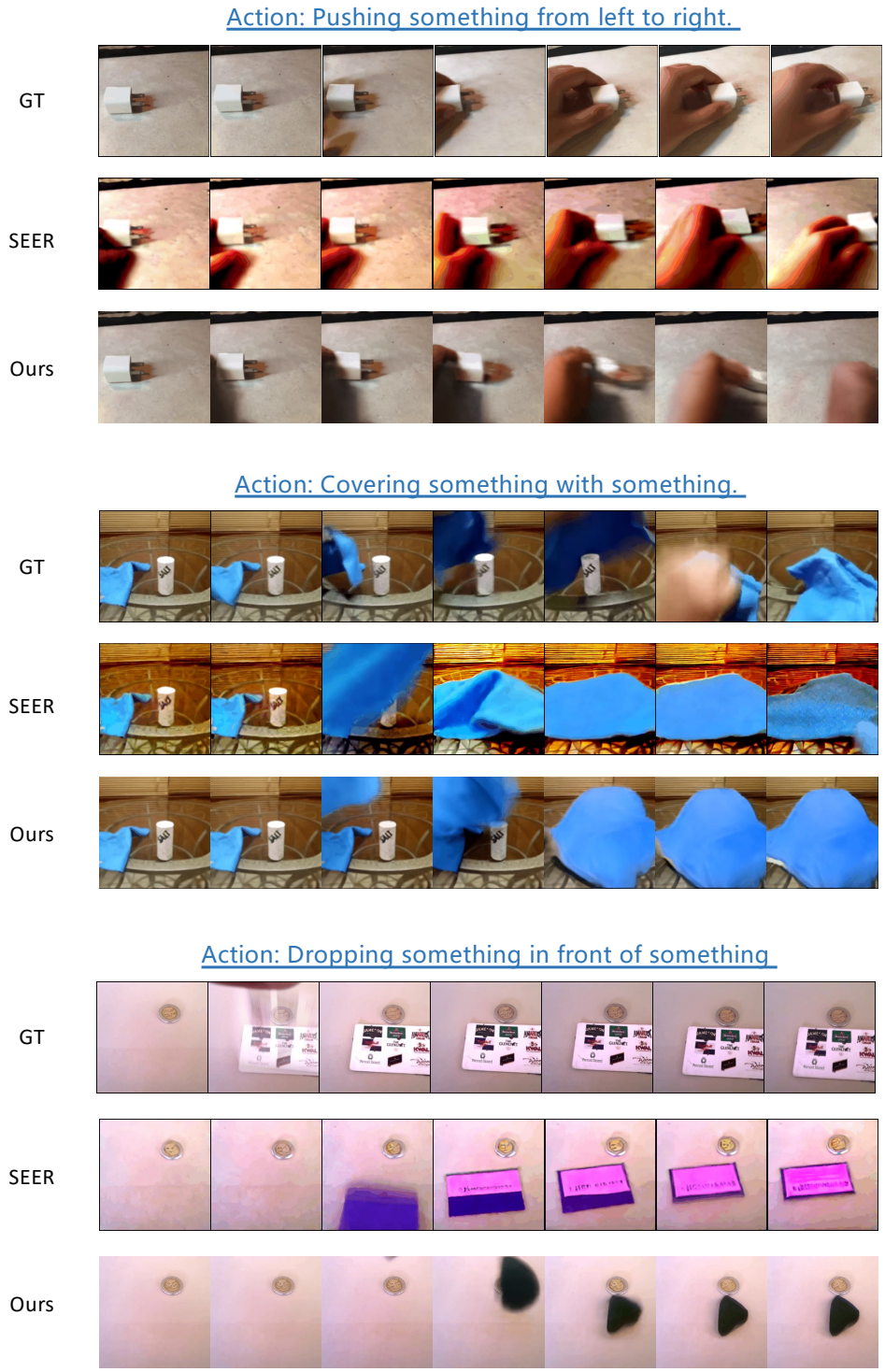

Figure 6: Video generation case study. We compare our methods with SEER [18] in the setting of generating videos given the first frame together with the classifier guidance. Cases are picked from the official website of SEER [18].

https://doi.org/10.52202/079017-3456

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We properly claim our contribution in abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations in Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper has provided the proof in main text.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided all the details in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-3456

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release the code in the feature.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details have been reported in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The cost of training required to report error bars is excessively high.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the computation resources in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

https://doi.org/10.52202/079017-3456

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper conform the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provided the broader impacts in appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper did not release data or models for the main contribution.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper follows the CC-BY 4.0 license in experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper did not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing experiments in this paper.

https://doi.org/10.52202/079017-3456

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing experiments in this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.