DISTILL VIDEO DATASETS INTO IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset distillation aims to synthesize compact yet informative datasets that allow models trained on them to achieve performance comparable to training on the full dataset. While this approach has shown promising results for image data, extending dataset distillation methods to video data has proven challenging and often leads to suboptimal performance. In this work, we first identify the core challenge in video set distillation as the substantial increase in learnable parameters introduced by the temporal dimension of video, which complicates optimization and hinders convergence. To address this issue, we observe that a single frame is often sufficient to capture the discriminative semantics of a video. Leveraging this insight, we propose **S**ingle-**F**rame **V**ideo set **D**istillation (SFVD), a framework that distills videos into highly informative frames for each class. Using differentiable interpolation, these frames are transformed into video sequences and matched with the original dataset, while updates are restricted to the frames themselves for improved optimization efficiency. To further incorporate temporal information, the distilled frames are combined with sampled real videos from real videos during the matching process through a channel reshaping layer. Extensive experiments on multiple benchmarks demonstrate that SFVD substantially outperforms prior methods, achieving improvements of up to 5.3% on MiniUCF, thereby offering a more effective strategy to video dataset distillation.

1 INTRODUCTION

Dataset distillation (DD) (Wang et al., 2018; Zhao et al., 2020; Cui et al., 2023; Wang et al., 2022; Yin et al., 2023; Su et al., 2024; Wang et al., 2025a) aims to distill a large dataset into a condensed informative synthetic dataset, so that the model trained on it could have a similar performance as the original dataset. This small, information-rich dataset drastically reduces training time and computational costs, significantly lowers storage requirements, enables rapid prototyping and hyperparameter tuning (Poyser & Breckon, 2024), and can facilitate easier data analysis and potentially even privacy-preserving data sharing (Yang et al., 2024b). While recent works (Cui et al., 2023; Guo et al., 2023; Su et al., 2024; Gu et al., 2024) have made significant progress on image dataset distillation, the application for video datasets remains relatively underexplored.

Videos extend images into the temporal domain, capturing not only spatial information in each frame but also motion and temporal dynamics over time. Furthermore, video recognition plays a crucial role in enabling machine learning models to understand the world, as it allows them to capture not only static appearances but also the evolving temporal patterns that characterize real-world phenomena. Recently, several studies have investigated video dataset distillation (Wang et al., 2024), but current approaches remain primarily empirical and still exhibit suboptimal performance.

In this paper, we argue that a fundamental challenge hindering effective video dataset distillation arises from the substantial increase in learnable parameters within synthetic datasets, which is caused by the added temporal dimensionality of video data compared to images. As illustrated in Figure 1(a), in image dataset distillation, each synthetic image is represented as a learnable tensor that is iteratively updated throughout the distillation process. In Figure 1(b), a straightforward extension of image distillation methods to video data results in treating each synthetic video as a fully learnable tensor. However, since a video inherently contains far more pixels than an image, this direct extension leads to an explosive growth of the parameter space. Such a heavy-parameter optimization makes it difficult for distillation algorithms to converge effectively (Nakkiran et al., 2021) and to comprehensively capture the complex information distribution of the original video dataset within a condensed dataset.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

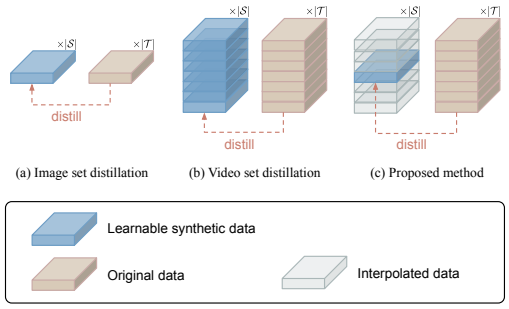
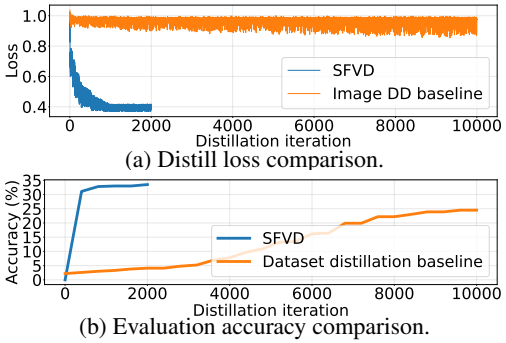


Figure 1: **Comparison of learnable parameter spaces.** $|\mathcal{S}|$ and $|\mathcal{T}|$ denote the synthetic and target dataset sizes. (a) In image dataset distillation, every synthesized image is fully learnable. (b) A naive extension to the video domain requires optimizing all frames within each synthetic video, leading to an intractable parameter space. (c) Our proposed method constrains learnability to a single frame per synthetic video, drastically reducing the parameter space to enable efficient and stable optimization.



(a) Distill loss comparison.

(b) Evaluation accuracy comparison.

Figure 2: **Convergence and performance comparison.** We highlight the limitations of naively applying state-of-the-art image distillation method (Guo et al., 2023) to video data. Tracking the optimization trajectory reveals that the baseline (orange) (a) struggles to converge and yields (b) suboptimal evaluation accuracy. In contrast, by effectively constraining the optimization space, our proposed SFVD (blue) achieves rapid convergence and significantly higher accuracy in substantially fewer iterations.

As illustrated in Figure 2, the direct application of image dataset distillation method (Guo et al., 2023) to video dataset encounters difficulties in convergence and yields suboptimal performance.

To address the above challenge, we build directly on our preliminary observation that a single frame is able to capture a substantial portion of the semantic content of a video. Motivated by this finding, we propose the Single-Frame Video set Distillation (SFVD) framework. We depart from directly matching synthetic videos with the original dataset and instead propose to match synthetic images with real videos. Specifically, we represent each video in the distilled dataset by a single frame, which is subsequently interpolated into a video sequence and match with the original dataset, while only updating the parameters within the frame. This design substantially reduces the optimization burden while enabling effective video dataset distillation. As illustrated in Figure 1(c), rather than optimizing the entire video, the objective of SFVD is to distill a small set of highly representative frames. This design substantially reduces the learnable parameter space during the distillation process, thereby enabling more effective and stable optimization.

To further incorporating temporal dynamics, we integrate the distilled frames with sampled original videos as temporal cues and employ a channel reshaping layer to fuse them into a combined representation. We then match the combined representation with the original dataset, where the distilled frames act as informative priors that guide parameter updates, ensuring robust initialization and faster convergence toward high-quality distilled video representations. We conduct extensive evaluations on four widely used video benchmarks: UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011) for human action recognition, Kinetics (Carreira & Zisserman, 2017) for large-scale video understanding, and Something-Something V2 (Goyal et al., 2017) for fine-grained temporal reasoning. Across all datasets and evaluation settings, our method consistently surpasses existing approaches by a substantial margin. On the Mini-UCF benchmark, our approach achieves substantial improvements, outperforming state-of-the-art methods by 4.5% in accuracy for IPC=1 and by up to 5.3% for IPC=5.

Our main contributions can be summarized as follows:

- We identify and analyze the significantly increased number of learnable parameters of the synthetic videos as a critical issue for video dataset distillation, impeding distillation convergence and information capture.
- We propose a novel Single-Frame Video set Distillation (SFVD) framework in which a single frame can be interpolated into a video, matched with the original dataset, while only updating the parameters within the frame. We further integrate temporal information alongside the distilled frames, creating a comprehensive representation.

- We demonstrate the effectiveness of our approach through extensive experiments on various benchmark datasets, showing that the proposed method significantly outperforms existing approaches.

2 Related work

Dataset distillation. Dataset distillation (Wang et al., 2018) compresses large datasets into smaller synthetic ones for efficient model training, benefiting applications like network architecture search (Elsken et al., 2019), federated learning (Li et al., 2020), and continual learning (Zenke et al., 2017). Foundational techniques primarily align models or data at different metric levels, including performance matching (Loo et al., 2022; Zhou et al., 2022; Nguyen et al., 2020; Loo et al., 2023; Nguyen et al., 2021; Chen et al., 2024), gradient matching (Zhao et al., 2020; Zhao & Bilen, 2021), trajectory matching (Cazenavette et al., 2022; Du et al., 2023; Guo et al., 2023; Li et al., 2023; Cui et al., 2023; Zhao et al., 2024; Lee & Chung, 2024; Yang et al., 2024a; Zhong et al., 2025; Wang et al., 2025b), and distribution matching (Zhao & Bilen, 2023; Wang et al., 2022; Shang et al., 2023; Zhao et al., 2023; Sajedi et al., 2023; Zhang et al., 2024a; Deng et al., 2024; Wang et al., 2025c; Cui et al., 2025c; Liu et al., 2023; Li et al., 2025; Cui et al., 2025b). Recently, to enable distillation at larger scales, decoupled optimization methods (Yin et al., 2023; Shao et al., 2024a; Sun et al., 2024; Shao et al., 2024b; Du et al., 2024; Zhang et al., 2024b; Shen et al., 2025; Tran et al., 2025; Cui et al., 2025a; Li et al., 2026) have improved computational efficiency, alongside the integration of generative models (Cazenavette et al., 2023; Gu et al., 2024; Su et al., 2024; Wang et al., 2025a) and soft labels (Qin et al., 2024; Shang et al., 2024) to enhance synthetic data quality. Beyond images, distillation has expanded to image-text (Wu et al., 2023; Xu et al., 2024; Guo et al., 2026), sequential (Zhang et al., 2025), medical (Li et al., 2024), and video (Wang et al., 2024) modalities. However, prior video dataset distillation work (Wang et al., 2024) remains primarily empirical, lacking a thorough analysis of why image-based methods often fail on video data. To bridge this gap, we identify the unique challenges of video model training dynamics and propose a novel method, SFVD, as detailed in Section 3.

Video recognition. Video classification has been extensively studied, and a wide range of architectures have been proposed to handle the spatial and temporal complexities of video data. Early approaches are based on 2D ConvNets, which is used for spatial feature extraction and a recurrent module, such as an LSTM or GRU (Donahue et al., 2015; Yue-Hei Ng et al., 2015), is stacked on top to model temporal dynamics across frames. To more effectively capture spatiotemporal features jointly, 3D ConvNets (Tran et al., 2015; Carreira & Zisserman, 2017; Feichtenhofer et al., 2019) were introduced. Hybrid architectures decompose 3D convolutions into spatial and temporal operations, as in P3D (Qiu et al., 2017), S3D (Xie et al., 2018), and R(2+1)D (Tran et al., 2018). More recently, transformer-based architectures have achieved strong performance on video tasks. Models like ViViT (Arnab et al., 2021) and Video Swin Transformer (Liu et al., 2022) adapt vision transformer structures to spatiotemporal input, while MViT (Multiscale Vision Transformer) (Fan et al., 2021) introduces a hierarchical structure for multi-scale temporal modeling. In this paper, following prior work on video dataset distillation (Wang et al., 2024), we primarily adopt C3D for evaluation, and additionally employ CNN+GRU/LSTM for cross-architecture validation.

3 Methods

3.1 Problem Formulation

The problem of Video Dataset Distillation is formulated as follows: given a large, real video dataset $\mathcal{T} = \{x_i, y_i\}_{i=1}^{|\mathcal{T}|}$, where $x_i \in \mathbb{R}^{f,c,h,w}$, video set distillation aims to generate a synthetic dataset $\mathcal{S} = \{\hat{x}_i, \hat{y}_i\}_{i=1}^{|\mathcal{S}|}$, where $|\mathcal{S}| \ll |\mathcal{T}|$, so that the model ϕ_θ trained on synthetic dataset $\mathcal{S}$ has the similar performance on the original dataset $\mathcal{T}$. It is worthy noting that $\hat{x}_i$ is not strictly constrained to $\hat{x}_i \in \mathbb{R}^{f,c,h,w}$, provided the preprocessing $g(\cdot)$ yields $g(\hat{x}_i) \in \mathbb{R}^{f,c,h,w}$. Typical dataset distillation methods (Liu & Du, 2025) use the following formula to optimize $\mathcal{S}$:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \ \mathbb{E}_{\theta^{(0)}\sim\Theta}[l\left(\mathcal{T}; \theta^*_{\mathcal{S}}\right)], \tag{1}$$

where $\theta^*_{\mathcal{S}} = \arg\min_{\theta} \ l(\mathcal{S}; \theta)$ is the optimal parameters trained on $\mathcal{S}$, $l(\cdot;\theta)$ represents the loss function. $\theta^{(0)}$ indicates that the model is trained from the first epoch, and Θ is the initial parameter

distribution. If we use a trajectory matching strategy (Cazenavette et al., 2022), the formulation of video set distillation could be written as:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbb{E}_{\theta^{(0)}\sim\Theta} \sum_{t=0}^{T} \mathcal{D}\left(\theta_{\mathcal{S}}^{(t)}, \theta_{\mathcal{T}}^{(t)}\right), \tag{2}$$

where $\mathcal{D}(\cdot,\cdot)$ calculates the distance between the parameter trained on synthetic dataset $\theta_{\mathcal{S}}$ and original dataset $\theta_{\mathcal{T}}$ at step t. T is the trajectory length.

3.2 Observation

Challenge of video dataset distillation. The principal challenge inherent in video dataset distillation comes from the significantly increased number of learnable parameters of the synthetic videos compared to that of synthetic images in image dataset distillation. Reviewing the dataset distillation on image dataset, backpropagation operates on a synthetic dataset $\mathcal{S} = \{\hat{x}_i, \hat{y}_i\}_{i=1}^{|\mathcal{S}|}$, where $\hat{x}_i \in \mathbb{R}^{c,h,w}$. However, for video set distillation, the back propagation takes effects on a synthetic video dataset $\mathcal{S} = \{\hat{x}_i, \hat{y}_i\}_{i=1}^{|\mathcal{S}|}$, where $\hat{x}_i \in \mathbb{R}^{f,c,h,w}$, with f denoting the number of frames. Because of this additional dimension f, the number of learnable parameters for each sample $\hat{x}_i$ increases substantially compared to image dataset distillation. This enlarged parameter space makes trajectory matching, which directly performs pixel-wise updates on the synthetic dataset, considerably more difficult to optimize and less stable. As a result, directly applying image dataset distillation methods (Cazenavette et al., 2022; Zhao & Bilen, 2023; Guo et al., 2023) often leads to convergence issues and ultimately yields suboptimal performance in the video domain, as empirically demonstrated in Figure 2.

Frame information observation. In light of the aforementioned challenges in video set distillation, identifying and prioritizing the distillation of the most salient information becomes crucial. The temporal dimension in video dataset distillation often contains redundant information for discriminative learning tasks. Building on previous research (Zhu et al., 2018), we hypothesize that the core semantic content necessary for model training might be preserved even with a significantly reduced set of frames. This hypothesis motivates an investigation into the efficacy of learning from video sequences constructed by interpolating frames, as a potential avenue to mitigate the challenges posed by high-dimensional video inputs. We designed preliminary experiments to quantify the information a model can acquire from sparsely sampled representations.

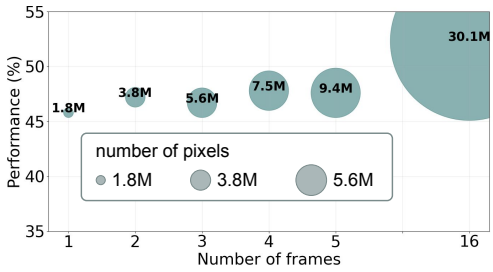


Figure 3: **Evaluation results on interpolated datasets.** With substantially fewer pixels, single-frame interpolated data retains major discriminative information comparable to full video datasets.

In our experimental setup, for each original video, we uniformly sampled n frames. These n frames are subsequently processed by an interpolation method, denoted $g(\cdot)$, to generate a new video sequence with the original temporal length f. This procedure results in an interpolated dataset $\mathcal{I} = \{(g(\tilde{x}_i), \tilde{y}_i)\}_{i=1}^{|\mathcal{T}|}$, where $\tilde{x}_i \in \mathbb{R}^{n\times c\times h\times w}$ represents the set of n sampled frames for the i-th video. A model is then trained on the dataset $\mathcal{I}$ and its performance is evaluated on the original dataset $\mathcal{T}$. The results of this evaluation are depicted in Figure 3. With videos interpolated from single frames, the model could already achieve 88% performance of the model trained on the entire 16-frame video dataset. However, the number of pixels in the dataset $\mathcal{I}$ is only about 6% of the original dataset. These findings indicate that even when a video is reconstructed from a single randomly sampled frame, a substantial portion of the information essential to the discriminative objective can be captured.This observation is essential in dataset distillation, as the majority of the dataset distillation methods (Cazenavette et al., 2022; Zhao & Bilen, 2023; Cui et al., 2023; Guo et al., 2023; Wang et al., 2024) directly updates the synthetic dataset pixel-wisely, and reducing the number of pixels is equal to reducing the number of learnable parameters during distillation process.

3.3 Single-frame Video Set Distillation

As discussed in Section 3.2, traditional dataset distillation approaches that synthesize complete video sequences suffer from an exceptionally large parameter space, making the optimization particularly

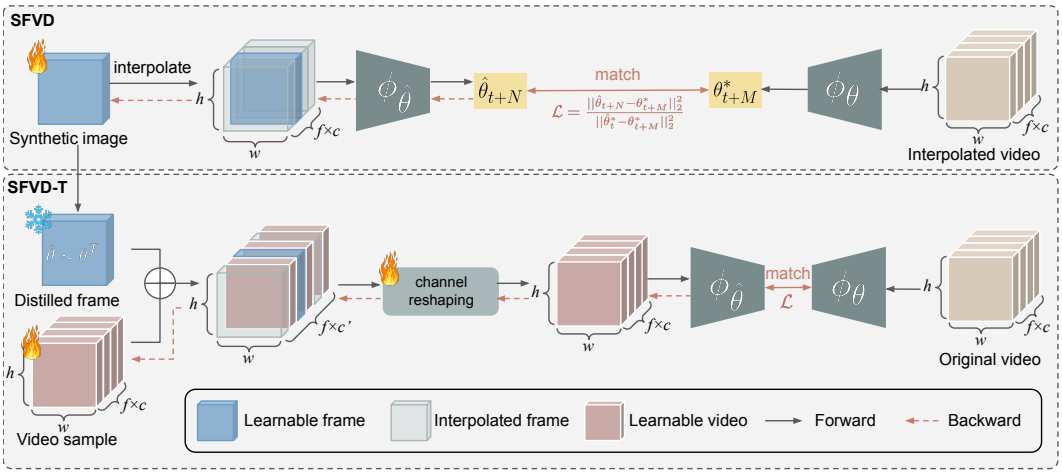


Figure 4: **The workflow of the proposed framework.** The proposed SFVD optimizes a set of synthetic images. These synthetic images are interpolated into videos and matched against interpolated videos during the distillation process. For integration of temporal information (SFVD-T), the distilled frames from SFVD is frozen. A fused representation is obtained by concatenating the distilled frames and video samples along the channel dimension. A channel reshaping layer is then employed to transform the integrate the distilled frames and video samples, and match with the original videos.

challenging. An alternative solution involves leveraging single static frames. Prior work (Wang et al., 2024) attempted to match training gradients using static frames but yielded suboptimal outcomes. We posit that a fundamental limitation of such static-frame methods is an objective mismatch. Optimizing a synthetic dataset composed of static images $\mathcal{S}_I$ to derive optimal parameters $\theta^*_{\mathcal{S}_I}$ for an image-centric model $\hat{\phi}_\theta$, does not inherently guarantee that these frames will be optimal for training a target video model ϕ_θ to achieve its optimal parameters on the original video dataset $\mathcal{T}$. As illustrated conceptually in Figure 4 (a), the model architectures and the nature of the data (static vs. dynamic) differ, creating an inconsistency. This discrepancy means that directly optimizing static frames fails to ensure their effective generalization when used to train video models.

Instead of matching between single images, we propose Single-Frame Video set Distillation (SFVD) to match videos with images. This is directly inspired by preliminary observations in Section 3.2, which demonstrate that substantial video classification performance can be preserved when models are trained on video sequences generated via differentiable interpolation $g(\cdot)$ from a single frame, which is $l(\mathcal{T};\theta^*_{\mathcal{T}}) \approx l(\mathcal{T};\theta^*_{\mathcal{I}})$. This finding suggests that a single frame is capable for conveying essential discriminative information. SFVD operationalizes this insight. The synthetic dataset in SFVD, denoted $\mathcal{S} = \{(g(\hat{x}_i), \hat{y}_i)\}_{i=1}^{|\mathcal{S}|}$, comprises learnable frames $\hat{x}_i \in \mathbb{R}^{c,h,w}$ and interpolation method $g(\cdot)$. The optimization objective aims to align the training dynamics induced by these interpolated synthetic videos with those of a target dataset. In this way, the Equation 1 becomes

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbb{E}_{\theta^{(0)}\sim\Theta}[l\left(\mathcal{I};\theta^*_{\mathcal{S}}\right)]. \tag{3}$$

Specifically, we adopt a crossfade-based interpolation strategy to interpolate video frames from a sequence of input frames. Given a sequence of sampled frames $x = \{x^j\}_{j=1}^{f'}$ where $f' \leq f$, we construct intermediate frames by linearly interpolating each adjacent pair (x^j, x^{j+1}) across $\lfloor f/(f'-1) \rfloor$ steps. The frame at step s is computed as

$$g_s(x^j, x^{j+1}) = (1-\alpha_s)x^j + \alpha_s x^{j+1}, \tag{4}$$

where $\alpha_s \in [0,1]$ increases uniformly from 0 to 1 over the interpolation steps. The complete interpolated sequence $g(x)$ is then obtained by concatenating the generated intermediate frames for all pairs (x^j, x^{j+1}) in temporal order.

Using $\mathcal{I}$ as a computationally more tractable target for trajectory matching, we adapt the trajectory matching objective from Equation 2 as follows:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbb{E}_{\theta^{(0)}\sim\Theta} \sum_{t=0}^{T} \mathcal{D}\left(\theta^{(t)}_{\mathcal{S}}, \theta^{(t)}_{\mathcal{I}}\right). \tag{5}$$

Explicitly, the objective becomes:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbb{E}_{\theta^{(0)}\sim\Theta} \sum_{t=0}^{T} \mathcal{D}\Big(\theta^{(t)}_{\{(g(\hat{x}_i),\hat{y}_i)\}_{i=1}^{|\mathcal{S}|}}, \theta^{(t)}_{\{(g(\tilde{x}_i),\tilde{y}_i)\}_{i=1}^{|\mathcal{T}|}}\Big) \tag{6}$$

The above formulation means that by optimizing synthetic frames with interpolated videos $\mathcal{I}$ using the same interpolation method, SFVD ensures that the synthetic data is directly optimized for its intended use in training video models, overcoming the limitations of the prior method (Wang et al., 2024) focusing solely on frame matching. Additionally, SFVD significantly reduces the parameter space compared to traditional distillation methods, thus facilitating stable convergence. In summary, SFVD provides an efficient and effective approach to video dataset distillation by leveraging differentiable interpolation to bridge the representational gap between the single frame and the video data, thereby mitigating the aforementioned challenge in video dataset distillation.

3.4 Integrating Temporal Information

While distilled static frames can effectively capture significant semantic information from videos, the temporal information is also crucial for comprehensive video understanding. Full video sequences contain rich temporal cues but may be computationally expensive or contain redundancies. Therefore, a critical challenge arises: how to synergistically combine the information density of distilled static representations with the essential temporal context derived from video sequences? This section introduces a methodology to achieve this integration.

Let $\mathcal{S} = \{g(\hat{x}_i), \hat{y}_i\}_{i=1}^{|\mathcal{S}|}$ denote the distilled dataset by SFVD through matching with the interpolated dataset $\mathcal{I}$. We further incorporate temporal dynamics using the workflow as shown in Figure 4 (c). Specifically, we sampled $|\mathcal{S}|$ videos from the original dataset, and for each instance, both $g(\hat{x}_i)$ and the corresponding video v_i are provided as input to a channel reshaping layer to obtained the fused representation $\{z_i\}_{i=1}^{|\mathcal{S}|}$, where $z_i \in \mathbb{R}^{f,c,h,w}$. We denote $\mathcal{M}$ as the channel reshaping layer. With distilled frame $\hat{x}_i$ and its corresponding video v_i, fused representation is obtained by

$$z_i = \mathcal{M}(g(\hat{x}_i) \oplus_c v_i)\,, \tag{7}$$

where $\oplus_c$ denotes concatenation along the channel dimension. In experiments, $\mathcal{M}$ is implemented as a lightweight 3D convolutional encoder applies a $3\times3\times3$ convolution to jointly model spatial and temporal correlations.

For both SFVD and SFVD-T, we adopt a trajectory matching objective. Let θ^*_{t+M} denote the parameters of a target model after M optimization steps on the original video dataset, starting from an initial parameter state θ^*_t. Let $\hat{\theta}_{t+N}$ denote the parameters of a model trained for N steps using the fused representations z_i, starting from the same initial state $\hat{\theta}_t = \theta^*_t$. The objective is to minimize the discrepancy between these parameter trajectories, formalized by the loss function:

$$\mathcal{L} = \frac{||\hat{\theta}_{t+N} - \theta^*_{t+M}||_2^2}{||\hat{\theta}^*_t - \theta^*_{t+M}||_2^2}, \tag{8}$$

where the parameters $\hat{\theta}$ are updated iteratively via gradient descent for $n = 0, 1, \cdots, N-1$:

$$\hat{\theta}_{t+n+1} = \hat{\theta}_{t+n} - \alpha\nabla l((z_i, \hat{y}_i); \hat{\theta}_{t+n}), \tag{9}$$

where $l(\cdot)$ represents the cross-entropy loss, and α is the learning rate. Crucially, during this optimization phase, the distilled representations $\hat{x}_i$ are treated as fixed inputs. In contrast, the parameters within the $\mathcal{M}$ and potentially any parameters involved in the processing of the video segments v_i are learnable. This allows the temporal integration mechanism to adapt and optimize the fusion process to effectively mimic the training trajectories on the original dataset.

4 Experiments

4.1 Datasets and Metrics

Our evaluation methodology aligns with prior research (Wang et al., 2024), employing the UCF101 (Soomro et al., 2012), HMDB51 (Kuehne et al., 2011), Kinetics-400 (Carreira & Zisserman, 2017), and Something-Something V2 (Goyal et al., 2017) datasets. UCF101 contains 13,320

Table 1: **Comparison with SOTA methods.** Top-1 accuracy is reported for MiniUCF and HMDB51 dataset. Top-5 accuracy is reported for Kinetics-400 and SSv2 dataset. **Bold** and Underlined values indicate the best and the second best results, respectively.

Dataset		MiniUCF		HMDB51		Kinetics-400		SSv2	
IPC		1	5	1	5	1	5	1	5
Full Dataset		57.2±0.1		28.6±0.7		34.6±0.5		29.0±0.6	
Coreset Selection	Random	9.9±0.8	22.9±1.1	4.6±0.5	6.6±0.7	3.0±0.1	5.6±0.0	3.3±0.1	3.9±0.1
	Herding (Welling, 2009)	12.7±1.6	25.8±0.3	3.8±0.2	8.5±0.4	-	-	-	-
	K-Center (Sener & Savarese, 2017)	11.5±0.7	23.0±1.3	3.1±0.1	5.2±0.3	-	-	-	-
Dataset Distillation	DM (Zhao & Bilen, 2023)	15.3±1.1	25.7±0.2	6.1±0.2	8.0±0.2	6.3±0.0	9.1±0.9	3.6±0.0	4.1±0.0
	MTT (Cazenavette et al., 2022)	19.0±0.1	28.4±0.7	6.6±0.5	8.4±0.6	3.8±0.2	9.1±0.3	3.9±0.1	6.3±0.3
	FRePo (Zhou et al., 2022)	20.3±0.5	30.2±1.7	7.2±0.8	9.6±0.7	-	-	-	-
	DATM (Guo et al., 2023)	14.6±0.3	24.9±1.1	-	-	-	-	-	-
	Static-VDSD (Wang et al., 2024)	13.7±1.1	24.7±0.5	5.1±0.9	7.8±0.4	4.6±0.2	6.6±0.2	3.9±0.1	4.1±0.0
	DM+VDSD (Wang et al., 2024)	17.5±0.1	27.2±0.4	6.0±0.4	8.2±0.1	6.3±0.2	7.0±0.1	4.0±0.1	3.8±0.1
	MTT+VDSD (Wang et al., 2024)	23.3±0.6	28.3±0.0	6.5±0.1	8.9±0.6	6.3±0.1	11.5±0.5	5.5±0.1	8.3±0.2
	FRePo+VDSD (Wang et al., 2024)	22.0±1.0	31.2±0.7	8.6±0.5	10.3±0.6	-	-	-	-
	SFVD (ours)	27.5±0.7	34.2±0.3	9.7±0.3	13.4±1.0	10.4±0.4	15.4±0.7	8.0±0.5	10.9±0.3
	SFVD-T	**27.8±0.2**	**36.5±1.2**	**10.9±0.6**	**15.6±0.5**	**11.4±0.2**	**16.2±0.6**	**8.3±1.1**	**11.7±0.1**

videos distributed across 101 action categories, whereas HMDB51 includes 6,849 videos categorized into 51 classes. Kinetics (Carreira & Zisserman, 2017) is a comprehensive collection of video clips spanning 400 human action classes. Something-Something v. 2 (SSv2) (Goyal et al., 2017) focuses on 174 motion-intensive classes. For the UCF101 and HMDB51 datasets, we report top-1 classification accuracy. In line with (Wang et al., 2024), we also utilize MiniUCF, a subset of UCF101 that includes its 50 most prevalent classes. For the SSv2 and Kinetics-400 datasets, top-5 accuracy is reported.

4.2 Implementation Details

Following previous work (Wang et al., 2024), the videos of MiniUCF and HMDB51 are sampled to 16 frames, with a sampling interval of 4 frames dynamically. Each frame is cropped and resized to 112×112. For Kinetics-400 and SSv2, the video are sampled to 8 frames and the frames are cropped to 64×64. For SFVD interpolation, the synthetic frames are duplicated 16 times on MiniUCF and HMDB51, 8 times on Kinetics-400 and Something-Something V2. Since HMDB51, SSv2, and Kinetics are highly class-imbalanced datasets, we oversample the less frequent samples within these datasets (Zhao et al., 2024). In experiments, we use crossfade for the differentiable interpolation $g(\cdot)$.

4.3 Main Results

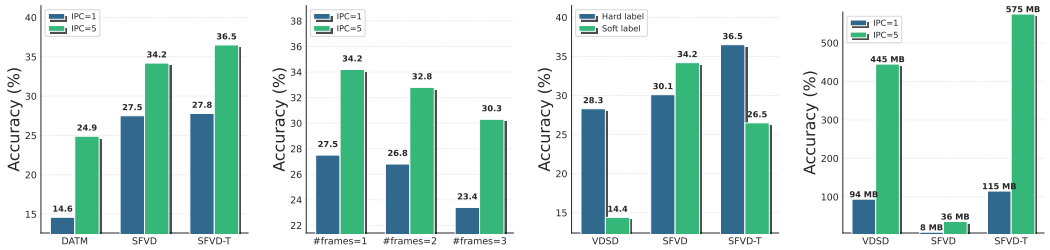
Compare with SOTA methods. We compared our method with state-of-the-art dataset distillation techniques, as well as coreset selection methods, on four widely used video datasets: MiniUCF, HMDB51, Kinetics-400, and Something-Something V2. As shown in Table 1, our SFVD method outperforms all existing approaches on the MiniUCF dataset. This is attributed to our method's more constrained parameter updating space, which helps capture discriminative features more effectively. SFVD is also highly storage-efficient, requiring only one image per instance (IPC). The integration of temporal information (SFVD-T) leads to an even better performance of the distilled dataset. We found that SFVD performs especially well with smaller IPC values, and similar performance characteristics and advantages of our methods were also observed on the HMDB51 dataset.

Table 2: **Cross-architecture evaluation.** Experiments are conducted on the MiniUCF dataset with one image per class (IPC = 1).

Method	Evaluation Model		
	ConvNet3D	CNN+GRU	CNN+LSTM
Random	9.9±0.8	6.2±0.8	6.5±0.3
DM	15.3±1.1	9.9±0.7	9.2±0.3
MTT	19.0±0.1	8.4±0.5	7.3±0.4
DM+VDSD	17.5±0.1	12.0±0.7	10.3±0.2
MTT+VDSD	23.3±0.6	14.8±0.1	13.4±0.2
ours	**27.8±0.2**	**19.3±0.5**	**18.4±0.4**

For larger video datasets such as Kinetics-400 and Something-Something V2, our proposed method also demonstrates leading performance. These more extensive and complex datasets often pose significant challenges for distillation techniques due to their increased diversity in actions, scenes, and temporal dynamics. Despite these challenges, our approach effectively identifies and preserves crucial semantic information, resulting in a distilled dataset that maintains a high level of accuracy. This indicates the robustness and scalability of our method, showcasing its capability to handle the increased complexity and data volume inherent in larger-scale video analysis tasks, while still achieving superior results.

Cross-architecture generalization. A critical challenge in dataset distillation is whether the synthesized data overfits the specific network architecture used during the distillation process. To



(a) Component ablation (b) Frames selection (c) Soft-label influence (d) Storage usage

Figure 5: **Ablation study and efficiency analysis on MiniUCF.** We systematically validate our design choices. (a) Compared to the image distillation baseline, our components significantly boost generation accuracy. (b) Counter-intuitively, constraining the synthesis to a single frame yields optimal performance while avoiding optimization collapse. (c) Soft-label guidance further improves distillation efficacy. (d) SFVD achieves this superior performance with a drastically reduced storage footprint compared to video-tensor baselines, while SFVD-T maintain a similar storage requirement.

evaluate the generalizability of our distilled MiniUCF dataset (IPC = 1), we follow the protocol of prior work (Wang et al., 2024) by testing on unseen architectures. While the data is distilled using ConvNet3D, we evaluate it on two widely used recurrent architectures: CNN+GRU (Donahue et al., 2015) and CNN+LSTM (Yue-Hei Ng et al., 2015). As shown in Table 2, all methods experience a performance drop when transferring to heterogeneous architectures. However, our proposed framework demonstrates superior robustness. Specifically, compared to the strongest baseline (MTT+VDSD), our method achieves an absolute accuracy improvement of 4.5% on CNN+GRU and 5.0% on CNN+LSTM. These results indicate that our approach synthesizes universally representative features rather than merely overfitting to the source architecture.

4.4 Ablation Studies

Ablation analysis of proposed components. To verify the individual contribution of each component in our method, we conducted a comprehensive ablation study. The empirical results, detailed in Figure 5a, demonstrate that using the SFVD module alone significantly improves performance over the baseline (Guo et al., 2023), indicating its effectiveness in capturing essential visual patterns even without temporal cues. Furthermore, when the integration of temporal information is involved, we observe a further enhancement in performance, suggesting that temporal dynamics play a complementary role in enhancing representation quality. The consistent gains across multiple evaluation metrics confirm that both components contribute meaningfully to the effectiveness of our approach.

Impact of varying the number of distilled frames. In an extension of our investigation into the distillation process, we explored alternative configurations to the single-frame distillation approach. Specifically, experiments were designed to evaluate the effects of distilling video sequences into a varying number of representative frames. The subsequent interpolation was performed consistently with the procedures outlined in Section 3.2. The quantitative evaluation outcomes are presented in Figure 5b. As the number of frames targeted for distillation increases, there is a corresponding degradation in overall performance. This empirical finding aligns with and substantiates the concerns articulated in Section 3. Specifically, an increase in the number of distilled frames leads to a significant expansion in the parameter space of the distillation model. This expansion can complicate the optimization landscape, potentially preventing the distillation process from achieving robust convergence and leading to suboptimal performance.

Investigation of soft-label integration. Recognizing the significant emphasis placed on the utility of soft labels within recent image dataset distillation works (Guo et al., 2023; Sun et al., 2024; Yin et al., 2023; Qin et al., 2024; Su et al., 2024; Shao et al., 2024a), we conducted a series of experiments to evaluate their integration to the video dataset distillation and our proposed methodology. For this investigation, the soft labels employed were derived directly from model logits, rather than utilizing more elaborate soft labels generated through complex knowledge distillation (KD) techniques (Sun et al., 2024; Yin et al., 2023; Shao et al., 2024a).

The results are presented in Table 5c. Notably, our proposed SFVD framework demonstrated effective compatibility with soft labels, maintaining or enhancing its performance. In contrast, both the baseline (Wang et al., 2024) and proposed SFVD-T variant exhibited a decline in performance



Figure 6: **Visualization of the distilled dataset.** We visualized 3 samples of the distilled datasets of MiniUCF with IPC=5. For simplicity we only show 8 frames of the samples. The exhibited samples from up to down are from "IceDancing", "Basketball", and "PommelHorse" categories.

when soft labels were introduced. This divergence in outcomes is hypothesized to stem from the increased complexity introduced by a larger number of learnable parameters during the distillation phase, particularly for VDSD and SFVD-T in conjunction with soft labels. An over-parameterized distillation process can lead to optimization challenges and instability. Consequently, to ensure robust convergence and facilitate the reliable determination of optimal parameters for temporal reshaping and video sample selection, hard labels were selected for the primary experiments, prioritizing stability and consistent results in those contexts.

Dataset storage efficiency. We evaluate the storage requirements of the distilled MiniUCF dataset, as illustrated in Figure 5d. Benefiting from its single-frame representation, SFVD maintains an exceptionally small memory footprint, requiring only 8 MB and 36 MB for IPC=1 and 5, respectively. This is a substantial reduction compared to the VDSD baseline (94 MB and 445 MB), which stores full video-sized tensors. While our enhanced variant, SFVD-T, also stores video-sized tensors and requires about 100 MB and 575 MB, its memory overhead remains comparable to VDSD. Importantly, even our most storage-intensive configuration (SFVD-T at IPC=5) consumes only a tiny fraction of the original dataset's 9.81 GB size, demonstrating the high compression efficiency of our framework.

4.5 Qualitative Results

To observe the temporal changes in the distilled videos, we sampled frames from the videos obtained using different methods and visualized their inter-frame differences. We show three samples in Figure 6 and more in the Appendix D. Although visually abstract, we can still conclude that the distilled videos contain information from more than a single video and indeed exhibit temporal variations. These variations suggest that the distilled videos are not static repetitions but carry meaningful transitions over time, capturing diverse motion patterns that contribute to effective video model training. This validates the effectiveness of the proposed method in preserving temporal information within the distilled dataset.

5 Conclusion

In this paper, we confronted the critical challenge of a drastically increased number of parameters in video dataset distillation. We introduced the novel Single-Frame Video set Distillation (SFVD) framework, premised on the insight that individual frames often contain substantial discriminative information. SFVD initially distills highly representative frames, reducing synthetic data dimensionality and simplifying optimization. Subsequently, it integrates essential temporal information using a channel reshaping layer that fuses these distilled frames with video samples, and then matches with the original video features. Extensive experiments confirmed that SFVD significantly outperforms existing methods, validating our strategy of first tackling dimensionality and then incorporating temporal information. This research underscores a successful approach to the unique challenges of video data, offering a path towards more efficient and effective distilled video datasets and scalable training for video recognition models, while opening avenues for future exploration in advanced temporal modeling and broader task applications.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist in the writing of proofs, and implement methods. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, find related work, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or interpret results. Generating synthetic datasets is not applicable to this work. We have reviewed all AI-assisted work. We proofread the LLM-polished writing, and checked the code implementation. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021.

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

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, 2022.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *CVPR*, 2023.

Yilan Chen, Wei Huang, and Lily Weng. Provable and efficient dataset distillation for kernel ridge regression. In *NeurIPS*, 2024.

Jiacheng Cui, Xinyue Bi, Yaxin Luo, Xiaohan Zhao, Jiacheng Liu, and Zhiqiang Shen. Fadrm: Fast and accurate data residual matching for dataset distillation. *arXiv preprint arXiv:2506.24125*, 2025a.

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *ICML*. PMLR, 2023.

Xiao Cui, Yulei Qin, Xinyue Li, Wengang Zhou, Hongsheng Li, and Houqiang Li. Rethinking long-tailed dataset distillation: A uni-level framework with unbiased recovery and relabeling. *arXiv preprint arXiv:2511.18858*, 2025b.

Xiao Cui, Yulei Qin, Wengang Zhou, Hongsheng Li, and Houqiang Li. Optical: Leveraging optimal transport for contribution allocation in dataset distillation. In *CVPR*, 2025c.

Wenxiao Deng, Wenbin Li, Tianyu Ding, Lei Wang, Hongguang Zhang, Kuihua Huang, Jing Huo, and Yang Gao. Exploiting inter-sample and inter-feature relations in dataset distillation. In *CVPR*, 2024.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *CVPR*, 2023.

Jiawei Du, Juncheng Hu, Wenxin Huang, Joey Tianyi Zhou, et al. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *NeurIPS*, 2024.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 2019.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

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.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *CVPR*, 2024.

Shengbin Guo, Hang Zhao, Senqiao Yang, Chenyang Jiang, Yuhang Cheng, Xiangru Peng, Rui Shao, and Zhuotao Tian. Multimodal dataset distillation via phased teacher models. In *ICLR*, 2026.

Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*, 2023.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011.

Yongmin Lee and Hye Won Chung. Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. *arXiv preprint arXiv:2406.18561*, 2024.

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Dataset distillation using parameter pruning. *IEICE Transactions*, 2023.

Longzhen Li, Guang Li, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Foreground-aware dataset distillation via dynamic patch selection. *arXiv preprint arXiv:2601.02727*, 2026.

Muyang Li, Can Cui, Quan Liu, Ruining Deng, Tianyuan Yao, Marilyn Lionts, and Yuankai Huo. Dataset distillation in medical imaging: A feasibility study. *arXiv preprint arXiv:2407.14429*, 2024.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 2020.

Wenyuan Li, Guang Li, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Hyperbolic dataset distillation. *arXiv preprint arXiv:2505.24623*, 2025.

Haoyang Liu, Yijiang Li, Tiancheng Xing, Vibhu Dalal, Luwei Li, Jingrui He, and Haohan Wang. Dataset distillation via the wasserstein metric. *arXiv preprint arXiv:2311.18531*, 2023.

Ping Liu and Jiawei Du. The evolution of dataset distillation: Toward scalable and generalizable solutions. *arXiv preprint arXiv:2502.05673*, 2025.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022.

Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *NeurIPS*, 2022.

Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. Dataset distillation with convexified implicit gradients. In *ICML*. PMLR, 2023.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. *arXiv preprint arXiv:2011.00050*, 2020.

Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. In *NeurIPS*, 2021.

Matt Poyser and Toby P Breckon. Neural architecture search: A contemporary literature review for computer vision applications. *Pattern Recognition*, 2024.

Tian Qin, Zhiwei Deng, and David Alvarez-Melis. A label is worth a thousand images in dataset distillation. *arXiv preprint arXiv:2406.10485*, 2024.

Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *ICCV*, 2023.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Xinyi Shang, Peng Sun, and Tao Lin. Gift: Unlocking full potential of labels in distilled dataset at near-zero cost. *arXiv preprint arXiv:2405.14736*, 2024.

Yuzhang Shang, Zhihang Yuan, and Yan Yan. Mim4dd: Mutual information maximization for dataset distillation. In *NeurIPS*, 2023.

Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *CVPR*, 2024a.

Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. In *NeurIPS*, 2024b.

Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, and Shitong Shao. Delt: A simple diversity-driven earlylate training for dataset distillation. In *CVPR*, 2025.

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.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ 4: Dataset distillation via disentangled diffusion model. In *CVPR*, 2024.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *CVPR*, 2024.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

Minh-Tuan Tran, Trung Le, Xuan-May Le, Thanh-Toan Do, and Dinh Phung. Enhancing dataset distillation via non-critical region refinement. In *CVPR*, 2025.

Haoxuan Wang, Zhenghao Zhao, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Cao _2: Rectifying inconsistencies in diffusion-based dataset distillation. *arXiv preprint arXiv:2506.22637*, 2025a.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *CVPR*, 2022.

Kai Wang, Zekai Li, Zhi-Qi Cheng, Samir Khaki, Ahmad Sajedi, Ramakrishna Vedantam, Konstantinos N Plataniotis, Alexander Hauptmann, and Yang You. Emphasizing discriminative features for dataset distillation in complex scenarios. In *CVPR*, 2025b.

Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective. In *CVPR*, 2025c.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Ziyu Wang, Yue Xu, Cewu Lu, and Yong-Lu Li. Dancing with still images: video distillation via static-dynamic disentanglement. In *CVPR*, 2024.

Max Welling. Herding dynamical weights to learn. In *ICML*, 2009.

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *arXiv preprint arXiv:2308.07545*, 2023.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018.

Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. *arXiv preprint arXiv:2406.03793*, 2024.

Shaolei Yang, Shen Cheng, Mingbo Hong, Haoqiang Fan, Xing Wei, and Shuaicheng Liu. Neural spectral decomposition for dataset distillation. In *ECCV*, 2024a.

William Yang, Ye Zhu, Zhiwei Deng, and Olga Russakovsky. What is dataset distillation learning? *arXiv preprint arXiv:2406.04284*, 2024b.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *NeurIPS*, 2023.

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*. PMLR, 2017.

Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. M3d: Dataset condensation by minimizing maximum mean discrepancy. In *AAAI*, 2024a.

Jiaqing Zhang, Mingjia Yin, Hao Wang, Yawen Li, Yuyang Ye, Xingyu Lou, Junping Du, and Enhong Chen. Td3: Tucker decomposition based dataset distillation method for sequential recommendation. In *ACM Web Conference 2025*, 2025.

Xin Zhang, Jiawei Du, Ping Liu, and Joey Tianyi Zhou. Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. *arXiv preprint arXiv:2408.06927*, 2024b.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*. PMLR, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *CVPR*, 2023.

Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, and Yan Yan. Distilling long-tailed datasets. *arXiv preprint arXiv:2408.14506*, 2024.

Wenliang Zhong, Haoyu Tang, Qinghai Zheng, Mingzhu Xu, Yupeng Hu, and Weili Guan. Towards stable and storage-efficient dataset distillation: Matching convexified trajectory. In *CVPR*, 2025.

Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *NeurIPS*, 2022.

Chen Zhu, Xiao Tan, Feng Zhou, Xiao Liu, Kaiyu Yue, Errui Ding, and Yi Ma. Fine-grained video categorization with redundancy reduction attention. In *ECCV*, 2018.

A Training Details

We report the most important hyperparameters in the SFVD distillation in Table 3, where `lr_img` stands for the learning rate to update the synthetic frames, `lr_y` denotes the learning rate for the soft-label optimizer, `lr_lr` represents the learning rate for adaptive lr, `batch_syn` indicates the number of synthetic frames to match at each iteration, and `syn_steps` is the steps of training trajectories of the synthetic dataset to match expert trajectories.

Table 3: Hyperparameters for SFVD on video datasets.

Dataset	IPC	lr_img	lr_y	lr_lr	batch_syn	syn_steps
miniUCF	1	1000	10	1×10^{-5}	50	40
	5	1000	10	1×10^{-5}	50	80
HMDB51	1	1000	10	1×10^{-5}	51	40
	5	1000	10	1×10^{-5}	51	80
Kinetics	1	1000	10	1×10^{-5}	32	40
	5	1000	10	1×10^{-5}	32	40
SSv2	1	1000	10	1×10^{-5}	32	40
	5	1000	10	1×10^{-5}	32	40

B Further Analysis

Computational cost. We evaluate the computational efficiency of our approach against existing methods by measuring the total GPU hours required for distillation. All computational costs are measured on NVIDIA A6000 GPUs. As shown in Table 4, our proposed SFVD framework achieves a significant reduction in training cost while maintaining strong performance. Specifically, while DATM (Guo et al., 2023) requires 58.8 GPU hours for distillation on MiniUCF, SFVD drastically reduces this cost to merely 14.1 GPU hours, yielding an over $4\times$ speedup. Furthermore, our enhanced variant, SFVD-T, requires an additional 7.4 GPU hours to incorporate soft labels, leading to a total of 21.5 GPU hours. Notably, even our most computationally intensive setting (SFVD-T) remains substantially more efficient than prior state-of-the-art methods, including MTT (Cazenavette et al., 2022) and VDSD (Wang et al., 2024). These results demonstrate the superior computational efficiency and scalability of our framework.

Table 4: **Computational cost comparison.** Experiments are conducted on the MiniUCF dataset.

Method	MTT	DATM	VDSD	SFVD	SFVD-T
GPU hours	29.2	58.8	48.8	14.1	21.5

Different sampling strategies. Explicit frame sampling is exclusively utilized in our multi-frame distillation ablations. For these experiments, we adopt uniform sampling to ensure consistency with standard video training pipelines and prior work. As detailed in Table 5, evaluating alternative sampling strategies yielded no meaningful performance improvements over uniform sampling. Crucially, our empirical results demonstrate that optimal performance is achieved when distilling a single frame per video. In this single-frame regime, heuristic frame selection becomes redundant, as the optimization directly synthesizes a representative frame rather than selecting from pre-existing ones. This observation reinforces our central premise: under current frameworks, constraining the learnable space is more critical than refining frame selection heuristics.

Table 5: **Ablation on frame sampling.** Alternative heuristics yield no significant gains over uniform sampling.

Sampling	frames=1	frames=2	frames=3
Uniform	34.2±0.3	32.8±0.4	30.3±0.6
Clip	34.2±0.3	31.6±0.4	29.7±0.5
Optical flow	33.6±0.5	32.2±0.3	30.7±0.4

C Limitations

Video dataset distillation is currently in its early stages. While image dataset distillation methods have achieved nearly lossless performance on small datasets (Guo et al., 2023), our approach to video set distillation still shows a performance gap compared to using the full dataset.

D More Qualitative Results

We included more visualization of the distilled dataset to provide a clearer understanding of the visual properties of the generated samples in Figure 7. The additional examples cover multiple action categories from UCF101, illustrating the temporal changes accross the frames. Each sample exhibits distinct motion cues and spatial semantics representative of its target class, showing that the distilled data effectively captures discriminative visual patterns despite its compactness. These qualitative results further validate that the proposed SFVD framework can synthesize visually coherent and semantically meaningful representations.

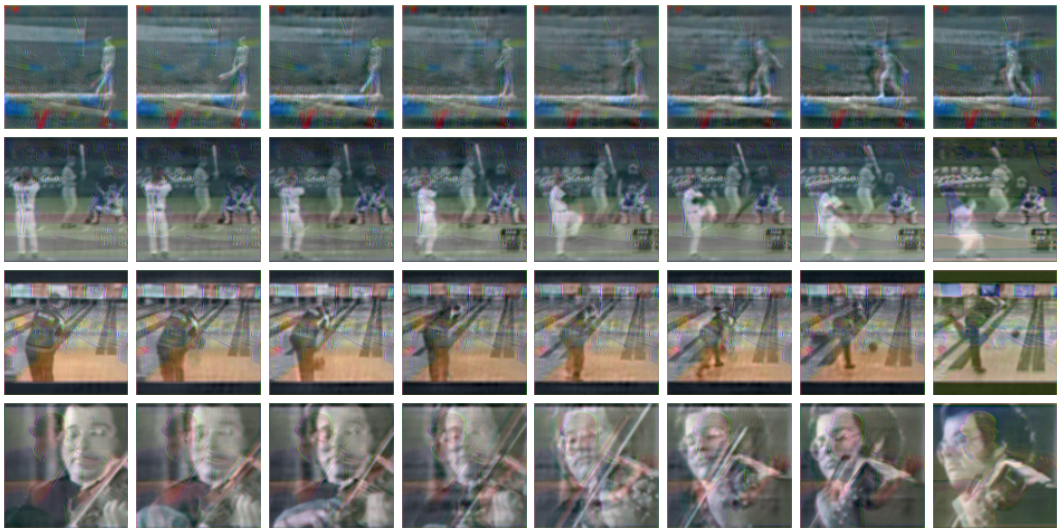


Figure 7: Qualitative results of the distilled dataset. From top to bottom, the classes depicted are "BalanceBeam", "BaseballPitch", "Bowling", and "PlayingViolin."

E Theoretical Justification.

In this section, we present a theoretical justification for the use of interpolated videos as surrogates for the original video dataset. The key idea is to bound the difference between the expected training loss on the true dataset and the corresponding loss on the interpolated dataset. This establishes that the interpolation process does not significantly distort the underlying optimization objective.

Our interpolation procedure replaces each original video x with a reconstructed video $g(\tilde{x})$ obtained from sampled frames. This raises a natural question: to what extent does replacing the dataset $\mathcal{T}$ with its interpolated counterpart $\mathcal{I}$ affect the expected loss used for training. The following lemma provides a formal bound on this difference.

Lemma 1. *Let $\mathcal{T}$ be the original dataset over (x, y) and let $\mathcal{I}$ be the interpolated dataset defined by $(x', y) = (g(\tilde{x}), y)$ for each (x, y). Let ϕ_θ be a network and let L be a loss. Assume both ϕ_θ and L are Lipschitz with constants C_ϕ and C_L. Then*

$$\big|L(\mathcal{T};\theta) - L(\mathcal{I};\theta)\big| \le C_L C_\phi \, \mathbb{E}_{(x,y)\sim\mathcal{T}}\big[\|x - g(\tilde{x})\|\big]. \tag{10}$$

In particular, if $\mathbb{E}_{(x,y)\sim\mathcal{T}}\big[\|x - g(\tilde{x})\|\big] \le \varepsilon$, then

$$\big|L(\mathcal{T};\theta) - L(\mathcal{I};\theta)\big| \le C\,\varepsilon \quad \textit{with } C := C_l C_\phi. \tag{11}$$

Proof. Fix any (x, y) drawn from $\mathcal{T}$ and let $x' = g(\tilde{x})$ be the corresponding interpolated video in $\mathcal{I}$. By the Lipschitz continuity of the loss L in its first argument we have

$$\big|L(\phi_\theta(x), y) - L(\phi_\theta(x'), y)\big| \le C_L \, \|\phi_\theta(x) - \phi_\theta(x')\|. \tag{12}$$

Applying the Lipschitz continuity of ϕ_θ with respect to the input yields

$$\|\phi_\theta(x) - \phi_\theta(x')\| \le C_\phi \, \|x - x'\| = C_\phi \, \big\|x - g(\tilde{x})\big\|. \tag{13}$$

Combining the two inequalities gives the pointwise bound

$$\big|L(\phi_\theta(x), y) - L(\phi_\theta(g(\tilde{x})), y)\big| \le C_L C_\phi \, \big\|x - g(\tilde{x})\big\|. \tag{14}$$

Taking expectation over $(x, y) \sim \mathcal{T}$ on both sides we obtain

$$\begin{aligned} &\big|L(\mathcal{T};\theta) - L(\mathcal{I};\theta)\big| \\ &= \Big|\mathbb{E}_{(x,y)\sim\mathcal{T}}\big[L(\phi_\theta(x), y) - L(\phi_\theta(g(\tilde{x})), y)\big]\Big| \\ &\le \mathbb{E}_{(x,y)\sim\mathcal{T}}\big[\big|L(\phi_\theta(x), y) - L(\phi_\theta(g(\tilde{x})), y)\big|\big] \\ &\le C_L C_\phi \, \mathbb{E}_{(x,y)\sim\mathcal{T}}\big[\|x - g(\tilde{x})\|\big], \end{aligned} \tag{15}$$

which is exactly Eq. equation 10. If the expected reconstruction error of g is bounded by ε then the last term is at most $C_L C_\phi \varepsilon$, hence the second statement follows with $C := C_L C_\phi$.