LATENT VIDEO DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset distillation has demonstrated remarkable effectiveness in high-compression scenarios for image datasets. While video datasets inherently contain greater redundancy, existing video dataset distillation methods primarily focus on compression in the pixel space, overlooking advances in the latent space that have been widely adopted in modern text-to-image and text-to-video models. In this work, we bridge this gap by introducing a novel video dataset distillation approach that operates in the latent space using state-of-the-art variational auto-encoders. Furthermore, we employ a diversity-aware data selection strategy to select both representative and diverse samples. Additionally, we introduce a simple, training-free method to further compress the distilled latent dataset. By combining these techniques, our approach achieves new state-of-the-art performances in video dataset distillation, outperforming prior methods on all datasets, e.g. on HMDB51 IPC 1, we achieve a 2.6% performance increase; on MiniUCF IPC 5, we achieve a 7.8% performance increase.

1 INTRODUCTION

Dataset distillation has emerged as a critical technique for compressing large-scale datasets into computationally efficient representations that retain their essential characteristics (Wang et al., 2018). While this technique has seen success in compressing image datasets (Cui et al., 2023; 2022; Loo et al., 2022; Nguyen et al., 2020; Wang et al., 2022; Zhao & Bilen, 2021), applications onto video datasets remain an underexplored challenge. Videos inherently possess temporal redundancy, as characterized by consecutive frames often sharing substantial similarity, presenting the potential for optimization via dataset distillation.

Existing video distillation methods predominantly focus on pixel-space compression. VDSD (Wang et al., 2024) addresses the temporal information redundancy by disentangling static and dynamic information. Method IDTD (Zhao et al., 2024) tackles the within-sample and inter-sample redundancies by leveraging a joint-optimization framework. However, these frameworks overlook the potential of latent-space compressions, which have proven transformative in generative models for images and videos (Tong et al., 2022; Zhao et al., 2025). Modern variational autoencoders (VAEs) (Welling, 2009; Ranganath et al., 2014) offer a pathway to address this gap by encoding videos into compact and disentangled representations in latent space.

In this work, we improve video distillation by operating entirely in the latent space of a VAE. Our framework distills videos into low-dimensional latent codes, leveraging the VAE's ability to model temporal dynamics (Zhao et al., 2025). Unlike previous methods, our approach encodes entire video sequences into coherent latent trajectories to model temporal dynamics through its hierarchical architecture. We compress the VAE itself through post-training quantization, largely reducing the model size, while retaining accuracy (Cui et al., 2023). After distillation, we apply Diversity-Aware Data Selection using Determinantal Point Processes (DPPs) (Kulesza & Taskar, 2012) to select both representative and diverse instances. Unlike clustering-based or random sampling methods, DPPs favor diversity by selecting samples that are well-spread in the latent space, reducing redundancy while ensuring comprehensive feature coverage (Nava et al., 2022). This leads to a more informative distilled dataset that enhances downstream model generalization.

Our method further introduces a training-free latent compression strategy, which uses high-order singular value decomposition (HOSVD) to decompose spatiotemporal features into orthogonal subspaces (Wang et al., 2024). This isolates dominant motion patterns and spatial structures, enabling further compression while preserving essential dynamics (Tong et al., 2022). By factorizing latent tensors, we dynamically adjust the rank of the distilled representations, allowing denser instance packing under fixed storage limits. Experiments on the MiniUCF dataset demonstrate that our method outperforms prior pixel-space approaches by 11.5% in absolute accuracy for IPC 1 and 7.8% for IPC 5. Overall, our contributions are:

1. We propose the first video dataset distillation framework operating in the latent space, leveraging a state-of-the-art VAE to efficiently encode spatiotemporal dynamics.

2. We address the challenge of spatiotemporal redundancy in the video latent space by integrating VAEs, Diversity-Aware Data Selection using DPPs and High-Order Singular Value Decomposition (HOSVD) into a structured compression framework.

3. Our method generalizes to both small-scale and large-scale video datasets, achieving a new state-of-the-art performance on all settings compared to existing methods.

2 Related Work

2.0.1 Coreset Selection

Coreset selection aims to identify a small but representative subset of data that preserves the essential properties of the full dataset. One of the foundational approaches utilizes k-center clustering (Sener & Savarese, 2018) to formulate coreset selection as a geometric covering problem, where a subset of data points is chosen to maximize the minimum distance to previously selected points. By iteratively selecting the most distant samples in feature space, this method ensures that the coreset provides broad coverage of the dataset's distribution, making it a strong candidate for reducing redundancy in large-scale datasets. Herding methods (Welling, 2009) take an optimization-driven approach to coreset selection by sequentially choosing samples that best approximate the mean feature representation of the dataset. Probabilistic techniques leverage Bayesian inference (Manousakas et al., 2020) and divergence minimization (Tiwary et al., 2023) to construct coresets that balance diversity and statistical representativeness. Influence-based selection methods (Yang et al., 2022) instead focus on quantifying the contribution of individual samples to generalization performance, retaining only the most impactful data points.

2.0.2 Image Dataset Distillation

Dataset distillation (Wang et al., 2018) has emerged as a powerful paradigm for compressing large-scale image datasets while preserving downstream task performance. DC improved dataset distillation with aligning the single-step gradients of synthetic and real data (Zhao et al., 2020). Further, meta-learning frameworks like Matching Training Trajectories (MTT) (Cazenavette et al., 2022) and Kernel Inducing Points (KIP) (Nguyen et al., 2021) advances performance by distilling datasets through bi-level optimization over neural architectures. Dataset condensation with Distribution Matching (DM) (Zhao & Bilen, 2023) synthesizes condensed datasets by aligning feature distributions between original and synthetic data across various embedding spaces.

Representative Matching for Dataset Condensation (DREAM) (Liu et al., 2023) improved sample efficiency by selecting representative instances that retained the most informative patterns from the original dataset. Generative modeling techniques have also been explored, with Distilling Datasets into Generative Models (DiM) (Wang et al., 2023) encoding datasets into latent generative spaces, allowing for smooth interpolation and novel sample generation. And Dataset Distillation via Disentangled Diffusion Model (D4M) (Su et al., 2024) created latent prototypes which were used to generate synthetic distilled images with diffusion models. Similarly, Hybrid Generative-Discriminative Dataset Distillation (GDD) (Li et al., 2024) balanced global structural coherence with fine-grained detail preservation by combining adversarial generative models with traditional distillation objectives. However, temporal redundancy and frame sampling complexities, as noted in (Huang et al., 2018; Liu et al., 2021), highlight the unique difficulties of extending image-focused distillation to video datasets.

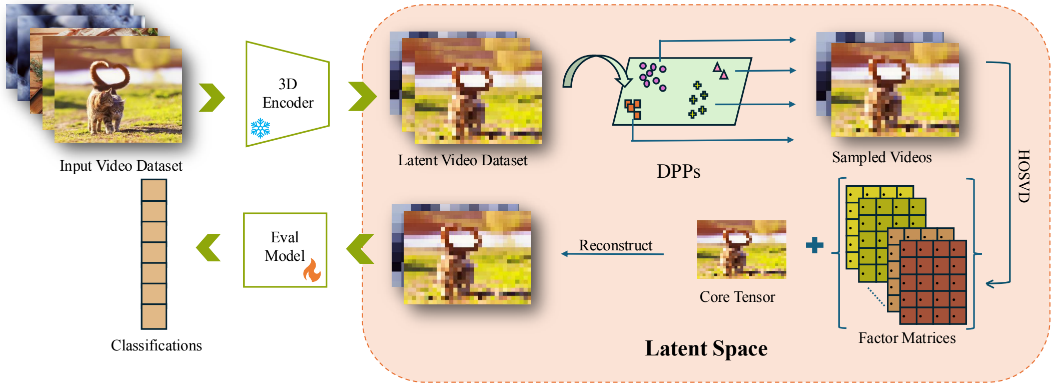

Figure 1: Our training-free latent video distillation pipeline. The entire video dataset is encoded into latent space with a VAE. We further employ the DPPs to select both representative and diverse samples, followed by latent space compression with HOSVD for efficient storage.

2.0.3 Video Dataset Distillation

While dataset distillation has achieved great improvements in static image datasets, direct application to videos presents unique challenges due to temporal redundancy and the need for efficient frame selection (Tong et al., 2022). Recent attempts to address video dataset distillation have primarily focused on pixel-space compression. Video Distillation via Static-Dynamic Disentanglement (VDSD) (Wang et al., 2024) tackles temporal redundancies between frames by separating static and dynamic components. VDSD partitions videos into smaller segments and employs learnable dynamic memory block that captures and synthesizes motion patterns, improving information retention while reducing redundancy. IDTD (Zhao et al., 2024) addresses the challenges of within-sample redundancy and inter-sample redundancy simultaneously. IDTD employs an architecture represented by a shared feature pool alongside multiple feature selectors to selectively condense video sequences. To retain the temporal information of synthesized videos, IDTD introduces a temporal fusor that integrates diverse features into the temporal dimension.

2.0.4 Text-to-Video Models and Their Role in Latent Space Learning

Latent-space representations have become a cornerstone of modern video modeling, offering structured compression while maintaining high-level semantic integrity (Tong et al., 2022; Zhao et al., 2025). Variational autoencoders enable efficient storage and reconstruction (Kingma et al., 2013). Extending this concept, hierarchical autoregressive latent prediction (Seo et al., 2022) introduces an autoregressive component that improves temporal coherence, leading to high-fidelity video reconstructions. Further enhancing latent representations, latent video diffusion transformers (Ma et al., 2024) incorporate diffusion-based priors to refine video quality while minimizing storage demands.

Building upon these latent space techniques, recent text-to-video models have presented their capability to generate high-resolution video content from textual descriptions. These methods employ a combination of transformer-based encoders and diffusion models to synthesize realistic video sequences. Imagen Video leverages cascaded video diffusion models to progressively upsample spatial and temporal dimensions, ensuring high-quality output (Ho et al., 2022). Meanwhile, zero-shot generation approaches utilize decoder-only transformer architectures to process multimodal inputs, such as text and images, without requiring explicit video-text training data (Kondratyuk et al., 2023). Hybrid techniques combining pixel-space and latent-space diffusion modeling further enhance computational efficiency while maintaining visual fidelity by leveraging learned latent representations during synthesis (Zhang et al., 2023). These advancements in latent space learning not only improve video compression but also drive the development of scalable and high-quality text-driven video generation.

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	3.2 ± 0.1	3.7 ± 0.0
	Herding	12.7 ± 1.6	25.8 ± 0.3	3.8 ± 0.2	8.5 ± 0.4	4.3 ± 0.3	8.0 ± 0.1	4.6 ± 0.3	6.8 ± 0.2
	K-Center	11.5 ± 0.7	23.0 ± 1.3	3.1 ± 0.1	5.2 ± 0.3	3.9 ± 0.2	5.9 ± 0.4	3.8 ± 0.5	4.0 ± 0.1
Dataset Distillation	DM	15.3 ± 1.1	25.7 ± 0.2	6.1 ± 0.2	8.0 ± 0.2	6.3 ± 0.0	9.1 ± 0.9	4.1 ± 0.4	4.5 ± 0.3
	MTT	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 ± 0.2	6.5 ± 0.2
	FRePo	20.3 ± 0.5	30.2 ± 1.7	7.2 ± 0.8	9.6 ± 0.7	–	–	–	–
	DM+VDSD	17.5 ± 0.1	27.2 ± 0.4	6.0 ± 0.4	8.2 ± 0.1	6.3 ± 0.2	7.0 ± 0.1	4.3 ± 0.3	4.0 ± 0.3
	MTT+VDSD	23.3 ± 0.6	28.3 ± 0.0	6.5 ± 0.1	8.9 ± 0.6	6.3 ± 0.1	11.5 ± 0.5	5.7 ± 0.2	8.4 ± 0.1
	FRePo+VDSD	22.0 ± 1.0	31.2 ± 0.7	8.6 ± 0.5	10.3 ± 0.6	–	–	–	–
	IDTD	22.5 ± 0.1	33.3 ± 0.5	9.5 ± 0.3	16.2 ± 0.9	6.1 ± 0.1	12.1 ± 0.2	–	–
	Ours	$\mathbf{34.8 \pm 0.5}$	$\mathbf{41.1 \pm 0.6}$	$\mathbf{12.1 \pm 0.3}$	$\mathbf{17.6 \pm 0.4}$	$\mathbf{9.0 \pm 0.1}$	$\mathbf{13.8 \pm 0.1}$	$\mathbf{6.9 \pm 0.6}$	$\mathbf{10.5 \pm 0.4}$

Table 1: Performance comparison between our method and existing baselines on both small-scale and large-scale datasets. Follow previous works, we report Top-1 test accuracies (%) for small-scale datasets and Top-5 test accuracies (%) for large-scale datasets.

3 Methodology

In this section, we first introduce the variational autoencoder (VAE) used to encode video sequences into a compact latent space. We then discuss our Diversity-Aware Data Selection method. Next, we present our training-free latent space compression approach using High-Order Singular Value Decomposition (HOSVD). Finally, we describe our two-stage dynamic quantization strategy. The entire pipeline of our framework is shown in Fig. 1.

3.1 Preliminary

3.1.1 Problem Definition

In video dataset distillation, given a large dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{T}|}$ consisting of video samples x_i and their corresponding class labels y_i, the objective is to construct a significantly smaller distilled dataset $\mathcal{S} = \{\tilde{x}_i, \tilde{y}_i\}_{i=1}^{|\mathcal{S}|}$, where $|\mathcal{S}| \ll |\mathcal{T}|$. The distilled dataset is expected to achieve comparable performance to the original dataset on action classification tasks while significantly reducing storage and computational requirements.

3.1.2 Latent image distillation

Latent image distillation has emerged as an effective alternative to traditional dataset distillation methods. Instead of distilling datasets at the pixel level, latent distillation leverages pre-trained autoencoders or generative models to encode images into a compact latent space. Latent Dataset Distillation with Diffusion Models (Moser et al., 2024), have shown that distilling image datasets in the latent space of a pre-trained diffusion model improves generalization and enables higher compression ratios. Similarly, Dataset Distillation in Latent Space (Duan et al., 2023) adapts conventional distillation methods like Gradient Matching, Feature Matching, and Parameter Matching to the latent space, significantly reducing computational overhead while achieving competitive performance. Different from these methods, we extend latent space distillation to video datasets by encoding both spatial and temporal information into the latent space.

3.1.3 Variational Autoencoder

Variational Autoencoders (VAEs) (Kingma et al., 2013) provide compact latent representations by encoding inputs into a probabilistic latent space while maintaining the ability to reconstruct the original data. Unlike deterministic autoencoders, VAEs learn a distribution $q_\phi(z|x)$ over latent variables, encouraging continuity and smooth interpolation in the latent space. This probabilistic structure is essential for dataset distillation, as it ensures that the compressed representations remain expressive and robust to variation.

In our framework, we employ a VAE to process full video clips. The encoder maps an input sequence $x \in \mathbb{R}^{T \times H \times W \times C}$ into a latent distribution $q_\phi(z|x)$ parameterized by mean and variance, where z is a spatiotemporal latent tensor. Samples from this distribution capture both spatial content and temporal dynamics in a compact form. The decoder then reconstructs the original clip from z

through $p_\theta(x|z)$, ensuring that the latent codes retain the essential motion and appearance patterns needed for downstream tasks.

Training follows the standard Evidence Lower Bound (ELBO) objective:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)] + \beta \cdot D_{\text{KL}}(q_\phi(z|x) \parallel p(z)) \tag{1}$$

where the first term enforces reconstruction fidelity and the second term regularizes the latent distribution towards the prior $p(z) = \mathcal{N}(0, 1)$. The balance parameter β controls the trade-off between compression and information preservation.

3.2 Diversity-Aware Data Selection

After encoding the entire video dataset into the latent space using a state-of-the-art VAE, an effective data selection strategy is crucial to maximize the diversity and representativeness of the distilled dataset. To this end, we employ Diversity-Aware Data Selection using Determinantal Point Processes (DPPs) (Kulesza & Taskar, 2012), a principled probabilistic framework that promotes diversity by favoring sets of samples that are well-spread in the latent space.

DPPs provide a mechanism for selecting a subset of latent embeddings that balance coverage and informativeness. Given the encoded latent representations of the dataset, we construct a similarity kernel matrix L, where each entry L_{ij} quantifies the pairwise similarity between latent samples z_i and z_j. The selection process then involves sampling from a determinantal distribution parameterized by L, ensuring that the chosen subset is both diverse and representative of the full latent dataset. We define a kernel matrix L using the following function:

$$L_{ij} = \exp(-\frac{\parallel z_i - z_j \parallel^2}{2\sigma^2}) \tag{2}$$

Then subset S is sampled according to:

$$P(S) = \frac{\det(L_S)}{\det(L + I)} \tag{3}$$

here L_S is the submatrix of L that corresponds to the rows and columns indexed by S. The denominator $\det(L + I)$ serves as a normalization factor, ensuring that the probabilities across all possible subsets sum to 1. This normalization stabilizes the sampling process by incorporating an identity matrix I, which prevents numerical instability in cases where L is near-singular.

Our approach is motivated by the observation that naive random sampling or traditional clustering-based selection strategies (Ikotun et al., 2023) tend to underperform in high-dimensional latent spaces (Ghilotti et al., 2023), where redundancy is prevalent. By leveraging DPPs, we effectively capture a more comprehensive distribution of video features. Furthermore, the computational efficiency of DPPs allows us to scale our selection process to large datasets without significant overhead.

Applying DPPs in the latent space instead of the pixel space offers several key advantages. First, latent representations encode high-level semantic features, making it possible to directly select samples that preserve meaningful variations in motion and structure, rather than relying on pixel-wise differences that may be redundant or noisy. Second, the latent space is significantly more compact and disentangled, allowing DPPs to operate more effectively with reduced computational complexity compared to pixel-space selection (Wang et al., 2024). Finally, in the latent space, similarity measures are inherently more structured, which makes DPPs better suited for ensuring diverse and representative selections.

3.3 Training-free Latent Space Compression

While our Diversity-Aware Data Selection reduces sample-level redundancy, the resulting latent tensors still contain substantial spatiotemporal redundancy, since consecutive frames often encode overlapping motion patterns.

A natural approach for redundancy reduction is Singular Value Decomposition (SVD), which compresses matrices by discarding low-energy components. However, applying SVD requires flattening video tensors into 2D matrices, thereby destroying spatial and temporal structure.

To address this, we employ High-Order Singular Value Decomposition (HOSVD), which generalizes SVD to multi-dimensional tensors while preserving correlations across modes. Given a latent tensor $Z \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_n}$, HOSVD decomposes it into a compact core tensor $\mathcal{G}$ and orthonormal factor matrices U_i:

$$Z = \mathcal{G} \times_1 U_1 \times_2 U_2 \times \cdots \times_n U_n. \tag{4}$$

By truncating along the temporal mode, we explicitly remove redundant motion information while retaining dominant dynamics. Similarly, spatial and channel-wise redundancy can be reduced by low-rank approximation in the corresponding modes. Importantly, this procedure is entirely training-free, scalable, and seamlessly integrates into our pipeline after DPP-based selection.

3.4 VAE Quantization

To improve storage efficiency, we apply a two-stage post-training quantization to the 3D-VAE (Zhao et al., 2025), combining dynamic quantization for fully connected layers and mixed-precision optimization for the remaining layers.

In the first stage, all fully connected layers are quantized from FP32 to INT8 using dynamic scaling of activations and weights. This significantly reduces memory and computation while preserving inference stability, since matrix multiplications in fully connected layers exhibit high redundancy and are well-suited for integer quantization (Hu et al., 2024).

In the second stage, convolutional and batch normalization layers are compressed from FP32 to FP16. Mixed-precision is preferred here because convolutional operations are more sensitive to precision loss, and FP16 provides sufficient dynamic range to maintain reconstruction quality (Yun et al., 2023).

This hybrid quantization yields over a 2.6× reduction in VAE model size with negligible loss in reconstruction fidelity, ensuring that the encoder remains compact while effectively modeling spatiotemporal dependencies in video sequences.

4 Experiments

4.1 Datasets and Metrics

Following previous works VDSD (Wang et al., 2024) and IDTD (Zhao et al., 2024), we evaluate our proposed video dataset distillation approach on both small-scale and large-scale benchmark datasets. For small-scale datasets, we utilize MiniUCF (Wang et al., 2024) and HMDB51 (Kuehne et al., 2011), while for large-scale datasets, we conduct experiments on Kinetics (Carreira & Zisserman, 2017) and Something-Something V2 (SSv2) (Goyal et al., 2017). MiniUCF is a miniaturized version of UCF101 (Soomro et al., 2012), consisting of the 50 most common action classes selected from the original UCF101 dataset. HMDB51 is a widely used human action recognition dataset containing 6,849 video clips across 51 action categories. Kinetics is a large-scale video action recognition dataset, available in different versions covering 400, 600, or 700 human action classes. SSv2 is a motion-centric video dataset comprising 174 action categories.

4.2 Baselines

Based on previous work, we include the following baseline: (1) coreset selection methods such as random selection, Herding (Welling, 2009), and K-Center (Sener & Savarese, 2018), and (2) dataset distillation methods including DM (Zhao & Bilen, 2023), MTT (Cazenavette et al., 2022), FRePo (Zhou et al., 2022), VDSD (Wang et al., 2024), and IDTD (Zhao et al., 2024). DM (Zhao & Bilen, 2023) ensures that the models trained on the distilled dataset produce gradient updates similar to those trained on the full dataset. MTT (Cazenavette et al., 2022) improves distillation by aligning model parameter trajectories between the synthetic and original datasets. FRePo (Zhou et al., 2022) focuses on generating compact datasets that allow pre-trained models to quickly recover their original performance with minimal training. VDSD (Wang et al., 2024) introduces a static-dynamic disentanglement approach for video dataset distillation. IDTD (Zhao et al., 2024) enhances video dataset distillation by increasing feature diversity across samples while densifying temporal information within instances.

4.3 Implementation Details

4.3.1 Dataset Details

For small-scale datasets, MiniUCF and HMDB51, we follow the settings from previous work (Wang et al., 2024; Zhao et al., 2024), where videos are dynamically sampled to 16 frames with a sampling interval of 4. Each sampled frame is then cropped and resized to 112×112 resolution. We adopt the same settings as prior work (Wang et al., 2024; Zhao et al., 2024) for Kinetics-400, each video is sampled to 8 frames and resized to 64×64, maintaining a compact representation suitable for large-scale dataset distillation. In Something-Something V2 (SSv2), which is relatively smaller among the two large-scale datasets, we sample 16 frames per video and resize them to 112×112, demonstrating the scalability of our method across datasets of varying sizes.

4.3.2 Evaluation Network

Following the previous works, we use a 3D convolutional network, C3D (Tran et al., 2015) as the evaluation network. C3D (Tran et al., 2015) is trained on the distilled datasets generated by our method. Similar to previous works, we assess the performance of our distilled datasets by measuring the top-1 accuracy on small-scale datasets and top-5 accuracy on large-scale datasets.

4.3.3 Fair Comparison

Throughout our experiments, we rigorously ensure that the total storage space occupied by the quantized VAE model and the decomposed matrices remain within the constraints of the corresponding Instance Per Class (IPC) budget. Specifically, on SSv2, our method utilizes no more than 68% of the storage space allocated to the baseline methods DM and MTT, guaranteeing a fair and consistent comparison. We made sure the combined size of the quantized VAE and latent tensor are within the storage budget for fair comparison, as shown in Tab. 2. Our storage is sublinear and more scalable to large datasets with higher resolutions.

Dataset	MiniUCF		HMDB51		Kinetics-400		SSv2	
IPC	1	5	1	5	1	5	1	5
DM	115	586	115	586	150	765	400	2039
MTT	115	586	115	586	150	765	400	2039
VDSD	94	455	94	455	123	591	327	1583
Ours	107	475	107	475	148	455	223	458

Table 2: Storage (in MB) analysis of prior methods and ours.

4.4 Experimental Results

In Tab. 1, we present the performance of our method across MiniUCF (Wang et al., 2024), HMDB51 (Kuehne et al., 2011), Kinetics-400 (Carreira & Zisserman, 2017), and SSv2 (Goyal et al., 2017) under both IPC 1 and IPC 5 settings.

On MiniUCF, our approach outperforms the best baseline (IDTD) by 12.3% under IPC 1, achieving 34.8% accuracy compared to 22.5%, and by 7.8% under IPC 5, reaching 41.1% accuracy. Similarly, on HMDB51, our method achieves 12.1% accuracy under IPC 1, surpassing the strongest baseline by 2.6%, while under IPC 5, it reaches 17.6%, a 1.4% improvement. These results highlight the effectiveness of our latent-space distillation framework, which provides superior compression efficiency and classification performance compared to pixel-space-based approaches. The consistent performance gains across both IPC settings demonstrate the robustness of our method in preserving essential video representations while achieving high compression efficiency.

Furthermore, the results in Kinetics-400 and SSv2 reinforce our findings, as our approach consistently outperforms all baselines. Improvements in low-IPC regimes (IPC 1) suggest that our training-free latent compression and diversity-aware data selection are particularly effective when dealing with extreme data reduction. Our method achieves 9.0% accuracy on Kinetics-400 IPC 1, outperforming the strongest baseline (IDTD) by 2.9%, and 6.9% accuracy on SSv2 IPC 1, surpassing VDSD by 2.2%. The trend continues in IPC 5, where our model achieves 13.8% on Kinetics-400 and 10.5% on SSv2, both establishing new state-of-the-art results in video dataset distillation.

4.5 Ablation Study

In this section, we systematically analyze the key components of our method to understand their contributions to overall performance. We evaluate on cross-architecture generalization, various sampling methods, different rank compression ratios in HOSVD, and different latent space compression techniques. We also provide comparison with traditional video compression methods and diversity analysis.

	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
DM + VDSD	17.5 ± 0.1	12.0 ± 0.7	10.3 ± 0.2
MTT	19.0 ± 0.1	8.4 ± 0.5	7.3 ± 0.4
MTT + VDSD	23.3 ± 0.6	14.8 ± 0.1	13.4 ± 0.2
Ours	$\mathbf{34.8 \pm 0.5}$	$\mathbf{19.9 \pm 0.7}$	$\mathbf{18.3 \pm 0.7}$

Table 3: Result of experiment on cross-architecture generalization for MiniUCF when IPC is 1.

4.5.1 Cross Architecture Generalization

To further evaluate the generalization capability of our method, we conduct experiments on cross-architecture generalization, as presented in Tab. 3. The results demonstrate that datasets distilled using our method consistently achieve superior performance across different evaluation models—ConvNet3D, CNN+GRU, and CNN+LSTM—compared to previous state-of-the-art methods.

Our approach achieves 34.8% accuracy with ConvNet3D, significantly surpassing all baselines, including MTT+VDSD (23.3%) and DM+VDSD (17.5%). Notably, our method also outperforms all baselines when evaluated on recurrent-based architectures (CNN+GRU and CNN+LSTM), obtaining 19.9% and 18.3% accuracy, respectively. This highlights the robustness of our distilled dataset in preserving spatiotemporal coherence, which is crucial for models leveraging sequential dependencies, and validates our advantage over traditional compression methods.

4.5.2 Rank Compression Ratio

We evaluate the impact of different rank compression ratios in HOSVD on overall performance in Tab. 4. Empirical results show that a rank compression ratio of $r = 0.75$ consistently provides a strong balance between storage efficiency and model accuracy across datasets. While increasing the compression ratio reduces storage requirements, overly aggressive compression can lead to significant information loss, negatively affecting downstream tasks. Notably, as shown in Tab. 4, when the rank compression ratio is set to $r = 0.1$, both datasets exhibit classification accuracy around 4.0%, suggesting that excessive compression leads to degraded latent representations, making the distilled dataset nearly indistinguishable from random noise.

	Rank Compression Ratio				
	0.10	0.25	0.50	0.75	1.00
MiniUCF	4.1 ± 0.1	19.0 ± 1.3	31.5 ± 0.7	$\mathbf{34.8 \pm 0.5}$	28.9 ± 0.5
HMDB51	3.9 ± 0.6	7.6 ± 1.0	11.5 ± 0.1	$\mathbf{12.1 \pm 0.3}$	8.9 ± 0.5

Table 4: Accuracies under different rank compression ratios. Both MiniUCF and HMDB51 datasets are evaluated under IPC 1.

4.5.3 HOSVD vs Classic SVD

To evaluate the effectiveness of our latent-space compression strategy, we compare truncated SVD with HOSVD under the same storage budget at IPC 5. Truncated SVD is a matrix factorization technique that approximates a data matrix by keeping only its largest singular values, thus reducing dimensionality while retaining the most informative components. However, SVD operates on flattened data matrices, leading to a loss of structural information, particularly in spatiotemporal representations.

As shown in Tab. 5, HOSVD consistently outperforms truncated SVD across all datasets, demonstrating its ability to better preserve spatial and temporal dependencies in the latent space. On Kinetics-400 and SSv2, HOSVD achieves higher classification accuracy (+1.4% and +1.2%, respectively), highlighting its advantage in handling large-scale datasets.

Dataset	MiniUCF	HMDB51	Kinetics-400	SSv2
SVD	38.5 ± 0.4	15.8 ± 0.2	12.4 ± 0.3	9.3 ± 0.2
HOSVD	$\mathbf{41.1 \pm 0.6}$	$\mathbf{17.6 \pm 0.4}$	$\mathbf{13.8 \pm 0.1}$	$\mathbf{10.5 \pm 0.4}$

Table 5: Classification accuracies comparison between different latent compression techniques under the same storage budget for each dataset at IPC 5.

4.5.4 Traditional Video Compression Methods

To further demonstrate the applicability of our method, we compared it with two widely used traditional video compression methods, VP9 (Google, 2013) and H.264 (Wiegand et al., 2003), in the context of video dataset distillation. As shown in Tab.6, our method consistently outperforms both VP9 and H.264 across all four datasets. This is likely because traditional video compression primarily aims to reduce bitrate while preserving perceptual visual quality, whereas our method explicitly selects representative samples and compresses them spatiotemporally to better retain task-relevant information for downstream learning.

Dataset	MiniUCF		HMDB51		Kinetics-400		SSv2	
IPC	1	5	1	5	1	5	1	5
VP9	25.8	37.5	7.3	14.8	4.8	7.2	4.3	5.6
H.264	24.6	38.2	8.5	15.4	5.5	7.8	3.7	4.5
Ours	**34.8**	**41.1**	**12.1**	**17.6**	**9.0**	**13.8**	**6.9**	**10.5**

Table 6: Performance of Traditional Video Compression Methods

5 Conclusion

In this work, we introduce a novel latent-space video dataset distillation framework that leverages VAE, DPPs, and HOSVD to achieve state-of-the-art performance with efficient storage. We carefully selected DPPs and HOSVD for video distillation after investigating existing baselines and exploring alternatives such as KDE, KMeans, SVD, and PCA. Our method provides a simple yet effective solution by significantly reducing both temporal and spatial redundancy. Unlike prior works such as VDSD, which trains a dedicated network for spatiotemporal modeling, we leverage a pre-trained VAE and HOSVD to efficiently compress this information, achieving better performance with lower computational cost. Moreover, our plug-and-play design enables seamless integration with future state-of-the-art VAEs, allowing continuous improvements as models evolve.

5.0.1 Limitations & Future Work

While our method shows strong performance, we plan to explore learning-based approaches to enhance dataset distillation, aiming to improve both efficiency and generalization in future work. We also intend to investigate non-linear decomposition techniques for latent-space compression, which could offer more compact and expressive representations than linear methods. Although our method is exclusive to video distillation given the nature of the temporal redundancy of video datasets, we plan to extend our method to image dataset distillation where temporal redundancy is not applicable.

6 Reproducibility statement

We have made significant efforts to ensure the reproducibility of our work. Detailed descriptions of our models, training procedures, dataset preprocessing step, and evaluation protocols are provided in the Section 4.3 and Appendix B. In addition, we provide an anonymous GitHub repository link in Appendix B containing our implementation and scripts to reproduce experiments.

References

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

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

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. *arXiv preprint arXiv:2207.09639*, 2022.

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6565–6590, 2023.

Yuxuan Duan, Jianfu Zhang, and Liqing Zhang. Dataset distillation in latent space. *arXiv preprint arXiv:2311.15547*, 2023.

Lorenzo Ghilotti, Mario Beraha, and Alessandra Guglielmi. Bayesian clustering of high-dimensional data via latent repulsive mixtures. *arXiv preprint arXiv:2303.02438*, 2023. URL https://arxiv.org/abs/2303.02438.

Google. Vp9 video codec. https://www.webmproject.org/vp9/, 2013. Accessed: 2025-07-26.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The ”something something” video database for learning and evaluating visual common sense, 2017.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Xing Hu, Yuan Cheng, Dawei Yang, Zhihang Yuan, Jiangyong Yu, Chen Xu, and Sifan Zhou. I-llm: Efficient integer-only inference for fully-quantized low-bit large language models. *arXiv preprint arXiv:2405.17849*, 2024.

De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *CVPR*, pp. 7366–7375, 2018. doi: 10.1109/CVPR.2018.00769.

Abiodun M Ikotun, Absalom E Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, pp. 2556–2563, 2011. doi: 10.1109/ICCV.2011.6126543.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012. doi: 10.1561/2200000044. URL https://doi.org/10.48550/arXiv.1207.6083.

Longzhen Li, Guang Li, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Generative Dataset Distillation: Balancing global structure and local details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Workshop*, pp. 7664–7671, 2024.

Xin Liu, Silvia L. Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C. van Gemert. No frame left behind: Full video action recognition. In *CVPR*, pp. 14892–14901, June 2021.

Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. Dream: Efficient dataset distillation by representative matching. *arXiv preprint arXiv:2302.14416*, 2023.

Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. *arXiv preprint arXiv:2210.12067*, 2022.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Dionysis Manousakas, Zuheng Xu, Cecilia Mascolo, and Trevor Campbell. Bayesian pseudocoresets. In *NeurIPS*, 2020.

Brian B Moser, Federico Raue, Sebastian Palacio, Stanislav Frolov, and Andreas Dengel. Latent dataset distillation with diffusion models. *arXiv preprint arXiv:2403.03881*, 2024.

Elvis Nava, Mojmir Mutny, and Andreas Krause. Diversified sampling for batched bayesian optimization with determinantal point processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 7031–7054. PMLR, 2022.

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.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *AISTATS*, 2014.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Younggyo Seo, Kimin Lee, Fangchen Liu, Stephen James, and Pieter Abbeel. Harp: Autoregressive latent video prediction with high-fidelity image generator. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3943–3947. IEEE, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

StabilityAI. Improved autoencoders ... https://huggingface.co/stabilityai/sd-vae-ft-mse, n.d. Accessed: 2025-07-26.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ 4: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5809–5818, 2024.

Piyush Tiwary, Kumar Shubham, Vivek Kashyap, et al. Constructing bayesian pseudo-coresets using contrastive divergence. *arXiv preprint arXiv:2303.11278*, 2023.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *NeurIPS*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015. doi: 10.1109/ICCV.2015.510.

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, Jianyang Gu, Daquan Zhou, Zheng Zhu, Wei Jiang, and Yang You. Dim: Distilling dataset into generative model. *arXiv preprint arXiv:2303.04707*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6296–6304, June 2024.

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

T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003. doi: 10.1109/TCSVT.2003.815165.

Shuo Yang, Zeke Xie, Hanyu Peng, Minjing Xu, Mingming Sun, and P. Li. Dataset pruning: Reducing training data by examining generalization influence. *ArXiv*, abs/2205.09329, 2022.

Juyoung Yun, Sol Choi, Francois Rameau, Byungkon Kang, and Zhoulai Fu. Standalone 16-bit training: Missing study for hardware-limited deep learning practitioners. *arXiv preprint arXiv:2305.10947*, 2023.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.

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

Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *Advances in Neural Information Processing Systems*, 37:12847–12871, 2025.

Yinjie Zhao, Heng Zhao, Bihan Wen, Yew-Soon Ong, and Joey Tianyi Zhou. Video set distillation: Information diversification and temporal densification. *arXiv preprint arXiv:2412.00111*, 2024.

Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *arXiv preprint arXiv:2206.00719*, 2022.