PRISM: Video Dataset Condensation with Progressive Refinement and Insertion for Sparse Motion

Supplementary Material

A. Proof of Lemma 1

Lemma 1 (Loss-Descent Blockage under Gradient Misalignment) *Let $s_t = \alpha s_{k_i} + (1-\alpha)s_{k_{i+1}}$, with $0 < \alpha < 1$, be a linearly interpolated frame between two key-frames s_{k_i} and $s_{k_{i+1}}$. Let the task-loss gradients be denoted as*

$$g_t = \nabla_{s_t}\mathcal{L}(s_t), \quad g_i = \nabla_{s_{k_i}}\mathcal{L}(s_{k_i}),$$

$$g_{i+1} = \nabla_{s_{k_{i+1}}}\mathcal{L}(s_{k_{i+1}}).$$

Suppose

$$\langle g_t,\, g_i\rangle < 0 \quad \textit{and} \quad \langle g_t,\, g_{i+1}\rangle < 0.$$

Then, for every convex combination

$$v = \lambda(-g_i) + (1-\lambda)(-g_{i+1}), \quad \lambda \in [0,1],$$

the following inequality holds:

$$\langle g_t,\, v\rangle > 0.$$

Consequently, no first-order update obtained by modifying only the two endpoint frames can decrease $\mathcal{L}$ at s_t; the loss is stationary or strictly increasing along every such direction. Therefore, s_t must be promoted to the key-frame set and directly optimized to enable further loss minimization.

Proof. By the bilinearity of the inner product,

$$\langle g_t, v\rangle = \lambda\langle g_t, -g_i\rangle + (1-\lambda)\langle g_t, -g_{i+1}\rangle.$$

Applying the assumption $\langle g_t, g_i\rangle < 0$, we obtain

$$\langle g_t, -g_i\rangle = -\langle g_t, g_i\rangle > 0,$$

and similarly,

$$\langle g_t, -g_{i+1}\rangle = -\langle g_t, g_{i+1}\rangle > 0.$$

Therefore,

$$\langle g_t, v\rangle = \lambda\cdot\langle g_t, -g_i\rangle + (1-\lambda)\cdot\langle g_t, -g_{i+1}\rangle > 0.$$

This shows that the directional derivative of $\mathcal{L}$ at s_t along any direction v formed by adjusting only the endpoints is positive:

$$D_v\mathcal{L}(s_t) = \langle\nabla\mathcal{L}(s_t), v\rangle > 0.$$

Thus, no first-order update along such directions can reduce the loss at s_t, and $\mathcal{L}(s)$ is strictly increasing along all directions spanned by $-g_i$ and $-g_{i+1}$. It follows that further loss minimization requires directly optimizing s_t as a key-frame. □

B. Hyperparameter

In Table A, we show the learning rate and batch size under each dataset and IPC. The ϵ is set to 0 for all experiments throughout the manuscript. The warm-up and cool-down phases are processed for 20% of the whole iteration each. In other words, if the condensation process is set to 100 iterations, the warm-up phase takes up the first 20 iterations and the cool-down phase takes up the last 20 iterations, leaving 80 iterations for the progressive refinement and insertion of frames. We follow the setting from the prior method [**?**] for evaluation and cross-architecture evaluation.

Justification for experimental scale. We emphasize that our experimental setup, using a lightweight backbone, miniC3D, low resolutions (e.g., 64x64), and a small number of frames (T=8 or T=16), follows the established protocol from prior work in video dataset condensation [**?**]. This constrained setting is a necessary consequence of the task's extreme computational demands. Unlike standard model training, dataset condensation requires an iterative optimization process to synthesize the data itself. This process, whether through gradient matching or distribution matching, requires repeated forward and backward passes through the network to update the synthetic data, making it orders of magnitude more costly than a standard training epoch. Therefore, using larger models (e.g., I3D), high resolutions (e.g., 224x224), or long sequences is computationally prohibitive for current condensation methods. Our goal is thus to demonstrate the relative efficacy and efficiency of the condensation method within this feasible, standardized environment.

Effect of Initialization Strategy We validate our "start-small" initialization approach by comparing it against a variant where a full-length video is interpolated from Gaussian noise endpoints, and all interpolated frames are treated as learnable parameters from the very beginning. Experimental results show that this fully-learnable initialization severely degrades performance, yielding only 5.69% on HMDB51 and 15.15% on miniUCF at 1 VPC. We observe that optimization becomes temporally imbalanced in this

Table A. Hyperparameters for PRISM under different datasets and IPC.

Method	Dataset	Train			Evaluation	
		IPC	LR	Batch Real	Epoch	LR
PRISM	MiniUCF	1	1	64	500	$1e^{-2}$
		5	25	64		
		10	50	64		
	HMDB51	1	0.7	64		
		5	25	64		
		10	75	64		
	Kinetics-400	1	1	64		
		5	50	128		
	SSv2	1	3	64		
		5	30	128		

setting; gradient updates concentrate heavily on the middle frames, leaving the endpoint frames weakly optimized. This supports our design choice to start with a minimal anchor set and allocate learnable capacity progressively only when gradient conflicts indicate true non-linear motion.

C. ϵ Sweep

The sweep over ϵ reveals a clear unimodal trend centered at $\epsilon = \mathbf{0}$ which produces the highest accuracy across both miniUCF and HMDB51 as shown in Table B. This value is critical because it represents the point where the cosine similarity transitions from a positive to a negative correlation. When we apply the insertion criterion $\cos_i^t < \epsilon$ with $\epsilon = \mathbf{0}$, the condition for insertion becomes $\cos(\cdot) < \mathbf{0}$. Geometrically, this requires the gradient $\mathbf{g_t}$ of the interpolated frame to be pointing into the opposite half-space (i.e., making an angle greater than $90°$) relative to both adjacent key-frame gradients $\mathbf{g_i}$ and $\mathbf{g_{i+1}}$. This directional disagreement is the most direct signal that the interpolated frame s_t lies in a region of the loss landscape where the current linear interpolation between the key-frames $(\mathbf{s_{k_i}}, \mathbf{s_{k_{i+1}}})$ cannot model the motion. The existence of an interpolated point whose optimal update direction is pointing away from the update directions of its anchors signifies a sharp, non-linear change in the spatiotemporal content, demanding the insertion of a new key-frame.

If ϵ is too negative (e.g., -0.3), the criterion becomes overly strict, allowing only the most severely misaligned gradients to trigger insertion. As a result, too few key-frames are added, limiting PRISM's ability to capture non-linear motion transitions, a phenomenon consistent with the "without insertion" ablation (Table 6(A) in the main manuscript) that demonstrated reduced performance. Conversely, when ϵ is positive (e.g., 0.2 or 0.3), the threshold becomes overly permissive, causing many interpolated frames with mildly differing gradients to be unnecessarily promoted. This leads to noisy or redundant key-frame sets and destabilizes optimization. Thus, $\epsilon = 0$ provides the ideal, non-arbitrary balance by defining the precise point of directional disagreement that signals non-linearity.

D. Warm-Up and Cool-Down Sensitivity Sweep

The warm-up and cool-down sensitivity sweep shows that intermediate scheduling (0.2) yields the strongest performance, while both extremes degrade results. This pattern reflects PRISM's reliance on stable insertion dynamics throughout training. A longer warm-up (i.e., 0.3) delays frame insertion excessively, preventing PRISM from capturing early non-linear transitions and resulting in under-refinement of the condensed video. On the other hand, removing warm-up entirely (i.e., 0) allows frames to be inserted before the initial key-frames have stabilized exposing insertion decisions to noisy early gradients that lead to suboptimal or redundant frames being selected. Similarly, an extended cool-down prematurely stops insertion and prevents late, meaningful refinements, while removing cool-down entirely allows late insertions that do not receive enough optimization steps, degrading the final representation. The peak at 0.2 arises because it provides a balanced curriculum—early enough stabilization to avoid noisy insertions and late enough freezing to prevent under-optimized or overly large key-frame sets. This balance mirrors the observed necessity of both warm-up and cool-down in the main ablation, further validating PRISM's temporal curriculum.

E. Qualitative Results

We visualize the condensed videos on HMDB51 and MiniUCF under the 1 VPC setting for maximal clarity. The visualized frames in Figure B and Figure C correspond to those retained after the condensation process, where the noise images are placeholders which does not get stored along with condensed data.

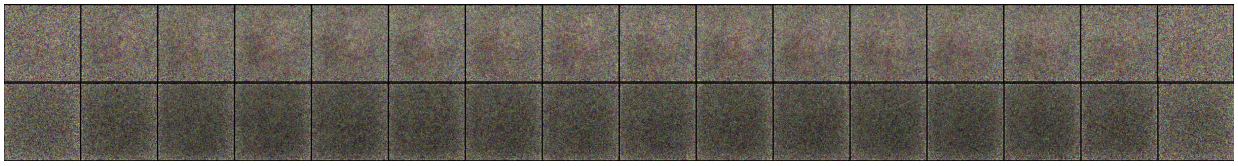

Figure A. Visualized sequences of the condensed frames.

Table B. ϵ Sweep on MiniUCF and HMDB51 dataset on 1 IPC.

Dataset	ϵ						
	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
MiniUCF	16.2	16.6	16.8	**17.9**	17.6	17.4	16.9
HMDB51	6.4	6.5	6.8	**7.5**	7.4	6.8	6.6

Table C. Warm-Up and Cool-Down Sensitivity on MiniUCF and HMDB51 dataset on 1 IPC.

Dataset	0.3	0.2	0.1	0
MiniUCF	17.6	**17.9**	17.4	17.3
HMDB51	7.3	**7.5**	7.1	6.8

Red rectangles highlight the negative effect when the warm-up phase is omitted. As consistently observed across both datasets, removing the warm-up leads to excessive frame selection, resulting in redundant and less informative synthetic frames while consuming more memory.

Blue rectangles indicate frames produced when the cool-down phase is omitted. Although overall results appear more stable than in the warm-up-removed case, we observe that some frames are added during the final few iterations of condensation. These late-added frames often lack sufficient training, reducing their utility for action recognition by being not fully trained.

F. The Use of Large Language Models (LLMs)

We used a Large Language Model (LLM) to assist with improving the clarity, grammar, and organization of the text. All scientific contributions, including the core methodology, experimental design, and analysis of results, are solely the work of the authors.

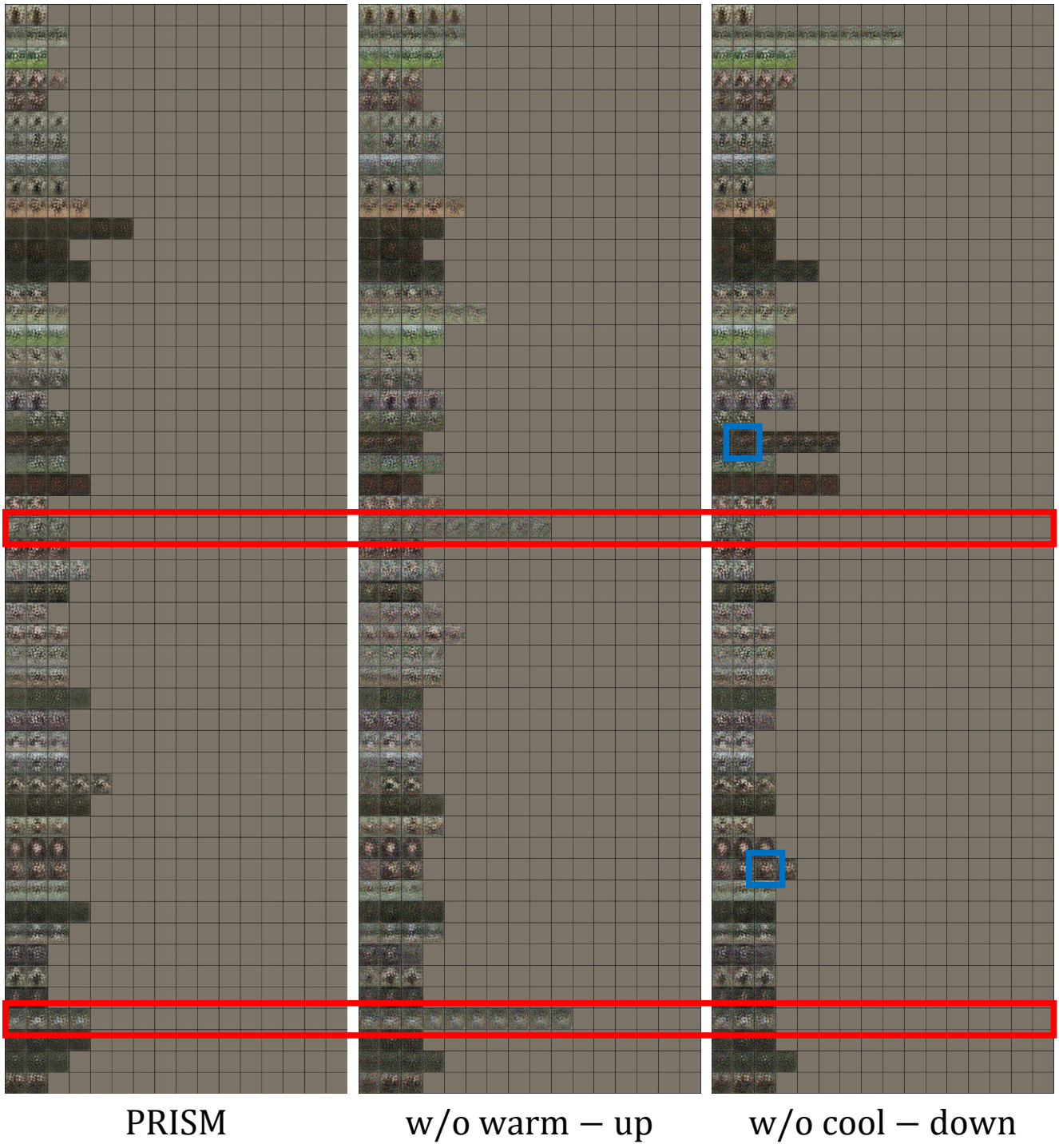

Figure B. Visualization of PRISM, PRISM without warm-up, and PRISM without cool-down on HMDB51 under 1 VPC. Red rectangles highlight the negative effects of omitting the warm-up phase, while blue rectangles indicate frames that may be under-trained due to the absence of a cool-down phase.

Figure C. Visualization of PRISM, PRISM without warm-up, and PRISM without cool-down on MiniUCF under 1 VPC. Red rectangles highlight the negative effects of omitting the warm-up phase, while blue rectangles indicate frames that may be under-trained due to the absence of a cool-down phase.