# Supplementary Material

PEEK: Picking Essential frames via Efficient Knowledge distillation

Anonymous BMVC Submission

## 1 Supplementary Material

This supplementary material provides additional implementation details for the teacher scoring stage and for the downstream captioning protocol used in our experiments.

### 1.1 Teacher scoring details (SigLIP 2)

We use the public `google/siglip2-so400m-patch14-384` checkpoint, with an SO400M backbone, $14\times14$ patches, and $384\times384$ input resolution. The model weights are frozen. Candidate frames are decoded at 2 fps and passed as RGB PIL images to the official SigLIP 2 processor, which resizes them to $384\times384$ and applies the model's default normalization. Each frame is encoded independently by the vision tower; the teacher therefore sees no temporal context. Segment captions are tokenized as-is, padded or truncated to a fixed length of 64 tokens, and encoded by the text tower.

Let $\mathbf{z}_t$ denote the pooled visual embedding of frame $f_t$, and let $\mathbf{u}$ denote the pooled text embedding of the segment caption. We explicitly L2-normalize both embeddings and compute the teacher score as their cosine similarity:

$$s_t = \frac{\langle \mathbf{z}_t, \mathbf{u} \rangle}{\|\mathbf{z}_t\|_2 \|\mathbf{u}\|_2}. \tag{1}$$

We do not use the SigLIP 2 sigmoid head. The resulting per-frame cosine scores are stored as teacher supervision. In the main ListMLE objective, these scores are used through their induced ranking; min–max normalized scores are used for visualization and analysis.

### 1.2 Downstream captioning protocol

For each segment and frame budget $k$, the $k$ selected frames are sorted in temporal order and passed as a single multi-image user message to the downstream VLM, followed by one short captioning prompt. We do not concatenate frames into a grid, do not insert timestamps, and do not add explicit "frame $i$ of $k$" delimiters.

**Checkpoints.** We use the public instruct checkpoints
`Qwen/Qwen2.5-VL-3B-Instruct`,
`Qwen/Qwen2.5-VL-7B-Instruct`,
`HuggingFaceTB/SmolVLM2-2.2B-Instruct`, and `Qwen/Qwen3.5-4B`.
All models are run in `bfloat16` with SDPA attention.

**Prompts.** The captioning prompt depends on the downstream VLM:

- **Qwen2.5-VL-3B:** "Describe this video in one sentence."
- **Qwen2.5-VL-7B and SmolVLM2-2.2B:** "Generate one short sentence (under 12 words) describing what is happening, like 'a man is playing guitar' or 'a woman is cooking'."
- **Qwen3.5-4B:** "Write a single short caption (5–10 words) describing the main action in the video."

**Decoding and generation settings.** Decoding is deterministic. We set `do_sample=False` and use the default `num_beams=1`, so generation is greedy and `temperature`, `top_p`, and `top_k` are not consulted. We otherwise keep each checkpoint's released `generation_config.json`. In particular, Qwen2.5-VL-3B and Qwen2.5-VL-7B use their default `repetition_penalty`$=1.05$, while SmolVLM2-2.2B and Qwen3.5-4B run pure argmax decoding, with no default `repetition_penalty`. We set `max_new_tokens`$=50$ for Qwen2.5-VL-3B, Qwen2.5-VL-7B, and SmolVLM2-2.2B, and `max_new_tokens`$=20$ for Qwen3.5-4B.

**Image resolution.** Image resolution follows each VLM processor's default settings, except for Qwen3.5-4B, where we cap `image_max_pixels` at $401{,}408$. If this causes an out-of-memory error, we retry with `image_max_pixels` set to $200{,}704$.

**Qwen3.5 "thinking" mode.** For Qwen3.5-4B, we disable "thinking" mode by setting `enable_thinking=False`. We also strip any residual `<think>...</think>` block from the decoded output before scoring.

## 1.3 ListMLE vs. MSE + pairwise ranking loss ablation

For the loss ablation, we keep the same frozen MobileCLIP2-S0 frame encoder, temporal scorer architecture, training data, and optimization hyperparameters. Both variants are trained on ActivityNet Captions training segments using the same SigLIP 2 teacher targets. The scorer uses $512$-D frame embeddings, a 2-layer Transformer with hidden size $256$, $4$ attention heads, FFN size $1024$, and dropout $0.15$. We train for $25$ epochs with batch size $1024$, AdamW learning rate $2{\times}10^{-4}$, weight decay $0.03$, $2$ warmup epochs, and gradient clipping at norm $1.0$.

Let $s_i$ be the raw SigLIP 2 cosine teacher score for frame $i$ in a segment. Training uses per-segment min–max normalized targets

$$y_i = \begin{cases} \dfrac{s_i - \min_j s_j}{\max_j s_j - \min_j s_j}, & \max_j s_j > \min_j s_j, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\hat{y}_i$ denote the student score for frame $i$ after the scorer head. For ListMLE, the output activation is the identity. For the MSE + pairwise baseline, the output activation is a sigmoid, so $\hat{y}_i \in (0, 1)$.

For ListMLE, frames are sorted by decreasing teacher target. If $\pi$ is the permutation such that $y_{\pi_1} \geq y_{\pi_2} \geq \cdots \geq y_{\pi_n}$, the per-segment loss is the negative Plackett–Luce log-likelihood, averaged over valid frames:

$$\mathcal{L}_{\text{ListMLE}} = \frac{1}{n} \sum_{r=1}^{n} \left[ \log \sum_{\ell=r}^{n} \exp(\hat{y}_{\pi_\ell}) - \hat{y}_{\pi_r} \right]. \tag{2}$$

The implementation computes this with `logcumsumexp` for numerical stability, then averages over segments in the batch.

The MSE + pairwise ranking baseline combines a framewise regression term with a margin ranking term:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{M}|} \sum_{(b,i) \in \mathcal{M}} (\hat{y}_{b,i} - y_{b,i})^2, \tag{3}$$

where $\mathcal{M}$ is the set of valid, non-padding frames in the batch. For each segment $b$, we form ordered teacher-preferred pairs

$$\mathcal{P}_b = \{(i, j) : y_{b,i} - y_{b,j} > 10^{-6}\}. \tag{4}$$

If more than $64$ such pairs exist, we sample $64$ pairs uniformly without replacement. The pairwise hinge loss uses margin $m = 0.05$:

$$\mathcal{L}_{\text{rank}} = \frac{1}{B'} \sum_{b:\mathcal{P}_b \neq \emptyset} \frac{1}{|\tilde{\mathcal{P}}_b|} \sum_{(i,j) \in \tilde{\mathcal{P}}_b} \max\left(0,\ m - (\hat{y}_{b,i} - \hat{y}_{b,j})\right), \tag{5}$$

where $\tilde{\mathcal{P}}_b$ is the sampled pair set and $B'$ is the number of segments with at least one valid pair. The total baseline loss is

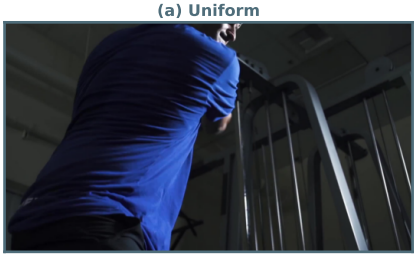
$$\mathcal{L}_{\text{MSE+pairwise}} = 0.3\, \mathcal{L}_{\text{MSE}} + 0.7\, \mathcal{L}_{\text{rank}}. \tag{6}$$

For downstream MSR-VTT captioning, both trained scorers are evaluated with the same protocol: frames decoded at 2 fps, stratified argmax selection, and Qwen2.5-VL-3B-Instruct with the prompt "Describe this video in one sentence."

## 1.4 Additional Qualitative Examples


Figure 1: Additional qualitative comparisons between PEEK and the SigLIP2 Oracle teacher on ActivityNet Captions test segments. Each page shows two additional examples (identified by the filenames). The visualization contrasts the frame-selection outcome for PEEK (query-free selector distilled from the Oracle's ranking) versus the Oracle scoring single frames against the ground-truth caption. For each method, we show the selected frame and the resulting downstream caption produced by Qwen2.5-VL-3B from that selection, highlighting cases where PEEK's temporally-aware selection yields frames that better preserve context (or avoids locally high-scoring but misleading frames) compared to the per-frame Oracle.

(a) Uniform

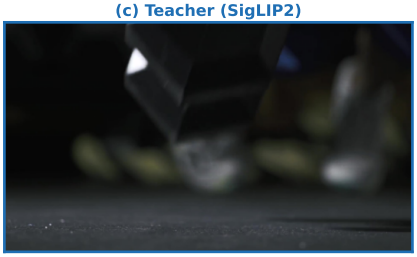


A man in a blue shirt is doing pull-ups on a bar.

(b) PEEK (ours)

A man is pulling on a rope while another person watches from behind him.

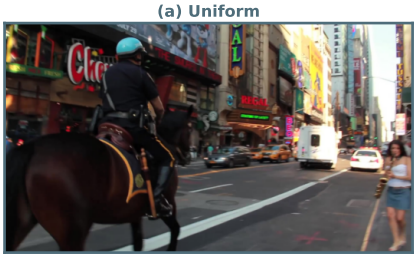
(c) Teacher (SigLIP2)



The video shows a close-up of a person's feet walking on a dark surface, with the focus on the movement and the texture of the shoes.

*GT: "A close up of a weight is shown followed by several clips of people working out intensely."*


(a) Uniform

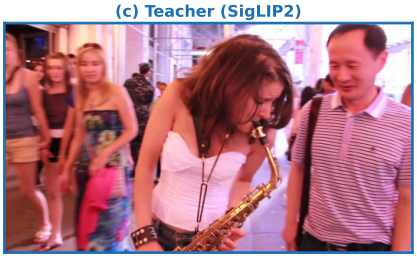


A police officer on horseback patrols the streets of New York City, passing by various shops and pedestrians.

(b) PEEK (ours)

A woman is walking down the street playing a saxophone while other people walk by.

(c) Teacher (SigLIP2)



A woman plays the saxophone while people walk by.

*GT: "The lady plays the saxophone in NYC."*

Figure 1: (Figure 1 continued.)

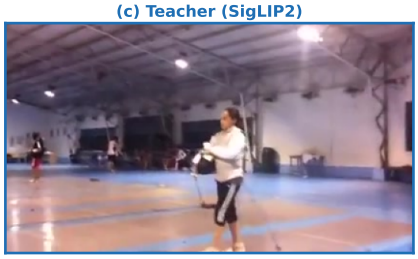
(a) Uniform

A woman is practicing boxing with gloves on.

(b) PEEK (ours)

A group of people are practicing fencing in an indoor arena.

(c) Teacher (SigLIP2)



A woman is standing in a gym holding a baton.

*GT: "A girl walks along the gym holding her fencing gear."*

(a) Uniform

A young boy is practicing gymnastics on a balance beam, with an adult watching him.

(b) PEEK (ours)

A young man is performing a gymnastics routine on parallel bars, while another man watches from the side.

(c) Teacher (SigLIP2)

A gymnast is performing on the parallel bars while a coach watches.

*GT: "A person is seen balance on a set of uneven bars while a man watches on the side."*

Figure 1: (Figure 1 continued.)

(a) Uniform

A group of people are standing on a grassy field, with some holding coolers and others standing around.

(b) PEEK (ours)

A group of shirtless men are playing drums on a grassy field.

(c) Teacher (SigLIP2)

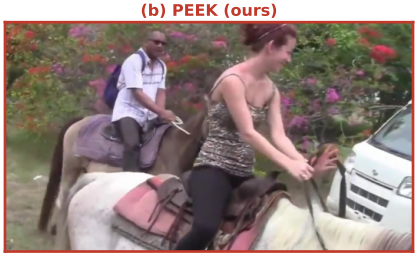
A group of drummers are practicing on the field at sunset.

*GT: "The drummers walk to form a line and then the other drummers came up to them and they continue to play drums."*

(a) Uniform

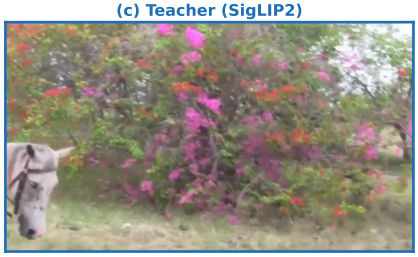
A man is riding a small horse on a dirt path surrounded by trees and grass.

(b) PEEK (ours)



A man and woman are riding horses in a park with colorful flowers in the background.

(c) Teacher (SigLIP2)



A cow is standing next to a bush of flowers.

*GT: "They continue riding through a grassy path with flowered bushes on the sides."*

Figure 1: (Figure 1 continued.)