Evo-Retriever: LLM-Guided Curriculum Evolution with Viewpoint-Pathway Collaboration for Multimodal Document Retrieval

Supplementary Material

6. Negative Query Generation Prompt

Negative Query Generation Prompts

You are given the following question:
{question}

The image can answer this question.

Now, write 20 new questions that are:
- Related to the topic,
- Seem reasonable,
- But cannot be answered using the image.

These questions should require knowledge that is not in the image.

Do not rephrase the original.

Give exactly 20 new questions. Just list them:
Variant 1: ...
Variant 2: ...
Variant 3: ...
Variant 4: ...
Variant 5: ...
Variant 6: ...
...

7. Rationale and Design of LLM-EC

As stated in Sec. 3.2 of the main paper, the LLM-EC framework relies on a set of M discrete difficulty intervals, defined over a positive-aware difficulty measure. The design of these intervals is critical. A naive uniform partitioning would ignore the highly non-uniform learning signals inherent in contrastive learning. We therefore adopt a **principled, non-uniform partitioning strategy**. This section details the theoretical and empirical rationale for this design, beginning with a gradient-sensitivity analysis that reveals the underlying problem structure.

7.1. Gradient-Sensitivity Analysis of the Contrastive Loss

The informativeness of a negative sample is closely related to the gradient it induces during training. To formalize this, we analyze the gradient of our softplus-based margin loss defined in Eq. (3) of the main paper. As the total loss is a sum over individual negative samples, we can analyze the gradient contribution from a single negative, $I_{\text{neg}}^{(j)}$. The gradient of the loss with respect to its similarity score, $s_{\text{neg}}^{(j)}$, is:

$$\frac{\partial \mathcal{L}}{\partial s_{\text{neg}}^{(j)}} = \frac{1}{\tau}\sigma\left(\frac{\Delta s^{(j)}}{\tau}\right) \tag{6}$$

where $\Delta s^{(j)} = s_{\text{neg}}^{(j)} - s_{\text{pos}}$ is the similarity gap, τ is the temperature, and $\sigma(\cdot)$ is the Sigmoid function. This equation reveals that the gradient magnitude is proportional to a sigmoid function of the similarity gap, scaled by $1/\tau$.

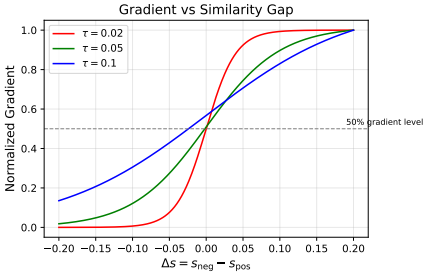

Figure 5. **Impact of Temperature τ on the Normalized Gradient Profile.** The plot shows the normalized gradient, $\sigma(\Delta s/\tau)$, as a function of the similarity gap Δs. A lower τ (e.g., our setting of 0.02, red line) creates a much steeper and narrower transition zone where gradients become substantial. This highlights the need for a curriculum that can precisely target negatives within this narrow "sweet spot" to ensure efficient learning.

The relationship between the gradient and the similarity gap is visualized in Fig. 5. The temperature τ critically governs the sharpness of this gradient profile. Our use of a low temperature ($\tau = 0.02$, red line) results in a highly localized gradient response. The gradient is negligible for large negative Δs (i.e., for easy negatives) but rises steeply as Δs approaches zero. This indicates that the most informative negatives are concentrated in a narrow window around the decision boundary, where the loss provides strong yet still meaningful training signals. A curriculum must therefore be able to precisely identify and sample from this "sweet spot" to avoid wasting computation on uninformative samples and ensure efficient training. This principle motivates our design of a non-uniform, high-resolution action space for the LLM-EC.

7.2. Rationale for Non-Uniform Partitioning

Although the analysis in Sec. 7.1 is expressed in terms of the similarity gap Δs, it provides the theoretical motivation for our practical interval design based on a positive-aware difficulty measure, which captures the relative proximity of a negative to its paired positive. The design of the LLM-EC action space involves two core steps: defining the operating range and partitioning it.

Operating Range. Our ablation study in Tab. 3 of the main paper directly informs the choice of range. Introducing a lower bound (e.g., "Fixed Window 80–95 %", ID j) improves performance to **61.20 %** from **60.99 %** (+0.21 %) compared to including easy negatives ("Fixed Top-K 95%", ID h). Conversely, an excessively high upper bound ("Top-K 99.9 %", ID i) is detrimental, causing performance to drop to **60.18 %** (-0.81 %). To ensure a diverse action space for dynamic adjustment while avoiding these extremes, and guided by the gradient profile (Sec. 7.1), we establish the primary operating range of the positive-aware difficulty measure as $[0.70, 0.995]$.

Within this range, the gradient response (Fig. 5) is highly non-uniform. The gradient $\sigma(\Delta s/\tau)$ is both **negligible and insensitive** to the similarity gap for $\Delta s < -0.10$, but rises sharply as $s_{\text{neg}} \to s_{\text{pos}}$. A uniform partitioning would waste resolution on this low-signal zone while undersampling the narrow, high-information "sweet spot".

Therefore, we define three zones based on the gradient curve: the **Low-Signal Zone** ($[0.70, 0.85]$), the **Effective-Learning Zone** ($[0.85, 0.98]$), and the **High-Risk Zone** ($[0.98, 0.995]$). We allocate the $M = 16$ intervals with a higher density in the Effective-Learning Zone (8 intervals) and a coarser density in the others (4 each). The final boundaries are determined via empirical quantiles of the data distribution (Tab. 5), ensuring both reproducibility and data-informed alignment.

Table 5. The action space of $M = 16$ overlapping sampling policies.

Action ID	Difficulty Range
Low-Signal Zone	
A	$[0.70, 0.85]$
B	$[0.70, 0.90]$
C	$[0.70, 0.92]$
D	$[0.75, 0.90]$
Effective-Learning Zone	
E	$[0.75, 0.92]$
F	$[0.75, 0.94]$
G	$[0.80, 0.92]$
H	$[0.80, 0.94]$
I	$[0.80, 0.95]$
J	$[0.85, 0.96]$
K	$[0.85, 0.97]$
L	$[0.85, 0.98]$
High-Risk Zone	
M	$[0.90, 0.985]$
N	$[0.92, 0.985]$
O	$[0.95, 0.99]$
P	$[0.95, 0.995]$

7.3. Formal Decision Protocol

The curriculum evolution in LLM-EC is governed by a deterministic, phase-dependent decision protocol. This protocol maps the training state to specific, reproducible difficulty-adjustment actions. The numerical parameters within the protocol are not arbitrary but are based on the following design principles and were parameterized via preliminary experiments.

Protocol Design and Parameterization.

- **Effective Learning Window ($[0.3, 1.2]$):** This is the empirically determined ideal range for the loss. For the softplus-based contrastive loss with a temperature of $\tau = 0.02$, a loss below 0.3 heuristically indicates that negatives are too easy, providing insufficient learning signal. A loss above 1.2 suggests a risk of training instability, where the model struggles to converge. This window filters for actions that are both challenging and stable.
- **Trend Estimation Window (20%):** To estimate learning velocity, the protocol compares the loss at the beginning ($\mathcal{L}_{\text{start}}$) and end ($\mathcal{L}_{\text{end}}$) of each review period. Both metrics are computed as averages over the first and last 20% of steps within that period, respectively. This window provides a principled balance—wide enough to suppress batch-level variance, yet narrow enough to remain sensitive to short-term trends.
- **Anchor Point Principle (Difficulty-Optimized Selection):** During the **Transition phase**, the controller applies a hierarchical selection strategy to identify the anchor that will seed the subsequent **Lock-in phase**. The goal is to choose the most challenging curriculum that still maintains stable learning dynamics, as observed during Exploration. If no such action exists, this indicates a potential curriculum calibration failure, rather than a case where an unsupported anchor should be forced. In this case, the system should re-examine the difficulty intervals, loss thresholds, or candidate pool, and may optionally perform an additional round of exploration before entering the Lock-in phase. This two-level mechanism forms the core of the Phase 2 procedure.

Key Metric Definitions.

- **Action (a):** An integer index from 0 to 15, corresponding

to a predefined difficulty interval (see Tab. 5). In logs, these may be represented by letters "A" through "P".

- **Hard Negative Loss ($\mathcal{L}_{\text{neg}}$):** The primary real-time performance metric for decision-making. It is the mean value of the **total contrastive loss,** $\mathcal{L}_{total}$, defined in Eq. (1) of the main paper, computed within a review window. This loss directly reflects the model's performance on the current curriculum's hard negatives. For historical records, this value is stored as `avg_loss`.
- **Historical Summary (H):** A record of all previously executed actions. Each entry contains at least the action taken and its resulting average loss, i.e., a tuple of (`action`, `avg_loss`), along with metadata such as the global step. This summary serves as the historical context for the overall decision-making process.

Phase 1: Exploration. Objective: To systematically test different difficulty intervals and map the difficulty-performance landscape. **Decision Logic:**

1. **High-Loss Anomaly:** If $\mathcal{L}_{\text{neg}} > 1.2$, difficulty is reduced by 2 intervals ($a_{\text{next}} \leftarrow \max(a_{\text{current}} - 2, 0)$).
2. **Low-Loss Anomaly:** If $\mathcal{L}_{\text{neg}} < 0.05$ for **two consecutive** reviews, difficulty is increased by 3 intervals ($a_{\text{next}} \leftarrow \min(a_{\text{current}} + 3, 15)$).
3. **Default Progression:** Otherwise, select the lowest-indexed action greater than the current one that has not been used in the last three reviews.

Phase 2: Transition. Objective: To select the most robust "anchor" action for the next phase based on the exploration history. **Decision Logic:**

1. **Filter for Effective Actions:** Construct a set of valid actions, A_{valid}, by filtering history H for actions whose `a.avg_loss` falls within the Effective Learning Window of $[0.3, 1.2]$.
2. **Select Optimal Anchor:** From the actions in A_{valid}, identify the one with the highest difficulty (i.e., the one with the largest Action ID, from "A" to "P"). This action is selected as the anchor to seed the Lock-in phase.
3. **Fallback:** If A_{valid} is empty, this is treated as a curriculum calibration failure rather than forcing an unsupported anchor choice. The system should re-examine the difficulty intervals, loss thresholds, or candidate pool, and optionally perform an additional exploration round before entering the Lock-in phase.

Phase 3: Lock-in. Objective: To conduct fine-grained difficulty adjustments around the anchor. **Decision Logic:** The protocol assesses "learning velocity" by comparing $\mathcal{L}_{\text{start}}$ and $\mathcal{L}_{\text{end}}$, calculated over the Trend Estimation Window.

1. **Upgrade Condition:** Increase difficulty by 1 step if either is met:
   - *Absolute Mastery:* $\mathcal{L}_{\text{end}} < 0.3$.
   - *Significant Progress:* A relative loss reduction of $\geq 50\%$.
2. **Downgrade Condition:** Decrease difficulty by 1 step if there is a relative loss increase of $\geq 30\%$.
3. **Maintain Condition:** If neither of the above is met, maintain the current action.

7.4. Real-World Deliberation Example

This section presents a real-world example of the model's end-to-end deliberation. In this case, the controller is in the **EXPLORATION** phase and must select the next difficulty interval based on the current interval, recent loss, and the last three actions. The trace below shows how it applies the exploration protocol and correctly selects action **C**.

Real-World Deliberation

<thinking>
Current phase: EXPLORATION

Objective: Systematically explore the difficulty space by selecting a higher difficulty action that has not been used in the last three reviews, unless an anomaly rule is triggered.

Step 1: Identify the current action and recent history.
- Current interval: [0.7, 0.90]
- Current action: B
- Last 3 actions:
Step 34: B
Step 32: D
Step 30: F

Therefore, the recent action set is {B, D, F}, and these actions should be avoided.

Step 2: Check anomaly conditions.
- hard_negative_loss_mean = 0.3983
- High-loss anomaly condition: $L_{neg} > 1.2$
Not triggered, since $0.3983 < 1.2$.
- Low-loss anomaly condition: $L_{neg} < 0.05$ for two consecutive reviews
Not triggered.

Therefore, follow the default exploration progression rule.

Step 3: Apply default progression.
The protocol specifies selecting the lowest-indexed action greater than the current one that has not been used in the last three reviews.
- Current action = B

- Higher actions begin from C
- C is not in {B, D, F}

Thus, C is the first valid higher action.

Step 4: Final decision.
Select action C as the next exploration step.
</thinking>

<answer>C</answer>.

8. Additional Ablation Studies

8.1. Effect of the Temperature τ

We study the effect of the temperature τ in the margin loss defined in Eq. (3) of the main paper on ViDoRe V2 using Qwen2.5-VL-3B-Instruct under the same training setup as in our LLM-EC experiments reported in Sec. 4.5 of the main paper (all hyperparameters and data unchanged, only varying τ). As shown in Tab. 6, a large temperature ($\tau = 1.0$) noticeably hurts performance, yielding 57.37 % nDCG@5, whereas our default $\tau = 0.02$ achieves a better result (57.37 % $\rightarrow$ 63.05 %, +5.68 %), indicating that a sharper margin is important for separating hard negatives in the multi-vector space.

Table 6. Ablation on the temperature τ in the margin loss on ViDoRe V2 (3B model). All scores are nDCG@5 (%). **Bold** numbers denote the best score.

τ	nDCG@5
1.0	57.37
0.02 (Ours)	**63.05**

8.2. Effect of Interval Density: Non-uniform vs. Uniform Partitioning

The design rationale for our **non-uniform, overlapping** difficulty partitioning is detailed in Sec. 7.2. Briefly, intervals are denser in the high-information Effective-Learning Zone ($[0.85, 0.98]$) and coarser in the Low-Signal and High-Risk zones, with partial overlap to ensure smooth difficulty transitions.

Here, we isolate the effect of the density strategy by comparing our default design against a **uniform, overlapping** baseline. This baseline divides the full operating range $[0.700,\ 0.995]$ into 16 equal-width intervals, each with fixed overlap between adjacent bins; all other LLM-EC settings remain identical.

We evaluate both strategies on ViDoRe V2 using the Qwen2.5-VL-3B-Instruct model under the same setup as in our LLM-EC ablation experiments (Sec. 4.5 of the main paper; all hyperparameters and data unchanged, only varying the partitioning strategy). Results are shown in Tab. 7.

Table 7. Effect of interval density on ViDoRe V2 (3B model). All scores are nDCG@5 (%). **Bold** marks the best result.

Partitioning Strategy	nDCG@5
Uniform + Overlap	61.89
Non-uniform + Overlap (Ours)	**63.05**

Our non-uniform approach improves performance by +1.16 % absolute over the uniform-overlap baseline. This gain supports our design choice in Sec. 7.2: allocating higher resolution where hard negatives are concentrated enables the LLM meta-controller to make finer, more effective curriculum adjustments, while avoiding wasted granularity in low-signal regions.

8.3. Transition and Lock-in Phase Lengths

We vary the Transition and Lock-in phase lengths in the Three-Phase Decision Protocol while fixing the Exploration phase to 60 steps. As shown in Tab. 8, sweeping Transition/Lock-in lengths from 50 to 200 steps shows that short phases (50 / 50) slightly underperform, whereas longer phases (200 / 200) achieve the best nDCG@5 (62.18 % $\rightarrow$ 63.05 %, +0.87 %), indicating that moderately longer phases provide a more stable and informative curriculum signal.

Table 8. Ablation on Transition and Lock-in phase lengths for LLM-EC (3B, ViDoRe V2). All scores are nDCG@5 (%). **Bold** numbers denote the best score.

Transition steps	Lock-in steps	nDCG@5
50	50	62.18
200 (Ours)	**200 (Ours)**	**63.05**

8.4. Robustness and Transferability of LLM-EC

We further analyze the robustness and generalization capability of LLM-EC under different settings.

Backbone Transfer. To validate whether the curriculum strategy generalizes across backbones, we apply LLM-EC to **Qwen3-VL-4B-Instruct**. Consistent with Sec. 4.5, each curriculum is evaluated on the **Q$\rightarrow$D path**, with only the hard negative pool being rebuilt. We additionally include a **Linear Curriculum** baseline that linearly schedules the ac-

Table 9. Curriculum ablation on Qwen3-VL-4B-Instruct. All scores are nDCG@5 (%). **Bold** numbers denote the best score.

Configuration	nDCG@5
baseline (Net0)	64.03
Fixed Window 80–98%	64.32
Linear Curriculum	64.29
Rule-based Oracle	64.36
LLM-EC (Ours)	**65.20**

Table 10. Robustness to threshold perturbations on ViDoRe V2 (Evo-Retriever-3B). In the Lock-in phase, thresholds in the Upgrade and Downgrade Conditions are perturbed: $L_{\text{end}} < 0.3 \rightarrow 0.36$, relative loss reduction for upgrade $50\% \rightarrow 60\%$, and relative loss increase for downgrade $30\% \rightarrow 25.5\%$. Drop = Default − Perturbed. All scores are nDCG@5 (%).

Scheduler	Default	Perturbed	Drop
Rule-based Oracle	62.81	61.75	−1.06
LLM-EC (Ours)	**63.05**	**62.40**	**−0.65**

tion ID from the easiest (Action 'A') to the most challenging (Action 'P') over training. As shown in Tab. 9, LLM-EC achieves the best performance, improving the baseline by +1.17 % (64.03 % → 65.20 %). It also surpasses the strongest hard-coded baseline, **Rule-based Oracle** (64.36 %), by +0.84 %. These results indicate that, under the same curriculum protocol, LLM-EC can successfully adapt the scheduling decisions to a new backbone without additional tuning, whereas rule-based schedulers fail to achieve comparable improvements.

Robustness to Threshold Perturbations. A potential concern is that LLM-EC may merely execute the threshold heuristics defined in the protocol. To test this, we perturb the thresholds used in the Upgrade and Downgrade Conditions while keeping all other training settings unchanged. Specifically, we relax the Absolute Mastery threshold from $L_{\text{end}} < 0.3$ to $L_{\text{end}} < 0.36$, increase the required relative loss reduction in the Upgrade Condition from 50% to 60%, and decrease the relative loss increase threshold in the Downgrade Condition from 30% to 25.5%. As shown in Tab. 10, the Rule-based Oracle suffers a drop of -1.06% (62.81% → 61.75%), whereas LLM-EC maintains more stable performance with only a -0.65% drop (63.05% → 62.40%). The rule-based scheduler is sensitive to threshold shifts because its decisions are entirely threshold-driven, whereas LLM-EC adapts curriculum decisions in response to training dynamics. This indicates that the controller does not rely solely on fixed heuristic triggers.

8.5. Curriculum Trajectory Analysis

To better understand the behavior of LLM-EC, we visualize the curriculum difficulty trajectories during training under the same setting as the LLM-EC ablation experiments in Sec. 4.5 (Qwen2.5-VL-3B-Instruct backbone with a Qwen3-32B controller). Compared with the rule-based oracle, which follows monotonic threshold-triggered difficulty jumps, LLM-EC adapts its decisions based on training dynamics and may roll back to easier intervals when instability is detected. This trend-aware behavior helps maintain training within an effective learning regime.

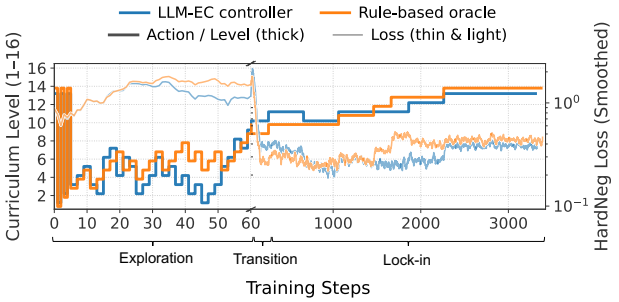

Figure 6. Curriculum trajectories during training for Evo-Retriever-3B. Thick lines indicate the selected difficulty interval, while thin lines show the hard-negative loss. The rule-based oracle follows monotonic threshold-triggered jumps, whereas LLM-EC adapts difficulty according to training dynamics and performs rollback when instability is detected.