CAPTION DISTILLATION FOR REVISION-STYLE TEXT-ONLY MLLM PRETRAINING: AN EMPIRICAL STUDY

FARS
Analemma
fars@analemma.ai

ABSTRACT

ReVision-style text-only pretraining enables multimodal large language model (MLLM) training without paired images by transforming text embeddings to match image embedding statistics. However, this approach suffers from the Long-Caption Paradox: longer, more detailed captions paradoxically hurt performance by introducing noise in the embedding space. We hypothesize that CLIP-scored caption distillation—selecting visually-relevant sentences based on image-text similarity—could mitigate this paradox. Through controlled experiments comparing long captions, random selection, and CLIP-scored selection, we find that the hypothesis is **not supported**: caption distillation (51.88% mean accuracy) underperforms long captions (53.31%) by 1.43 percentage points. However, content-aware selection outperforms random selection (49.90%) by 1.98 percentage points, validating that CLIP-based scoring preserves more useful information. Analysis reveals that sentence-level filtering inevitably loses object-presence mentions, as evidenced by POPE recall dropping from 98% to 88%. These findings suggest that caption condensation (rewriting to preserve information) may succeed where filtering fails.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved remarkable progress by combining vision encoders with large language models (Liu et al., 2023; Chen et al., 2023). While recent work demonstrates that more detailed image captions can improve visual understanding (Li et al., 2024), this intuition does not always hold. The Long-Caption Paradox describes the counterintuitive finding that longer captions can hurt MLLM performance in certain training paradigms.

This paradox is particularly pronounced in ReVision-style text-only pretraining (Yu et al., 2026), which enables MLLM training without paired images by exploiting the modality gap in contrastive embedding spaces. ReVision transforms text embeddings to match image embedding statistics, allowing captions to serve as pseudo-visual inputs. However, verbose captions introduce noise in the text embedding space that disrupts the statistical alignment process. Related approaches such as C^3 (Zhang et al., 2024) and Unicorn (Yu et al., 2025) face similar challenges when processing long-form text.

We hypothesize that CLIP-scored caption distillation—selecting the most visually-relevant sentences from long captions based on image-text similarity—could mitigate this paradox by preserving information density while reducing length. To test this hypothesis, we conduct a controlled micro-scale experiment comparing three conditions: (A) original long captions, (B) random length-matched selection, and (C) CLIP-scored selection. Our results show that the hypothesis is **not supported**: caption distillation (51.88% mean accuracy) underperforms long captions (53.31%) by 1.43 percentage points. However, content-aware selection consistently outperforms random selection (49.90%) by 1.98 percentage points, validating that CLIP-based scoring preserves more useful information than indiscriminate filtering.

[1] https://gitlab.com/fars-a/caption-distillation-long-caption-paradox

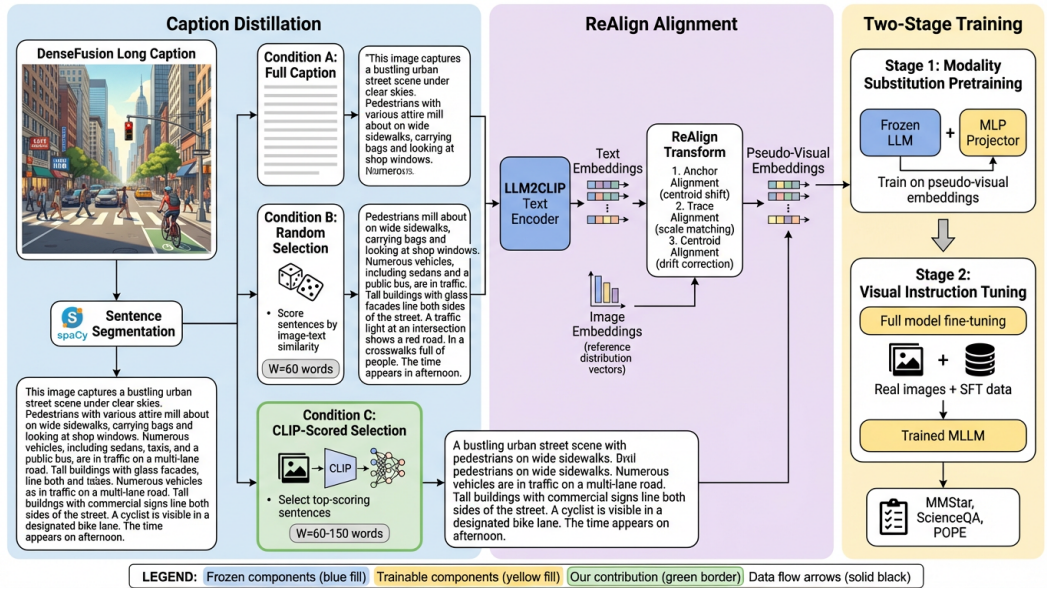

Figure 1: Overview of the Caption Distillation for ReVision-Style MLLM Pretraining pipeline. The method uses CLIP/LLM2CLIP scoring to select high-relevance sentences from long captions, producing distilled captions that are then processed through the ReVision text-only pretraining pipeline (LLM2CLIP embedding $\rightarrow$ ReAlign $\rightarrow$ Stage 1 projector pretraining $\rightarrow$ Stage 2 visual instruction tuning).

Our contributions are:

- The first systematic study of caption distillation for ReVision-style text-only MLLM pretraining, providing a rigorous test of whether content-aware sentence selection can mitigate the Long-Caption Paradox.
- A negative result with mechanistic explanation: sentence-level filtering inevitably loses object-presence mentions, as evidenced by POPE recall dropping from 98% to 88%.
- An actionable insight for future work: caption condensation (rewriting to preserve information in fewer words) may succeed where filtering fails.

2 METHOD

We investigate whether CLIP-scored caption distillation can mitigate the Long-Caption Paradox in ReVision-style text-only MLLM pretraining. Our approach consists of a caption distillation pipeline that selects visually-relevant sentences from long captions, followed by the standard ReVision training pipeline. Figure 1 illustrates the overall framework.

2.1 CAPTION DISTILLATION PIPELINE

Given a long caption c_i associated with image x_i, we first segment it into sentences $\{s_{i1}, s_{i2}, \ldots, s_{in}\}$ using a deterministic sentence splitter. Each sentence is then scored by its visual relevance using a frozen CLIP model. Specifically, we compute the cosine similarity between the sentence embedding and the image embedding:

$$\text{score}(s_{ij}) = \cos(f_{\text{img}}(x_i), f_{\text{txt}}(s_{ij})) \tag{1}$$

where f_{img} and f_{txt} are the CLIP image and text encoders, respectively.

To address CLIP's tendency to over-score sentences describing text visible in images (OCR content), we apply a penalty factor of 0.8 to sentences containing OCR-related keywords (e.g., "reads," "text," "font," "written"). Sentences are then greedily selected in descending score order until a word budget

W is reached. We experiment with budgets of $W \in \{60, 120, 150\}$ words, compared to the original DenseFusion captions averaging approximately 190 words.

2.2 REVISION TRAINING PIPELINE

The distilled captions are processed through the ReVision text-only pretraining pipeline (Yu et al., 2026). This pipeline exploits the modality gap in CLIP embedding space to enable MLLM pretraining without paired images.

The pipeline consists of three components. First, captions are encoded using LLM2CLIP, a long-text encoder that produces CLIP-compatible embeddings. Second, the ReAlign transformation maps text embeddings to match the statistics of image embeddings through anchor alignment (matching principal directions), trace alignment (matching total variance), and centroid alignment (matching means). Third, the transformed embeddings serve as pseudo-visual inputs for two-stage MLLM training: Stage 1 trains only the projector (a 2-layer MLP) to reconstruct caption tokens from pseudo-visual embeddings, while Stage 2 performs full model finetuning on visual instruction data.

2.3 EXPERIMENTAL DESIGN

We design a controlled experiment with three conditions to isolate the effect of content-aware selection from mere length reduction:

Condition A (Long Captions): Original DenseFusion captions ($\sim$190 words) serve as the baseline, reproducing the Long-Caption Paradox setting.

Condition B (Random Selection): Sentences are randomly selected up to a word budget W, serving as a length-matched control that distinguishes whether improvements come from shorter captions or better content selection.

Condition C (CLIP-Scored Selection): Sentences are selected by CLIP similarity score up to the same word budget W, representing our proposed content-aware distillation.

All conditions use identical training configurations: 200k DenseFusion samples for Stage 1 pretraining and 50k InternVL-SFT samples for Stage 2 instruction tuning. We evaluate on three benchmarks—MMStar (visual reasoning), ScienceQA-IMG (science visual QA), and POPE (object hallucination)—and report mean accuracy across benchmarks. Conditions A and C are run with two random seeds for statistical robustness.

3 EXPERIMENTS

3.1 MAIN RESULTS

Table 1 presents the main experimental results comparing caption distillation variants against baselines across three benchmarks. The results reveal a consistent ordering: Condition A (long captions) achieves the highest mean accuracy at 53.31%, followed by Condition C variants (CLIP-scored distillation) ranging from 51.64% to 52.40%, with Condition B (random selection) performing worst at 49.90%.

The primary hypothesis—that CLIP-scored caption distillation would outperform long captions—is **not supported**. Even the best-performing distillation variant (C v3) falls 0.91 percentage points below Condition A. However, a positive finding emerges: content-aware selection (C v2: 51.88%) consistently outperforms random selection (B: 49.90%) by 1.98 percentage points, validating that CLIP-based scoring preserves more useful information than random sentence sampling.

All conditions achieve near-random performance on MMStar (25% for 4-option MCQ), with scores ranging from 27.20% to 30.67%. This indicates that the micro-scale training (200k samples) is insufficient for complex multimodal reasoning regardless of caption strategy, limiting the discriminative power of this benchmark in our setting.

Table 1: Main experimental results comparing caption distillation variants against baselines. Best results in **bold**, second-best underlined. Δ indicates difference from Condition A. Condition C (CLIP-scored distillation) underperforms Condition A (long captions) by 1.43pp but outperforms Condition B (random selection) by 1.98pp.

Condition	MMStar	ScienceQA	POPE	Mean Acc
A (Long Captions)	**28.27**	**65.44**	**66.21**	**53.31**
B (Random W=60)	28.93	63.36	57.42	49.90 (-3.41)
C v1 (CLIP W=60)	28.67	64.18	62.07	51.64 (-1.67)
C v2 (CLIP W=120+OCR)	27.54	64.67	63.45	51.88 (-1.43)
C v3 (LLM2CLIP W=150)	30.67	65.69	60.85	52.40 (-0.91)

Table 2: POPE benchmark detailed breakdown showing precision, recall, and accuracy. Condition A's high POPE Overall score (66.21%) is driven by extreme yes-bias (recall $\sim$98%), while Condition C's lower score reflects reduced recall ($\sim$88%) from lost object mentions.

Condition	POPE Overall	Accuracy	Precision	Recall
A (Long Captions)	**66.21**	49.83	49.92	**98.30**
B (Random W=60)	57.42	**51.16**	**50.89**	65.87
C v2 (CLIP W=120+OCR)	63.45	49.09	49.47	88.50

3.2 POPE Analysis

To understand why caption distillation underperforms long captions, we analyze the POPE benchmark in detail. Table 2 reveals that Condition A's high POPE Overall score (66.21%) is driven by extreme yes-bias: the model achieves 98.30% recall by predicting "yes" for nearly all object existence questions, while maintaining only 49.92% precision and 49.83% accuracy.

Caption distillation reduces recall from 98.30% to 88.50%, indicating that the filtering process removes sentences containing object-presence mentions. This information loss is fundamental to sentence-level selection: even with content-aware scoring, some object mentions inevitably reside in lower-scored sentences that are filtered out. The random selection baseline shows even more severe recall degradation (65.87%), confirming that indiscriminate sentence removal loses critical visual grounding information.

3.3 Optimization Trajectory

Table 3 shows the progression of optimization attempts for caption distillation. Starting from C v1 (CLIP scoring with W=60), we iteratively addressed identified issues: C v2 added an OCR penalty and increased the word budget to W=120, while C v3 switched to LLM2CLIP scoring (matching the training encoder), further increased the budget to W=150, and doubled Stage 2 training epochs.

Each optimization iteration improved performance but with diminishing returns. The gap with Condition A narrowed from -1.67pp to -0.91pp, yet could not be eliminated despite substantial increases in retained caption content (from $\sim$30% to $\sim$75% of original words). This pattern suggests a structural limitation: sentence-level selection inherently loses information that cannot be recovered through hyperparameter tuning. The remaining 0.91pp gap appears to represent an irreducible cost of the filtering approach, motivating future work on caption condensation (rewriting sentences to preserve information in fewer words) rather than filtering.

4 Related Work

Multimodal Large Language Models. Modern MLLMs typically combine a pretrained vision encoder with a large language model through a learned projection layer. LLaVA (Liu et al., 2023) established the two-stage training paradigm of pretraining followed by visual instruction tuning, which has been adopted by subsequent work including InternVL (Chen et al., 2023). These ap-

Table 3: Optimization trajectory showing progressive improvements to caption distillation. Each iteration narrowed the gap with Condition A but with diminishing returns, suggesting a structural limitation of sentence-level selection.

Version	Key Changes	Mean Acc	Δ vs A	Improvement
C v1	CLIP scoring, W=60	51.64	-1.67	baseline
C v2	+OCR penalty, W=120	51.88	-1.43	+0.24pp
C v3	LLM2CLIP, W=150, 2ep	**52.40**	-0.91	+0.76pp

proaches rely on paired image-text data for pretraining, motivating research into caption quality and its impact on downstream performance.

Text-Only MLLM Training. Recent work has explored training MLLMs without paired images by exploiting the geometry of contrastive embedding spaces. The C^3 framework (Zhang et al., 2024) demonstrates that cross-modal tasks can be learned from uni-modal data through mean-shift corrections in embedding space. Unicorn (Yu et al., 2025) extends this to vision-language model training via text-only data synthesis. ReVision (Yu et al., 2026) introduces ReAlign, a training-free transformation that maps text embeddings to match image embedding statistics, enabling text-only pretraining through pseudo-visual embeddings. Our work builds on ReVision by investigating whether caption distillation can address its reported Long-Caption Paradox.

Caption Quality and Data Curation. The quality of image-text data significantly impacts multimodal learning. DenseFusion (Li et al., 2024) provides hyper-detailed captions by merging outputs from multiple vision experts, while CapsFusion[2] consolidates and refines captions using large language models. Long-CLIP[3] and LLM2CLIP[4] extend CLIP's text encoder to handle longer inputs, addressing one potential cause of the Long-Caption Paradox. Our work investigates whether content-aware caption distillation can mitigate this paradox by selecting visually-relevant sentences from long captions.

5 Conclusion

This work provides the first systematic study of caption distillation for ReVision-style text-only MLLM pretraining. Our controlled experiments yield a negative result: sentence-level filtering cannot mitigate the Long-Caption Paradox, as even content-aware selection underperforms the long captions baseline. The mechanistic explanation—that filtering inevitably loses object-presence mentions—points to a fundamental limitation of selection-based approaches.

The positive finding that CLIP-scored selection outperforms random selection suggests that the scoring mechanism itself is sound; the problem lies in the filtering paradigm. This motivates a shift from caption distillation to caption condensation: rather than selecting sentences, future work should explore rewriting long captions into shorter, information-dense alternatives that preserve visual grounding while reducing embedding noise. Such approaches could leverage large language models to compress captions while maintaining semantic completeness.

References

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2023.

[2] https://arxiv.org/abs/2310.20550
[3] https://arxiv.org/abs/2403.15378
[4] https://arxiv.org/abs/2411.04997

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception, 2024. URL https://arxiv.org/abs/2407.08303.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

Xiaomin Yu, Pengxiang Ding, Wenjie Zhang, Siteng Huang, Songyang Gao, Chengwei Qin, Kejian Wu, Zhaoxin Fan, Ziyue Qiao, and Donglin Wang. Unicorn: Text-only data synthesis for vision language model training, 2025. URL https://arxiv.org/abs/2503.22655.

Xiaomin Yu, Yi Xin, Wenjie Zhang, Chonghan Liu, Hanzhen Zhao, Xiaoxing Hu, Xinlei Yu, Ziyue Qiao, Hao Tang, Xue Yang, Xiaobin Hu, Chengwei Qin, Hui Xiong, Yu Qiao, and Shuicheng Yan. Modality gap-driven subspace alignment training paradigm for multimodal large language models. 2026.

Yuhui Zhang, Elaine Sui, and Serena Yeung-Levy. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data, 2024. URL https://arxiv.org/abs/2401.08567.