JoTS: Joint Token- and Segment-Level Prediction for Self-Supervised Multimodal Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Multimodal dense retrieval learns representations for semantic matching across textual and visual content. However, existing multimodal retrievers largely rely on manually annotated or synthetically generated relevance pairs, overlooking the supervision naturally present in naturally occurring multimodal documents, where text and images form coherent sequences. This natural document order can be regarded as annotation-free supervision.

We introduce **JoTS** (**Jo**int **T**oken- and **S**egment-Level Prediction), a self-supervised framework that jointly optimizes retrieval-conditioned next-token prediction (NTP) and segment-continuation prediction (SCP). NTP uses Retriever-weighted context to guide textual token prediction, while SCP identifies the succeeding text or image segment from a dynamic candidate set. These parallel objectives capture fine-grained linguistic dependencies and higher-level multimodal continuity. We pretrain JoTS on OBELICS using vision–language backbones of different scales. On general-purpose MMEB and reasoning-intensive MM-BRIGHT, JoTS consistently improves its base backbones and outperforms representative supervised and contrastively trained retrievers, demonstrating the effectiveness of predictive self-supervision for multimodal retrieval.

1 Introduction

Multimodal dense retrieval enables semantic search across textual, visual, and interleaved image–text queries and documents, where multiple text spans and images occur within a shared semantic context. Recent multimodal Retrievers, often adapted from vision–language models, have expanded from conventional image–text matching to instruction-conditioned retrieval over heterogeneous inputs. Yet most are trained with curated positive pairs, task-specific relevance annotations, or synthetically generated examples and explicit negatives. Such supervision is costly and often needs to be reconstructed for each new domain or collection. This motivates learning retrieval representations directly from naturally occurring multimodal documents and their intrinsic structure.

Raw multimodal documents offer a largely untapped source of such supervision. HTML pages and visually rich PDF or DOCX files organize multiple text spans, images, and figures into ordered semantic contexts (Laurençon et al., 2023; Awadalla et al., 2024; Faysse et al., 2025). Their token order exposes linguistic dependencies, while their text–image succession exposes higher-level multimodal continuity. Revela shows that next-token prediction (NTP) can supervise text Retriever learning without conventional contrastive relevance labels (Cai et al., 2025). However, NTP does not transfer symmetrically across modalities: visual inputs are continuous features inserted into the language-model input rather than targets of its textual vocabulary head (Bai et al., 2025; Liquid AI, 2025). Images may condition subsequent text prediction, but they are not themselves direct NTP targets. We therefore ask: *can a multimodal Retriever be trained directly from raw interleaved documents by jointly predicting textual tokens and higher-level multimodal continuation?*

Inspired by JEPA-style prediction of semantic units in representation space (Assran et al., 2023), we introduce JoTS, a self-supervised framework that combines retrieval-conditioned NTP with segment-continuation prediction (SCP). A segment is one image or one contiguous text span retained in document order. NTP predicts observed text using Retriever-weighted cross-chunk context,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

whereas SCP uses the Reference representation at a segment boundary to identify the true succeeding text or image segment. We construct training groups from OBELICS by combining semantically near, intermediate, and distant document bundles while preserving their contents, interleaving, order, modality composition, and computational load (Laurençon et al., 2023). Within each group, the Retriever encodes complete chunks under query and passage prompts, and their similarities weight the Reference model's in-batch attention, allowing both losses to train the Retriever. The same Retriever separately encodes individual segments to form the dynamic SCP candidate set, providing an additional gradient path through candidate representations. At inference, only the trained Retriever is retained.

We evaluate JoTS on MMEB, which covers 36 datasets across four multimodal embedding task families (Jiang et al., 2024b), and MM-BRIGHT, a reasoning-intensive benchmark spanning multimodal-to-text, multimodal-to-image, and multimodal-to-multimodal retrieval (Abdallah et al., 2026). With LFM2.5-VL-450M, JoTS achieves **32.558** macro P@1 on MMEB, compared with 2.610 for the unadapted backbone and 29.186 for same-data SimCSE-style training, and obtains 12.739, 27.304, and 17.998 nDCG@10 on the three MM-BRIGHT tasks. With Qwen2.5-VL-3B-Instruct, it reaches 34.089 P@1 and 15.711, **40.120**, and 20.553 nDCG@10, demonstrating transfer across model scales and architectures. Ablations further examine the two predictive objectives, SCP weighting, entropy regularization, and visual evidence. Together, these results establish intrinsic document continuity as effective supervision for multimodal Retriever pretraining.

Our contributions are threefold:

- We introduce constraint-preserving semantic difficulty grouping, which reshapes group-local contexts and candidates while preserving document content, multimodal structure, and computational load.
- We jointly optimize textual NTP and modality-agnostic SCP to extend language-model-based Retriever learning to cross-chunk text and image evidence.
- We demonstrate consistent improvements across different vision–language backbones on general-purpose and reasoning-intensive multimodal retrieval evaluations.

2 Related Work

2.1 Self-Supervised Dense Retrieval

Self-supervised text representations are learned through contrastive or reconstruction objectives (Gao et al., 2021; Izacard et al., 2022; Gao & Callan, 2022; Xiao et al., 2022). Language-model predictions also supervise retriever learning (Guu et al., 2020; Shi et al., 2023); Revela and DREAM more directly connect retriever similarities to autoregressive prediction (Cai et al., 2025; Tang & Yang, 2026). We extend predictive retriever pretraining to interleaved image–text documents, jointly exploiting token-level supervision and multimodal segment continuity.

2.2 Universal Multimodal Retrieval

Beyond paired image–text representation learning (Radford et al., 2021), universal multimodal retrievers support diverse modality configurations and tasks (Wei et al., 2023; Jiang et al., 2024b; Lin et al., 2025; Zhang et al., 2024). Related approaches use text-pair supervision (Jiang et al., 2024a) or unlabeled image–text pairs (Yang et al., 2026), while interleaved-document methods emphasize contextual encoding and representation aggregation (Lee et al., 2026). Our focus is instead on the pretraining signal: deriving retriever supervision from observed text–image segment order rather than preconstructed retrieval pairs or changes to document aggregation.

2.3 Predictive Representation Learning

Quick Thoughts and CPC identify context or future candidates (Logeswaran & Lee, 2018; van den Oord et al., 2019), while I-JEPA predicts masked visual representations (Assran et al., 2023). Multimodal approaches combine contrastive learning with captioning (Yu et al., 2022) or autoregressively predict text tokens and visual embeddings (Sun et al., 2024). We jointly optimize next-token prediction and identification of the succeeding text or image segment among group-local candidates

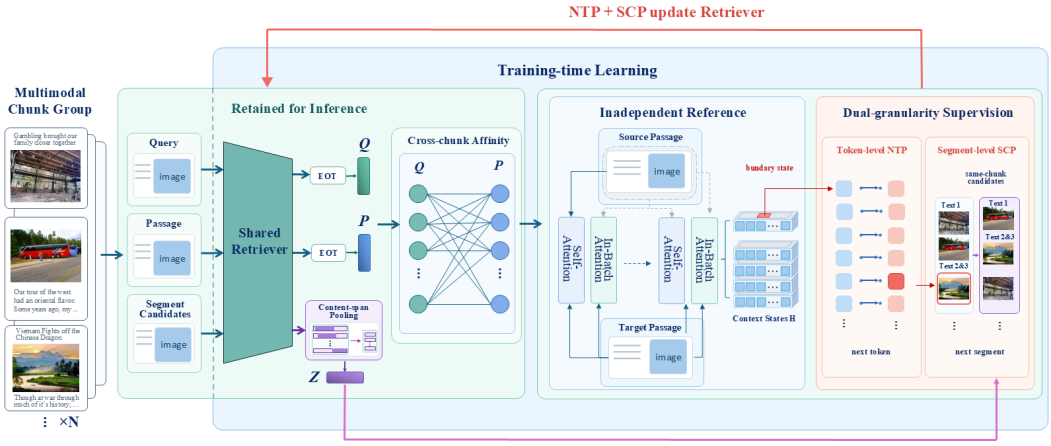


Figure 1: **The frame work of JoTS.** The shared Retriever produces query, passage, and candidate representations. Cross-chunk affinity guides the independent Reference model, whose context states support token-level NTP and same-chunk segment-level SCP. Both objectives update the Retriever, which is retained at inference. Blue/teal arrows denote forward computation, magenta denotes candidate embeddings (Z), and red denotes joint supervision.

encoded by the shared Retriever, rather than regressing visual embeddings or relying on a separate target encoder.

2.4 Interleaved Data and Training-Group Construction

Interleaved corpora preserve document-level image–text context (Zhu et al., 2023; Laurençon et al., 2023; Awadalla et al., 2024). Training composition is further studied through joint example selection (Evans et al., 2024), and hard-negative mining (Xiong et al., 2021; Qu et al., 2021). Our construction instead reassigns complete document bundles while preserving their multimodal structure and computational load. We control the mixture of semantically near, intermediate, and distant documents relative to an anchor. These roles guide grouping only, not model inputs or supervision labels; prediction targets remain determined by within-document order.

3 Method

We learn a multimodal dense retriever from unlabeled interleaved documents. The Retriever routes cross-chunk context to a Reference model, whose token- and segment-level predictions provide the training signal.

3.1 Framework and Training Groups

A training group is $\mathcal{G} = \{D_1, \ldots, D_B\}$ with $B = 16$, where each chunk preserves the original order of its text and image segments. We use independent Retriever and Reference copies, R_ϕ and F_θ, of the same vision–language backbone. For chunk-level encoding, we apply separate query and passage prompts, denoted by π_q and π_p, and use the final-layer hidden state at the end-of-text (EOT) token as the sequence representation. Let $\mathrm{Pool}_{\mathrm{EOT}}$ denote this pooling operation. The same Retriever is applied at two granularities:

$$q_i = \mathrm{Pool}_{\mathrm{EOT}}\left(R_\phi(\pi_q(D_i))\right), p_i \quad = \mathrm{Pool}_{\mathrm{EOT}}\left(R_\phi(\pi_p(D_i))\right), c_m = \mathrm{Pool}_{\mathrm{content}}\left(R_\phi(\pi_s(s_m))\right). \tag{1}$$

Here, s_m is exactly one image or one text span produced during chunking. The whole-chunk embeddings (q_i, p_i) produce the cross-chunk weight matrix used by both predictive objectives, whereas the independently encoded segment embeddings c_m form the dynamic candidate vocabulary used only by SCP.

To vary semantic difficulty without changing modality load, we regroup indivisible document bundles only within identical resource signatures:

$$\rho(b) = (n_{\text{chunk}}, n_{\text{image}}, n_{\text{text-only}}, n_{\text{multi-image}}, \mathbf{1}_{\text{single-image}}), \qquad \rho(b'_{g,r}) = \rho(b_{g,r}). \tag{2}$$

Within matched slots, the anchor-to-bundle similarity ranges follow

$$\frac{\#\{r > 1 : \kappa_{g,r} = c\}}{n_g - 1} \approx \pi_c, \qquad (\pi_{\text{near}}, \pi_{\text{mid}}, \pi_{\text{far}}) = (0.35, 0.35, 0.30). \tag{3}$$

Here, $\kappa_{g,r}$ is the anchor-to-bundle cosine-similarity range. Bundles are used without replacement, while their contents and within-document order remain unchanged. Details are in Appendix B.4.

3.2 Cross-Chunk Context Aggregation via Retriever-Weighted In-Batch Attention

Following Revela, we use in-batch attention (IBA) inside the Reference model to incorporate information from the other chunks in the same training group. The Retriever does not perform IBA itself; instead, it produces chunk-level weights that determine the contribution of each source chunk to the Reference context.

For whole-chunk encoding, query and passage prompts produce normalized EOT representations:

$$q_i = \text{norm}(R_\phi(D_i; \pi_q)), \qquad p_i = \text{norm}(R_\phi(D_i; \pi_p)). \tag{4}$$

Their similarities define the Retriever-produced cross-chunk weights:

$$A_{ij} = \frac{\mathbf{1}[i \neq j] \exp(q_i^\top p_j / \tau_{\text{ret}})}{\sum_{k \neq i} \exp(q_i^\top p_k / \tau_{\text{ret}})}, \qquad \tau_{\text{ret}} = 0.01. \tag{5}$$

The Reference model first constructs detached key–value caches for the source chunks. During its second causal forward pass, IBA uses A_{ij} to aggregate the source-specific attention contexts:

$$\text{IBA}_{i,t}^{(\ell)}(A) = \sum_{j \neq i} A_{ij} \ \text{Attn}\left(Q_{i,t}^{(\ell)}, K_j^{(\ell)}, V_j^{(\ell)}\right). \tag{6}$$

The resulting cross-chunk context is added to the within-chunk self-attention output:

$$Y_{i,t}^{(\ell)} = \text{SelfAttn}_{i,t}^{(\ell)} + \alpha_{\text{IBA}} W_O^{(\ell)} \, \text{IBA}_{i,t}^{(\ell)}(A), \qquad \alpha_{\text{IBA}} = 0.1. \tag{7}$$

Thus, IBA denotes the complete Retriever-weighted cross-chunk attention operation inside the Reference model, while cross-chunk context aggregation describes its function. Predictive losses update the Retriever through $R_\phi \to A \to \text{IBA}(A)$.

3.3 Token- and Segment-Level Predictive Learning

NTP supervises fine-grained textual continuation. Let $\mathcal{T}$ contain only positions whose next token is genuine document text, excluding prompts, templates, visual features, special tokens, and padding. We optimize

$$\mathcal{L}_{\text{NTP}} = -\frac{1}{|\mathcal{T}|} \sum_{(i,t) \in \mathcal{T}} \log p_{\theta,\phi}\left(x_{i,t+1} \mid x_{i,\leq t}, C_i(A)\right). \tag{8}$$

Images influence NTP through causal attention and IBA but are not language- vocabulary targets. We therefore add segment-continuation prediction (SCP) to directly cover both modalities in representation space.

For every within-chunk transition $s_{i,k} \to s_{i,k+1}$, the final Reference state at the current segment boundary is h_e, and the immediate next segment is target y_e. All group-local text and image

segments form the dynamic candidate set $\mathcal{V}_{\text{seg}} = \{s_m\}_{m=1}^{M}$. Using real-content mean pooling $c_m = \text{Pool}(R_\phi(s_m))$, we define

$$g_{e,m} = \frac{\text{norm}(W_{\text{context}} h_e)^\top \text{norm}(W_{\text{key}} c_m)}{\tau_{\text{SCP}}}, \quad \tau_{\text{SCP}} = 0.07,$$
$$\mathcal{L}_{\text{SCP}} = -\frac{1}{E} \sum_{e=1}^{E} \log \frac{\exp(g_{e,y_e})}{\sum_{m \in \mathcal{M}_e} \exp(g_{e,m})}. \tag{9}$$

The current segment is masked; no EOS or cross-chunk target is introduced. SCP updates the Retriever both through $A \to \text{IBA} \to h_e$ and directly through the candidate encoder $R_\phi(s_m)$.

Finally, with normalized routing entropy $\bar{H}(A) = [B \log(B-1)]^{-1} \sum_i [-\sum_{j \neq i} A_{ij} \log A_{ij}]$, the joint objective is

$$\mathcal{L}(t) = \mathcal{L}_{\text{NTP}} + \lambda_{\text{SCP}}(t) \mathcal{L}_{\text{SCP}} + 0.02 \, [\max(0, 0.35 - \bar{H}(A))]^2,$$
$$\lambda_{\text{SCP}}(t) = 0.03 \min(1, t/50). \tag{10}$$

At inference, the Reference model and segment projections are removed, and only R_ϕ is used for cosine-similarity retrieval. Further details are provided in Appendix A.

4 Experiments

4.1 Experimental Setups

Evaluation Benchmarks. We evaluate JoTS on two complementary multimodal benchmarks,MMEB and MM-BRIGHT. MMEB (Jiang et al., 2024b) contains 36 datasets spanning four meta-task categories: classification, visual question answering, multimodal retrieval, and visual grounding. MMEB primarily measures the general-purpose embedding capability of a model over heterogeneous combinations of image and text inputs. we evaluate reasoning-intensive retrieval on MM-BRIGHT (Abdallah et al., 2026), which covers 29 technical domains and more than 2.5 million candidate documents. Tasks 3 and 4 constitute our primary reasoning-intensive evaluation, while Task 2 provides a complementary assessment of whether a multimodal query representation retains information useful for text-document retrieval.We report nDCG@10 for each domain and macro-average the scores over all 29 domains of these 3 tasks. For every comparison, we use the same candidate construction, relevance judgments, and metric implementation within each task.

Training Data. We train on a curated subset of the naturally interleaved image–text documents from OBELICS (Laurençon et al., 2023). Our preprocessing preserves the original order of textual and visual elements within each document. After filtering and resource-aware chunking, the resulting training corpus contains 65,792 multimodal chunks from 26,049 documents, organized into 4,112 groups of 16 chunks each.

We construct the groups using semantic-difficulty grouping (SDG). Specifically, we use frozen document embeddings to arrange semantically near, intermediate, and distant documents within the same group, while preserving every selected chunk, its within-document order, and its image-load characteristics. The grouping procedure does not use retrieval annotations, benchmark queries, or relevance labels. Consequently, JoTS is trained exclusively from unlabeled interleaved documents rather than query–document pairs. Within a chunk, each complete text field or image is treated as one segment. Next-segment targets are defined only for adjacent segments within the same chunk.

Models. We instantiate both the retriever and the reference language model from the same model, like LFM2.5-VL-450M (Liquid AI, 2025) and Qwen2.5-VL (Bai et al., 2025). The retriever maps each multimodal chunk to query and passage representations and additionally encodes individual segments to form a group-specific candidate set. The reference model performs language modeling with retriever-weighted in-batch attention and supplies the contextual states used for next-segment prediction.

For chunk-level retrieval, we prepend role-specific query or passage instructions, preserve the original image–text order, and use the hidden state of the final native chat end token as the chunk rep-

resentation. No natural-language retrieval suffix is appended. At inference time, only the trained retriever adapter is retained; the reference model and the next-segment prediction heads are used only during training.

Baselines. We compare JoTS with the following groups of baselines. *Matched backbone controls* include the pretrained LFM2.5-VL-450M with the same query/passsage prompts, EOT readout, and evaluation pipeline but without retrieval training. This control separates improvements due to training from those due to the pretrained vision–language backbone and the embedding readout. *NTP-only baselines* retain the multimodal in-batch-attention architecture and next-token prediction objective but remove the next-segment loss. We include both our earlier multimodal NTP implementation and a controlled no- variant that uses the same Prompt-D input and pooling protocol as JoTS. *Objective ablations* remove next-token prediction or the retriever-entropy regularizer while holding the remaining training configuration fixed. We additionally compare alternative prompt and readout designs to separate the effect of the segment objective from that of representation extraction. The original text-only Revela model (Cai et al., 2025) is not used as a direct numerical baseline because it cannot consume the visual inputs required by MMEB and MM-BRIGHT.

For reference against existing supervised multimodal embedding systems, we include CLIP ViT-L/14 (Radford et al., 2021) and GME-Qwen2-VL-2B/7B (Zhang et al., 2024). These models are locally re-evaluated using the same benchmark-specific candidate and relevance protocols. Our LFM-based retriever consumes all available query images. For external models whose released interfaces support only one image, we use the first available query image and disclose this difference rather than treating the input protocols as identical.

Experimental Details. Each training group contains $B = 16$ chunks. We train with four-way distributed data parallelism on 4 A800 80GB GPUs, where every GPU processes one complete group per optimization step. Thus, a global step processes four independent groups, or 64 chunks, while retrieval similarities and in-batch attention are computed separately within each 16-chunk group. Both the retriever and reference model are adapted using LoRA with rank 16, scaling factor 32, and zero adapter dropout. The visual encoder and projector remain frozen.

We optimize all trainable parameters with AdamW using a peak learning rate of 1×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, and global gradient clipping at 1.0. The learning rate is linearly warmed up for 50 steps and then decayed over a 1,000-step horizon. Training uses `bfloat16` precision and seed 42 on four NVIDIA A800 GPUs. The text budget is 256 tokens per chunk, and each image is represented using between 96 and 256 visual tokens. During evaluation, query and candidate inputs use the same role prompts and native EOT readout as training. The reference model, segment projection layers, and training-time dynamic candidate set are not used during retrieval inference.

4.2 Main Results

Self-supervised JoTS is competitive with pair-supervised multimodal retrievers. Table 1 compares JoTS with unadapted backbones, self-supervised SimCSE-style adaptation, and external retrievers trained using paired or constructed query–candidate supervision. Despite using neither downstream relevance labels nor constructed query–document training pairs, JoTS surpasses CLIP ViT-L/14 on the MMEB average with both backbones and on MM-BRIGHT Task 2. More notably, JoTS with the 3B Qwen backbone reaches 40.120 on Task 3, outperforming CLIP by 8.165 points and the larger GME-Qwen2-VL-7B by 5.626 points. It also exceeds GME-Qwen2-VL-7B on Task 4 by 2.555 points. These comparisons show that joint token- and segment-level prediction can learn competitive multimodal retrieval representations without pair-level retrieval supervision.

JoTS substantially strengthens broad multimodal transfer on MMEB. MMEB evaluates transfer across classification, VQA, retrieval, and visual grounding rather than optimizing for a single retrieval format. With the compact LFM backbone, JoTS improves the MMEB-36 average from 29.186 to 32.558 over SimCSE-style adaptation. The largest gain appears in retrieval, where P@1 increases from 13.690 to 32.150, a gain of 18.460 points; visual grounding also improves by 2.700 points. Consequently, the 450M-parameter JoTS model exceeds CLIP ViT-L/14 by 3.325 points in the MMEB average and by 5.800 points in its retrieval category. With Qwen2.5-VL-3B, JoTS

Table 1: Retrieval results (%) on MMEB-36 (P@1) and MM-BRIGHT (nDCG@10). MMEB reports the overall macro average and four meta-task averages; MM-BRIGHT reports Tasks 2–4 averaged over 29 domains. Bold marks the better self-supervised adaptation within each backbone for each metric; external pair-supervised models are shown for reference and excluded from highlighting.

		MMEB-36 (P@1, %)					MM-BRIGHT (nDCG@10, %)		
Model	Objective	Avg.	Cls.	VQA	Retr.	VG	T2	T3	T4
LFM2.5-VL-450M backbone									
LFM2.5-VL-450M[†]	None	2.610	6.850	0.300	0.140	5.080	0.000	2.627	1.275
LFM SimCSE-style[‡]	InfoNCE	29.186	**38.310**	**28.400**	13.690	47.475	4.863	24.417	12.381
JoTS (Qwen)[‡]	NTP + SCP	**32.558**	37.360	21.200	**32.150**	**50.175**	**12.739**	**27.304**	**17.998**
Qwen2.5-VL-3B backbone									
Qwen2.5-VL-3B[†]	None	3.033	5.970	0.070	0.208	11.575	0.061	4.512	3.132
Qwen SimCSE-style[‡]	InfoNCE	**37.753**	**44.283**	**18.900**	**44.670**	**48.000**	8.615	22.959	10.977
JoTS (LFM)[‡]	NTP + SCP	34.089	41.380	13.040	41.950	44.900	**15.711**	**40.120**	**20.553**
External pair-supervised contrastive retrievers									
CLIP ViT-L/14[§]	Paired contrastive	29.233	40.392	9.160	26.350	53.150	10.422	31.955	28.969
GME-Qwen2-VL-2B[§]	Constructed-pair	52.919	54.460	67.617	31.650	58.150	19.345	42.446	22.945
GME-Qwen2-VL-7B[§]	Constructed-pair	57.597	58.250	72.467	36.880	63.150	22.500	34.494	17.998

[†] Unadapted backbone. [‡] Self-supervised adaptation on our multimodal corpus without downstream relevance labels. [§] Paired or constructed-pair contrastive supervision. T2 uses query images when available. Each row uses one fixed checkpoint; both JoTS rows use step 200.

Table 2: Retrieval performance on MMEB and MM-BRIGHT. MMEB reports macro-averaged Precision1 over 36 datasets and the corresponding averages for classification (Cls.), visual question answering (VQA), retrieval (Retr.), and visual grounding (VG). MM-BRIGHT reports macro-averaged nDCG10 over 29 domains for multimodal-to-text (T2), multimodal-to-image (T3), and multimodal-to-multimodal retrieval (T4). All values are percentages, and every row uses one fixed checkpoint across all reported columns. Within each backbone group, bold denotes the better result between the two self-supervised adaptation methods for each metric. External pair-supervised retrievers are reported for context and are not included in this highlighting.

		MMEB-36 (P@1, %)					MM-BRIGHT (nDCG@10, %)		
Variant	Step	Avg.	Cls.	VQA	Retr.	VG	T2	T3	T4
Full JoTS	200	**32.558**	37.360	21.200	32.150	50.175	12.739	27.304	17.998
w/o	400	29.167	29.180	15.980	34.175	47.075	15.629	28.702	19.420
w/o NTP	800	17.156	22.450	4.100	16.017	39.975	2.136	20.450	8.042
w/o entropy regularization	600	32.344	37.840	19.010	33.800	47.575	16.025	27.910	18.248

raises the unadapted model from 3.033 to 34.089, demonstrating that the proposed self-supervised objective transfers effectively across substantially different backbone scales and architectures.

JoTS exhibits its strongest advantage on reasoning-intensive cross-modal retrieval. MM-BRIGHT evaluates retrieval over heterogeneous domains while changing the target modality across Tasks 2–4. On LFM, JoTS improves over SimCSE-style adaptation by 7.876, 2.887, and 5.617 points on Tasks 2, 3, and 4, respectively. The advantage becomes larger with Qwen: the corresponding gains are 7.096, 17.161, and 9.576 points. In particular, the strong Task 3 result indicates that the learned representation is effective when a multimodal query must identify a relevant visual target, while the Task 4 result shows that this advantage extends to multimodal candidates. Together with the MMEB results, these findings indicate that NTP and provide complementary supervision: token prediction supports transferable semantics, whereas segment prediction is especially beneficial for structurally demanding cross-modal retrieval.

Table 3: Sensitivity to the segment-prediction coefficient λ on the LFM backbone. All configurations optimize $\mathcal{L}_{\text{NTP}} + \lambda\mathcal{L} + 0.02\mathcal{L}_{\text{ent}}$ with otherwise identical settings. For each coefficient, we report the best MMEB-36 P@1 among the common step-200, step-400, and step-600 checkpoints. Bold and underline denote the best and second-best results, respectively.

λ	Selected step	MMEB-36 P@1 (%)
0	400	29.167
0.03	200	**32.558**
0.10	600	<u>29.297</u>
0.20	600	28.958
0.50	200	23.764

4.3 Analysis and Ablations

The ablations validate the coarse-to-fine design of JoTS. We analyze three questions: whether token- and segment-level prediction provide complementary training signals, how strongly the segment objective should influence optimization, and whether entropy regularization improves the balance of the learned representation. First, removing NTP reveals that fine-grained token prediction remains essential for semantic acquisition, whereas removing shows that token prediction alone does not attain the same broad transfer performance. Second, the weight sweep demonstrates that is most effective as a lightweight auxiliary objective. Finally, the entropy and image-input ablations show that the complete objective promotes a more balanced representation and makes available visual evidence useful to retrieval.

NTP provides semantic grounding, while improves cross-task transfer. Removing NTP decreases the MMEB average from 32.558 to 17.156 and reduces MM-BRIGHT Tasks 2–4 from 12.739, 27.304, and 17.998 to 2.136, 20.450, and 8.042, respectively. Thus, segment prediction alone preserves some coarse multimodal structure, most visibly on Task 3, but cannot replace token-level semantic learning. Conversely, removing produces a much smaller but systematic loss of broad transfer: the MMEB average decreases by 3.391 points, with drops of 8.180, 5.220, and 3.100 points on classification, VQA, and visual grounding. Although the NTP-only variant remains strong on several dedicated retrieval columns, the joint objective performs better across the heterogeneous MMEB task distribution. therefore complements rather than duplicates NTP by encouraging the model to preserve segment-scale structure alongside token-level semantics.

A small coefficient provides the best optimization balance. Table 3 shows a clear non-monotonic relationship between segment supervision and downstream retrieval. Introducing with $\lambda = 0.03$ improves MMEB-36 from 29.167 to 32.558, confirming that segment-level supervision contributes beyond NTP alone. Increasing the coefficient to 0.10, 0.20, or 0.50 progressively reduces the benefit, with the strongest weighting yielding only 23.764. Because operates over a compact dynamic segment vocabulary, an excessive coefficient can prioritize within-group continuation discrimination over the transferable token-level semantics required by heterogeneous retrieval tasks. The selected value of 0.03 allows to shape the representation without overwhelming the principal NTP signal.

Entropy regularization improves representation balance rather than a single benchmark column. Removing $\mathcal{L}_{\text{ent}}$ changes the MMEB average only modestly, from 32.558 to 32.344, but redistributes performance toward the dedicated retrieval columns. The complete objective instead improves VQA by 2.190 points and visual grounding by 2.600 points over the no-entropy variant, while retaining a higher overall MMEB average. This pattern is consistent with the role of entropy regularization: it discourages overly concentrated inter-chunk routing and preserves information useful across multiple task types. We therefore retain the entropy term in the main model because the target is a generally transferable multimodal retriever rather than a configuration tuned to one evaluation format.

Joint prediction makes visual query evidence actionable. Table 4 separates Task 2 performance with and without query images. The complete JoTS model gains 0.724 nDCG@10 from the avail-

Table 4: Effect of query images on MM-BRIGHT Task 2 for the LFM ablations. "Images available" uses every accessible query image, whereas "Text only" removes all query images while keeping the candidate collection unchanged. We report the 29-domain macro nDCG@10. Δ_{img} is defined as *Images available* minus *Text only*; positive values denote a benefit from visual query evidence.

Variant	Images available	Text only	Δ_{img}
Full JoTS	12.739	12.015	$+0.724$
w/o	15.629	15.664	-0.035
w/o NTP	2.136	2.629	-0.493
w/o entropy regularization	16.025	14.625	$+1.400$

able images, and the variant that retains both NTP and while removing only entropy regularization gains 1.400 points. In contrast, removing either or NTP eliminates this positive image contribution. This result is particularly informative because the candidate collection is identical in both settings: the difference arises from whether the learned query representation can exploit its visual content. The comparison therefore supports the central design choice of combining token-level semantic modeling with segment-level multimodal structure prediction.

5 Limitations

Our study focuses on self-supervised multimodal retrieval using an OBELICS-derived training corpus and two complementary evaluation suites, MMEB and MM-BRIGHT. Although these benchmarks cover diverse retrieval formats and domains, they cannot represent every language, document structure, or image-intensive retrieval scenario encountered in practice. In addition, JoTS treats each complete text or image element as an atomic segment and constructs the segment-prediction vocabulary within the current training group. This design provides a controlled and efficient learning signal, but does not explicitly model hierarchical relations among sentences, images, chunks, and documents over longer contexts. Our experiments further concentrate on two vision-language backbones spanning 450M to 3B parameters; extending the analysis to larger architectures and broader pretraining corpora would clarify how the benefits of joint token- and segment-level prediction scale. Promising directions include adaptive or hierarchical segmentation, cross-group segment memories, and training corpora with greater linguistic and visual diversity.

6 Conclusion

We introduced JoTS, a self-supervised framework that trains multimodal Retrievers directly from interleaved image–text documents without annotated or synthetic query–document pairs. JoTS combines retrieval-conditioned next-token prediction with segment-continuation prediction, supported by constraint-preserving semantic difficulty grouping. Across MMEB and MM-BRIGHT, JoTS improves vision–language backbones of different scales and performs strongly against representative supervised and contrastively trained Retrievers. These results establish JoTS as an effective approach for transforming native token- and segment-level document continuity into supervision for stronger standalone multimodal Retrievers. Future work will examine its scalability to larger and more diverse corpora and backbones.

AI Use Statement

During the preparation of this manuscript, we used large language models (LLMs) as general-purpose assistants. The LLMs were used to improve textual clarity, wording, and grammatical structure; assist with the formatting of LaTeX tables and document layouts; and support code refinement and figure preparation.

Reproducibility Statement

The Appendix provides detailed descriptions of data construction, model architecture, training objectives, hyperparameters, prompts, and evaluation protocols. It also reports dataset statistics and additional experimental results to support reproduction and verification of our findings.

References

Abdelrahman Abdallah, Mohamed Darwish Mounis, Mahmoud Abdalla, Mahmoud SalahEldin Kasem, Mostafa Farouk Senussi, Mohamed Mahmoud, Mohammed Ali, Adam Jatowt, and Hyun Soo Kang. MM-BRIGHT: A multi-task multimodal benchmark for reasoning-intensive retrieval. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 8604–8612, New York, NY, USA, 2026. Association for Computing Machinery. doi: 10.1145/3770855.3817457. URL https://doi.org/10.1145/3770855.3817457.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens, 2024. URL https://arxiv.org/abs/2406.11271.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Fengyu Cai, Tong Chen, Xinran Zhao, Sihao Chen, Hongming Zhang, Sherry Tongshuang Wu, Iryna Gurevych, and Heinz Koeppl. Revela: Dense retriever learning via language modeling, 2025. URL https://arxiv.org/abs/2506.16552.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier J. Henaff. Data curation via joint example selection further accelerates multimodal learning, 2024. URL https://arxiv.org/abs/2406.17711.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2025. URL https://arxiv.org/abs/2407.01449.

Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2843–2853, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.203. URL https://aclanthology.org/2022.acl-long.203/.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020. URL https://arxiv.org/abs/2002.08909.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=jKN1pXi7b0.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models, 2024a. URL https://arxiv.org/abs/2407.12580.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2024b. URL https://arxiv.org/abs/2410.05160.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. URL https://arxiv.org/abs/2306.16527.

Jaewoo Lee, Joonho Ko, Jinheon Baek, Soyeong Jeong, and Sung Ju Hwang. Unified multimodal interleaved document representation for retrieval, 2026. URL https://arxiv.org/abs/2410.02729.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms, 2025. URL https://arxiv.org/abs/2411.02571.

Liquid AI. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations, 2018. URL https://arxiv.org/abs/1803.02893.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.466. URL https://aclanthology.org/2021.naacl-main.466/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Replug: Retrieval-augmented black-box language models, 2023. URL https://arxiv.org/abs/2301.12652.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality, 2024. URL https://arxiv.org/abs/2307.05222.

Yixuan Tang and Yi Yang. Dream: Dense retrieval embeddings via autoregressive modeling, 2026. URL https://arxiv.org/abs/2606.24667.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers, 2023. URL https://arxiv.org/abs/2311.17136.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 538–548, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.35. URL https://aclanthology.org/2022.emnlp-main.35/.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Wenjie Yang, Hang Yu, Yuyu Guo, and Peng Di. Solar: Self-supervised joint learning for symmetric multimodal retrieval, 2026. URL https://arxiv.org/abs/2605.15868.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL https://arxiv.org/abs/2205.01917.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms, 2024. URL https://arxiv.org/abs/2412.16855.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text, 2023. URL https://arxiv.org/abs/2304.06939.

A Additional Method Details

This appendix provides additional details of the interaction between the Retriever and the Reference model, including the construction of the retriever-derived routing distribution, its use in in-batch attention (IBA), and the token- and segment-level predictive objectives. Both next-token prediction (NTP) and egment-continuation prediction (SCP) are computed from the same second Reference forward pass augmented with IBA; they differ in the granularity and location at which prediction targets are defined.

A.1 Retriever-Derived Chunk Routing

Each training group contains (B=16) multimodal chunks. The Retriever encodes every chunk twice using two role-specific prompts. The query-side prompt asks what external context would help interpret the current chunk, whereas the passage-side prompt asks what information the chunk can provide to another chunk.

Let $q_i, p_i \in \mathbb{R}^d$ denote the query and passage representations of chunk i, respectively. Both representations are extracted from the final native end-of-turn (EOT) token after the complete interleaved chunk and are ℓ_2-normalized:

$$q_i = \frac{h^q_{i,\mathrm{EOT}}}{\left\| h^q_{i,\mathrm{EOT}} \right\|_2}, \qquad p_i = \frac{h^p_{i,\mathrm{EOT}}}{\left\| h^p_{i,\mathrm{EOT}} \right\|_2}. \tag{11}$$

The directed chunk-level similarity matrix is

$$S_{ij} = q_i^\top p_j. \tag{12}$$

Because the representations are normalized, the dot product is equivalent to cosine similarity. We mask the diagonal to prevent a chunk from reading its own source cache through IBA:

$$\widetilde{S}_{ij} = \begin{cases} -\infty, & i = j, \\ S_{ij}, & i \neq j. \end{cases} \tag{13}$$

The routing distribution is then

$$A_{ij} = \frac{\exp(\widetilde{S}_{ij}/\tau_{\text{ret}})}{\sum_{k=1}^{B} \exp(\widetilde{S}_{ik}/\tau_{\text{ret}})}, \qquad \tau_{\text{ret}} = 0.01. \tag{14}$$

Thus, $A_{ii} = 0$ and $\sum_j A_{ij} = 1$. The matrix A is not trained with document-identity labels or an explicit query–document contrastive objective. Instead, its learning signal is induced by the downstream predictive objectives through the Reference model.

A.2 Reference Forward Passes and In-Batch Attention

The Reference model processes each group twice. The first forward pass encodes the unprompted interleaved chunks and constructs source key–value caches $\{K_j^{(\ell)}, V_j^{(\ell)}\}_{j=1}^{B}$, where ℓ indexes the IBA-enabled attention layers. This source pass is executed without gradient tracking. Consequently, the cached source representations are constants within the current optimization step, although they are recomputed using the current Reference parameters at the next step.

The second Reference forward pass processes the same chunks with ordinary causal self-attention and additionally reads the source caches of the other chunks. For target token (t) in chunk (i), the token-level context from source chunk (j) is

$$C_{ijt}^{(\ell)} = \text{softmax}\left(\frac{Q_{it}^{(\ell)} {K_j^{(\ell)}}^{\top}}{\sqrt{d_h}}\right) V_j^{(\ell)}. \tag{15}$$

The chunk-level routing distribution combines these contexts as

$$C_{it}^{(\ell)} = \sum_{j \neq i} A_{ij} C_{ijt}^{(\ell)}, \tag{16}$$

and the attention output is

$$Y_{it}^{(\ell)} = \text{SelfAttn}_{it}^{(\ell)} + \gamma W_O^{(\ell)} C_{it}^{(\ell)}, \qquad \gamma = 0.1. \tag{17}$$

IBA therefore contains two nested selection mechanisms: token-level attention selects information within each source chunk, while A_{ij} distributes attention mass across source chunks. Both textual and visual source positions remain accessible to IBA. The diagonal mask is applied only to the cross-chunk IBA branch, while ordinary causal self-attention within the current chunk is always retained.

A.3 Selective Next-Token Prediction

NTP is defined only over genuine textual content from the source documents. Let $(\mathcal{T})$ denote the set of positions whose next element is a real text token. Image features, chat-template tokens, padding positions, and special tokens are excluded from the target set. The token-level objective is

$$\mathcal{L}_{\text{NTP}} = -\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \log p_\theta \left(x_{t+1} \mid x_{\leq t}, C_t^{\text{IBA}}\right). \tag{18}$$

This selective formulation is mathematically equivalent to shifted causal language-model cross-entropy on the supervised positions, while avoiding the materialization of vocabulary logits for unsupervised visual and template positions.

Visual features do not serve as discrete NTP targets. They nevertheless condition subsequent text prediction through causal self-attention and can be accessed by other chunks through IBA. NTP gradients therefore update the Reference parameters used in the second forward pass and affect the Retriever through (A), but do not define a pixel- or image-token reconstruction loss.

A.4 NEXT-SEGMENT PREDICTION

Here, SCP denotes *next-segment prediction*, a multi-way segment continuation objective, rather than BERT's binary next-sentence prediction. A segment corresponds to one non-empty element of the aligned multimodal sequence stored in a chunk. It is either one complete text span or one image; adjacent elements are not merged solely because they share a modality.

For one training group, we flatten all real segments from its 16 chunks into a group-local candidate set

$$\mathcal{V}_{\text{seg}} = \{s_1, \ldots, s_M\}. \tag{19}$$

The number (M) varies with group contents. Every segment instance retains a separate candidate index even when two segments have identical contents.

Each candidate is independently encoded by the Retriever. If $\mathcal{I}_m$ denotes the hidden-state span corresponding to the real content of segment (s_m), its representation is

$$c_m = \frac{1}{|\mathcal{I}_m|} \sum_{u \in \mathcal{I}_m} H^{\text{ret}}_{m,u}, \qquad z_m = \text{norm}(W_{\text{key}} c_m). \tag{20}$$

Prompt tokens, chat-template tokens, and the final EOT position are excluded from this pooling operation. The candidate matrix is

$$Z = [z_1; \ldots; z_M] \in \mathbb{R}^{M \times d}. \tag{21}$$

Consider two adjacent segments $s_{i,k} \to s_{i,k+1}$ within chunk (i). The second Reference forward provides a causal hidden state at the final safe content position $b_{i,k}$ of the current segment:

$$h_{i,k} = H^{\text{ref},2}_{i,b_{i,k}}, \qquad r_{i,k} = \text{norm}(W_{\text{context}} h_{i,k}). \tag{22}$$

Because the Reference model is causal, this state can observe the current segment and its preceding within-chunk context but cannot observe the target segment. It can additionally access the other 15 chunks through IBA.

The segment logits are

$$g_{e,m} = \frac{r_e^\top z_m}{\tau_{\text{seg}}}, \qquad \tau_{\text{seg}} = 0.07, \tag{23}$$

where e indexes SCP events. The candidate corresponding to the current segment is masked, while all other text and image segments in the group remain eligible. Let y_e be the candidate index of the immediate next segment and $\mathcal{M}_e$ the valid candidate set. The SCP loss is

$$\mathcal{L}_{\text{SCP}} = -\frac{1}{E} \sum_{e=1}^{E} \log \frac{\exp(g_{e,y_e})}{\sum_{m \in \mathcal{M}_e} \exp(g_{e,m})}. \tag{24}$$

If chunk i contains m_i segments, it produces $m_i - 1$ SCP events. The final segment does not produce an event because the current implementation contains neither an EOS candidate nor a cross-chunk continuation target.

A.5 JOINT OBJECTIVE AND GRADIENT PATHS

To discourage premature collapse of the chunk-routing distribution, we use a lower-bound entropy penalty. Let

$$\overline{\mathcal{H}}(A) = \frac{1}{B \log(B-1)} \sum_{i=1}^{B} \left(-\sum_{j \neq i} A_{ij} \log A_{ij} \right). \tag{25}$$

The unweighted regularizer is

$$\mathcal{L}_{\text{ent}} = \left[\max\left(0, h_0 - \overline{\mathcal{H}}(A)\right) \right]^2, \qquad h_0 = 0.35. \tag{26}$$

Table 5: Gradient paths induced by the predictive objectives. The detached source cache is recomputed at every optimization step but does not receive gradients within the current step.

Objective	Retriever matrix A	Retriever candidate encoder	Reference second forward	Source K/V cache	Segment projections
$\mathcal{L}_{\mathrm{NTP}}$	Yes	No	Yes	No	No
$\mathcal{L}_{\mathrm{SCP}}$	Yes	Yes	Yes	No	Yes
$\mathcal{L}_{\mathrm{ent}}$	Yes	No	No	No	No

Table 6: The 36 dataset-level tasks in MMEB-V1, grouped by the four official meta-task categories. Dataset identifiers follow the official evaluation implementation.

Meta-task	Count	Dataset-level tasks
Classification	10	ImageNet-1K; N24News; HatefulMemes; VOC2007; SUN397; Place365; ImageNet-A; ImageNet-R; ObjectNet; Country211.
Visual question answering	10	OK-VQA; A-OKVQA; DocVQA; InfographicsVQA; ChartQA; Visual7W; ScienceQA; VizWiz; GQA; TextVQA.
Multimodal retrieval	12	VisDial; CIRR; `VisualNews_t2i`; `VisualNews_i2t`; `MSCOCO_t2i`; `MSCOCO_i2t`; NIGHTS; WebQA; FashionIQ; Wiki-SS-NQ; OVEN; EDIS.
Visual grounding	4	MSCOCO; RefCOCO; RefCOCO-Matching; Visual7W-Pointing.
Total	36	—

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{NTP}} + \lambda_{\mathrm{SCP}}(t)\mathcal{L}_{\mathrm{SCP}} + \lambda_{\mathrm{ent}}\mathcal{L}_{\mathrm{ent}}, \tag{27}$$

where

$$\lambda_{\mathrm{SCP}}(t) = \lambda_{\mathrm{SCP}}^{\max} \min\left(1, \frac{t}{T_{\mathrm{warm}}}\right), \qquad T_{\mathrm{warm}} = 50, \qquad \lambda_{\mathrm{ent}} = 0.02. \tag{28}$$

B Experimental Details

B.1 Evaluation Benchmarks

MMEB-V1. We evaluate general-purpose multimodal embedding quality on the original Massive Multimodal Embedding Benchmark (MMEB-V1) (**?**). MMEB formulates classification, visual question answering, multimodal retrieval, and visual grounding as instruction-guided ranking problems. A model embeds a query and a set of task-specific candidates and returns the candidate with the highest similarity. The benchmark contains 36 dataset-level tasks: 20 have in-distribution training data, while 16 are out-of-distribution evaluation tasks.

MM-BRIGHT. MM-BRIGHT is a reasoning-intensive multimodal retrieval benchmark containing 2,803 real-world queries from 29 technical domains (Abdallah et al., 2026). Its queries include diagrams, charts, screenshots, scientific figures, and photographs whose relevance may not be recoverable through surface-level lexical matching. The benchmark defines four retrieval tasks with progressively richer query and target modalities. We evaluate Tasks 2–4 because our study focuses on retrieval settings with multimodal queries. Task 1, whose queries and candidates are both text-only, is excluded from our primary evaluation.

We use nDCG@10 following the official MM-BRIGHT evaluation design. Tasks 1–3 use binary relevance. Task 4 uses graded relevance: a gold passage paired with an annotated positive image receives relevance 2, the text-only version of a gold passage receives relevance 1, and an incorrect candidate receives relevance 0.

Table 7: The four MM-BRIGHT retrieval tasks. Our evaluation uses Tasks 2–4. MM denotes a joint text–image input.

Task	Query modality	Target modality	Used	Description
1	Text	Text document	No	Text-to-text reasoning-intensive retrieval baseline.
2	Text + image	Text document	Yes	Retrieves relevant textual passages using a multimodal query.
3	Text + image	Image	Yes	Retrieves relevant images using textual intent and visual query information.
4	Text + image	Text + image	Yes	Retrieves multimodal candidates whose textual and visual components are jointly evaluated for relevance.

Table 8: The 29 technical domains included in MM-BRIGHT.

Domain group	Domains
STEM and life sciences	Academia, Bioacoustics, Bioinformatics, Biology, Chemistry, Earth Science, Mathematics, Medical Sciences, and Physics.
Software and technical systems	Apple, Ask Ubuntu, Bitcoin, Cryptography, GIS, Quantum Computing, Robotics, and Salesforce.
Social sciences and humanities	Christianity, Economics, Islam, Law, Philosophy, and Psychology.
Applied domains	Aviation, Gaming, Project Management, Quantitative Finance, Sustainability, and Travel.

We use MM-BRIGHT revision `[TODO: dataset revision]` and the `[TODO: paper or legacy]` Task-4 protocol. All compared models use the same qrels, candidate construction, and evaluation implementation.

B.2 Source Corpus: OBELICS

Our pretraining data are derived from OBELICS (Laurençon et al., 2023), an open web-scale corpus of naturally interleaved image–text documents extracted from Common Crawl. Unlike image–caption corpora that isolate individual images and their alternative text, OBELICS preserves the ordering of textual and visual elements within the simplified HTML document structure. This property enables us to derive token- and segment-level predictive targets from naturally occurring multimodal sequences.

B.3 Chunk and Group Construction

We construct chunks as non-overlapping contiguous slices of the original interleaved document sequence. The construction preserves the order of text spans and images. Each chunk contains at most 256 text tokens and at most four images, and consecutive chunks do not overlap.

A segment is one non-empty element of the aligned text–image arrays. For example, the stored representation

$$\begin{aligned} \texttt{images} &= [\varnothing, I_1, \varnothing], \\ \texttt{texts} &= [T_1, \varnothing, T_2] \end{aligned} \tag{29}$$

is restored at runtime as

$$T_1 \to I_1 \to T_2. \tag{30}$$

Null entries are alignment placeholders and are not converted into model tokens. We neither move images to the beginning of a chunk nor merge stored elements into a single SCP segment.

Table 9: Official corpus statistics of OBELICS.

Statistic	Value
Common Crawl snapshots	25
Crawl period	Feb. 2020–Feb. 2023
Final documents	141M
Text tokens	115B
Associated images	353M
Unique images	298M
Unique-image ratio	84.3%
Median tokens per document	677
Median images per document	1
Maximum images per document	30

Table 10: Main stages of the official OBELICS construction pipeline. These statistics describe the source corpus rather than our selected training subset.

Stage	Scale after stage	Main operations
Initial web-document processing	41.2B documents	Common Crawl collection and preliminary text extraction.
Language, quality, and early deduplication	1.1B documents	English filtering, MinHash deduplication, repetition filtering, and text-quality classification.
Multimodal extraction	Approximately 2B downloaded images	HTML DOM simplification and extraction of text and image nodes while preserving document order.
Node-, paragraph-, and document-level filtering	365M documents; 1.4B images	Image-size and aspect-ratio filters, text-quality filters, and removal of boilerplate and low-quality documents.
Responsible filtering and final deduplication	141M documents; 353M images	Opt-out handling, NSFW filtering, and image-, document-, and paragraph-level deduplication.

B.4 Structure-Preserving Semantic Difficulty Grouping

The semantic difficulty grouping procedure operates only on the chunks already selected in the base manifest. It neither modifies chunk contents nor changes their within-document order. We collect the ordered chunks from each source document into an indivisible bundle b_i and associate it with the load signature

$$\ell(b_i) = \left(n_i^{\text{chunk}}, n_i^{\text{image}}, n_i^{\text{text}}, n_i^{\text{multi}}, t_i^{\text{single}}\right). \tag{31}$$

where the entries denote the number of chunks, total images, text-only chunks, multi-image chunks, and whether the bundle belongs to the single-image source category.

Each original group defines an ordered signature template

$$T_g = (\ell_{g,1}, \ldots, \ell_{g,m_g}). \tag{32}$$

The bundle assigned to slot (k) must exactly match $(\ell_{g,k})$. Per-slot replacement therefore preserves the number of chunks, images, text-only chunks, multi-image chunks, and source-document types for every group. In particular, every group continues to contain exactly 16 chunks. The selected data contain 69 observed signatures; 11 bundles have singleton signatures and therefore have no alternative same-signature assignment.

We use a frozen Qwen3-Embedding-0.6B encoder to obtain a normalized 1,024-dimensional representation (z_i) from the textual content of each source document (Bai et al., 2025). The resulting cosine similarity

$$s(b_i, b_j) = z_i^\top z_j \tag{33}$$

is a text-derived proxy for document-level semantic relatedness rather than a direct measurement of multimodal similarity. Semantic representations are available for 25,213 of the 26,049 bundles; the remaining 836 bundles follow a recorded fallback path. HDBSCAN clusters are used only as a weak auxiliary assignment signal and are not treated as semantic labels.

Table 11: Statistics and constraints of the selected training manifest.

Property	Value
Selected chunks	65,792
Document bundles	26,049
Training groups	4,112
Chunks per group	16
Maximum chunks per document	8
Maximum text tokens per chunk	256
Maximum images per chunk	4
Maximum text-only chunks per group	4
Multi-image chunks	approximately 39.1%
Images per group	14–35
Median images per group	23
Mean images per group	23.285

Table 12: Fulfillment of requested anchor-to-document roles.

Requested role	Number of slots	Fulfilled (%)
Near	7,914	99.90
Intermediate	7,872	96.20
Far	6,151	82.39

We estimate the empirical similarity distribution using 200,000 randomly sampled document pairs under the fixed construction seed. Its 30th and 70th percentiles define three intervals:

$$\begin{aligned} \mathcal{R}_{\text{far}} &= \{s : s \le 0.1666\}, \\ \mathcal{R}_{\text{intermediate}} &= \{s : 0.1666 < s < 0.2441\}, \\ \mathcal{R}_{\text{near}} &= \{s : s \ge 0.2441\}. \end{aligned} \tag{34}$$

The first document slot of each group is used as the anchor. The remaining slots are assigned requested near, intermediate, and far roles according to a 35%/35%/30% policy, with integer allocations obtained by the largest-remainder method. This policy is defined over anchor-to-document slots aggregated across the dataset; it is neither an exact per-group constraint nor an exact constraint on all within-group document pairs.

Let $\mathcal{U}$ be the unused bundle set. The feasible candidates for slot (g, k) are

$$\mathcal{C}_{g,k} = \{b \in \mathcal{U} : \ell(b) = \ell_{g,k}\}. \tag{35}$$

We consider at most 512 feasible candidates and select

$$b^*_{g,k} = \arg\max_{b \in \mathcal{C}_{g,k}} \left[u_{r_{g,k}}(s(a_g, b)) - P(a_g, b) \right], \tag{36}$$

where (u_r) rewards compatibility with the requested interval and (P) discourages same-host, duplicate, near-identical, or same-original-group assignments. Assignment is global and without replacement. Requested near roles are processed before intermediate and far roles because they have the most restricted feasible candidates.

Only anchor-to-document slots receive requested roles. Across all within-group document pairs, including pairs between two non-anchor bundles, the realized distribution is 32.54% near, 37.96% intermediate, and 29.50% far. Because a bundle may contain multiple chunks, these bundle-level proportions should not be interpreted as exact constraints on the chunk-pair distribution used by the model.

The semantic roles and similarity values used during construction are not exposed to the model and do not enter either predictive loss. They determine only which immutable document bundles coexist in a group, thereby reshaping the cross-document context available through IBA and the candidate distribution used by SCP. This procedure is consequently distinct from supervised hard- negative mining and from curriculum learning with a time-varying difficulty schedule.

Table 13: Integrity checks for the regrouped training manifest.

Audit item	Result
Number of groups	4,112, unchanged
Chunks used exactly once	65,792 / 65,792
Bundles used exactly once	26,049 / 26,049
Within-bundle chunk order preserved	Yes
Per-slot load signatures preserved	Yes
Cross-document pairs with cosine (¿0.95)	0

B.5 PROMPTS

The same role prefixes are used during training and downstream evaluation. Benchmark-provided task instructions and query contents are retained and wrapped by our query- or passage-side role prefix. We do not additionally apply the generic Qwen3-Embedding instruction template.

Retriever query prompt.

```
Represent the following multimodal chunk by the external
context that would help predict and understand it.
Multimodal chunk:
```

Retriever passage prompt.

```
Represent the following multimodal chunk by the information
it can provide to help predict and understand another chunk.
Multimodal chunk:
```

Segment-candidate prompt.

```
Represent the following segment as a candidate continuation
of multimodal context.
Segment:
```

For chunk-level encoding, the input is

$$[\text{role prefix}] + [\text{original interleaved chunk}] + [\text{native chat EOT}]. \tag{37}$$

No natural-language suffix is appended, and query and passage embeddings are read from the final native EOT hidden state. For SCP candidates, the input is

$$[\text{segment prompt}] + [\text{one complete text or image segment}] + [\text{native chat EOT}], \tag{38}$$

and the representation is obtained by masked mean pooling over the real segment-content span rather than the EOT position. The Reference model receives the original interleaved chunk without any Retriever role prefix or segment prompt.

B.6 TRAINING CONFIGURATION

Each DDP rank processes one complete group and constructs its own 16×16 routing matrix and group-local segment candidate set. There is no cross-rank candidate all-gather and no (64×64) routing matrix. DDP synchronizes and averages parameter gradients after backward propagation.

C AI USE DISCLOSURE

Generative AI tools, including OpenAI ChatGPT and Codex, were used to assist with literature discovery, methodological discussion, experiment planning, the organization of technical documentation, and the drafting and editing of manuscript text and LaTeX tables. These tools were also used for software implementation and debugging where applicable.

Table 14: Training hyperparameters of the primary configuration.

Hyperparameter	Setting
Backbone	LFM2.5-VL-450M
Distributed training	4 DDP ranks
Groups per rank per step	1
Chunks per group	16
Effective chunks per global step	64
Numerical precision	bfloat16
Optimizer	AdamW
Peak learning rate	1×10^{-4}
Learning-rate schedule	Linear decay
Schedule horizon	1,000 steps
Warmup	50 steps
Weight decay	0
Adam coefficients	$(0.9, 0.999)$
Adam epsilon	10^{-8}
Gradient clipping	1.0
Training seed	42
Retriever temperature	0.01
SCP temperature	0.07
IBA scale	0.1
Entropy weight	0.02
SCP maximum weight	$\lambda_{\mathrm{SCP}}^{\max}$

All AI-assisted outputs were reviewed by the authors. Bibliographic claims were checked against the cited papers or official project resources, and all reported experiments were executed and inspected by the authors. Generative AI tools were not treated as authoritative sources of scientific evidence. The authors take responsibility for the final methodology, implementation, experimental results, citations, and manuscript content.