AMUSE: Abstracting Modalities Using Schematized Embeddings

Anonymous ACL submission

Abstract

We introduce a training-free, modality-agnostic retrieval framework: every input – image, audio clip, or document – is converted into a schematized textual proxy and retrieved through a single off-the-shelf text-embedding index, with no task-specific training or fine-tuning. Our method employs multimodal large language models (MLLMs) to transform unstructured inputs into these proxies: structured descriptions that adhere to domain-specific schemas. The result is what we term *schematized embeddings*: text embeddings derived from schema-driven representations rather than raw inputs or free-form captions. Unlike opaque learned representations, schematized embeddings are fully interpretable: each retrieval decision can be traced to explicit textual attributes, enabling transparent reasoning about why items match. The framework adapts to new modalities and domains through schema-level adjustments rather than retraining, offering rapid deployment without data collection or fine-tuning. We evaluate AMUSE across image retrieval (CUB, Cars196, SOP), text retrieval (FiQA, LoCo), and text-to-audio retrieval (DCASE) benchmarks. On Cars196, our approach establishes a new state-of-the-art, surpassing the best trained method by nearly 2 percentage points in Recall@1. On text retrieval, AMUSE outperforms strong baselines including BM25 and BGE-large, achieving the highest Recall@100 on FiQA. On audio retrieval, AMUSE reaches approximately 80% of trained baseline performance without any audio-specific training. Our analysis reveals that schema design is a primary lever for improvement, particularly on fine-grained image retrieval where well-crafted schemas dominate model choice; on text and audio the underlying MLLM also contributes substantially. These findings suggest that schematized text can serve as a useful unifying medium for retrieval across modalities.

1 Introduction

Retrieval systems increasingly need to operate across heterogeneous modalities – text, images, audio, video, and beyond – yet most production pipelines remain modality-specific. The dominant response has been to align modalities in a shared latent space through large-scale pre-training (Radford et al., 2021; Girdhar et al., 2023; Han et al., 2024) or through retrieval-specific fine-tuning of multimodal LLMs (MLLMs) (Lin et al., 2025; Kong et al., 2025; Zhang et al., 2025). These approaches achieve strong cross-modal alignment but require costly data curation, careful negative sampling, and continuous re-training as new modalities or domains appear. They also produce opaque dense embeddings whose retrieval decisions cannot be inspected by a practitioner.

We propose an alternative that decouples modality understanding from retrieval mechanics. Our framework, **AMUSE** (*Abstracting Modalities Using Schematized Embeddings*), translates each input (image, audio clip, document, etc.) into a *schematized textual proxy*: a structured natural-language description that adheres to a domain-specific schema. The proxy is then embedded with an off-the-shelf text encoder and indexed in a standard vector database. We use the term *schematized embeddings* to denote text embeddings derived from schema-driven descriptions rather than raw inputs or free-form captions, emphasizing that the semantic content has been organized according to an explicit template specifying fields, granularity, and format. This design is entirely *training-free*, modality-agnostic, and *interpretable* by construction – every retrieval decision can be traced to a small set of human-readable attributes.

Three findings emerge from a controlled evaluation across image (CUB, Cars196, SOP), text (FiQA, LoCo), and text-to-audio (DCASE) retrieval. First, schematized proxies match or surpass

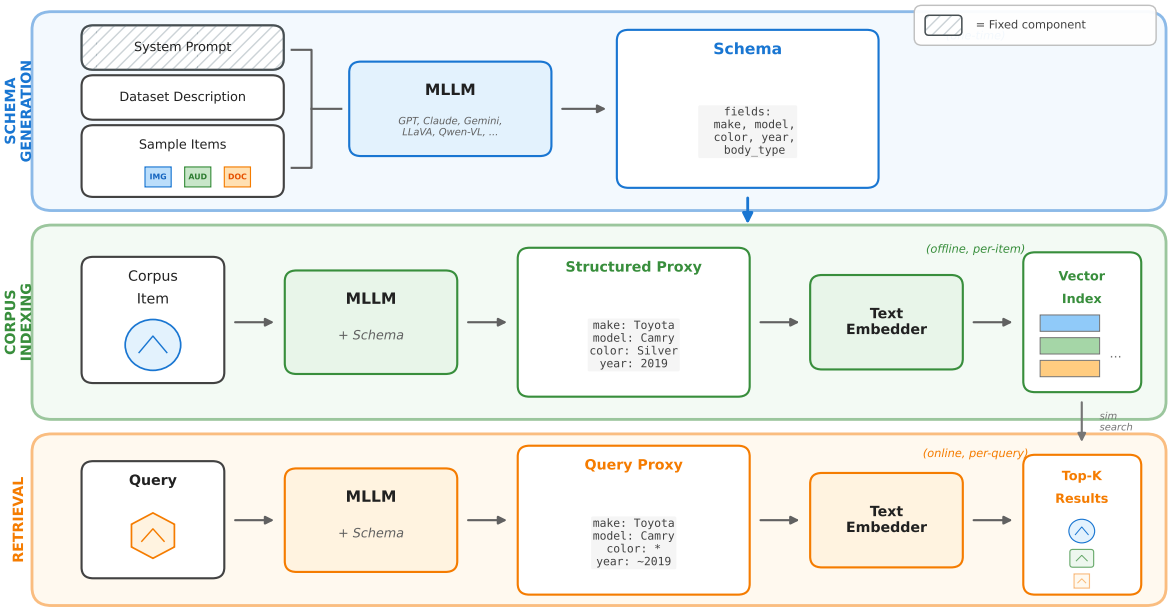

Figure 1: Overview of the AMUSE pipeline. Schema generation (top) produces a reusable schema from a fixed system prompt, dataset description, and sample items. Corpus indexing (middle) converts each item into a schematized proxy via an MLLM, then embeds it into a vector index. Retrieval (bottom) applies the same transformation to the query and performs similarity search to return top-k results.

trained baselines on fine-grained image retrieval, setting a new state of the art on Cars196 despite using no task-specific data. Second, controlled ablations decompose the schema's contribution into two effects – *task awareness* and *structural canonicalization* – and show that on fine-grained settings only the second is hard to approximate with a free-form prompt. Third, field-level ablation reveals that nearly all of the retrieval signal is carried by a small set of domain-core fields; affective and contextual descriptors function as decorative filler, so schema engineering effort should concentrate on identifying load-bearing attributes.

Contributions. (i) We introduce schematized embeddings as a training-free, interpretable substrate for modality-agnostic retrieval, decoupling modality understanding from indexing. (ii) We provide the first controlled comparison of schemas against free-form captions (with and without task instruction) on fine-grained image retrieval, separating task awareness from structural canonicalization. (iii) We perform a field-level ablation that pinpoints *which* schema fields carry retrieval signal on Cars196. (iv) We demonstrate the framework across three modalities, achieving state of the art on Cars196 and viable zero-shot performance on text-to-audio retrieval.

2 Related Work

Cross-modal and multimodal retrieval. Two paradigms dominate the literature. Contrastive pre-training methods learn paired-modality encoders such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), FLAVA (Singh et al., 2022), and AudioCLIP (Guzhov et al., 2022), achieving strong within-pair retrieval but requiring dedicated data per modality combination. MLLM-based methods take a single architecture and project additional modalities into a shared language space (Peng et al., 2024; Lu et al., 2023; Girdhar et al., 2023; Han et al., 2024); recent work refines this through modality-aware bi-encoders (Lin et al., 2025) and balanced contrastive objectives (Kong et al., 2025; Lee et al., 2025; Zeng et al., 2025). Both families produce opaque embeddings and require re-training for new modalities or domains. Lightweight prompt-based reuse of pretrained MLLMs (Koh et al., 2023; Jiang et al., 2025) and multimodal retrieval-augmented generation (Modarressi et al., 2024; Cong et al., 2023; Abootorabi et al., 2025; Drushchak et al., 2025) reduce the training burden but inherit the interpretability gap.

Concept-based explainability. Concept-Based Explainability (C-XAI) (Poeta et al., 2023) describes model behavior through human-readable abstractions. Concept Bottleneck Models (Koh et al., 2020) introduce an explicit layer of interpretable concepts and allow *concept intervention* at inference time, but discrete bottlenecks typically incur an accuracy-interpretability trade-off (Yuksekgonul et al., 2022; Oikarinen et al., 2023). AMUSE sits within C-XAI but replaces fixed learned concepts with open-ended schema-driven text generated by MLLMs, which preserves intervention while – as our experiments show – closing or eliminating the accuracy gap.

3 Method

Problem setting. We study retrieval over a corpus $\mathcal{C} = \{x_i\}_{i=1}^{N}$ where all items share a single modality (e.g., all images or all audio), and a query q that may be in the same or a different modality. The goal is to rank items in $\mathcal{C}$ by semantic relevance to q. AMUSE maps every x_i and q to a textual proxy in a shared text-embedding space, so the corpus modality and the query modality only need to share that proxy space; in this work we evaluate single-modality corpora with same-modality or cross-modal queries. We use text as the proxy because LLMs, VLMs, and text embedding models are by far the most reliable and most interpretable of the available components.

Pipeline. The pipeline (Figure 1) has three stages.

Schema generation. Once per dataset, an MLLM produces (a) a dataset-specific system prompt and (b) a structured schema specifying fields, allowed values, and ordering. The MLLM is fed a fixed global extraction guide, a user-provided dataset description, and 3–5 representative samples. The resulting schema is the template that every subsequent proxy must follow, ensuring corpus-wide structural consistency.

Indexing. Each corpus item x_i is passed to the MLLM along with the dataset-specific system prompt, producing a schematized textual proxy s_i that is restricted to the schema-defined fields. The proxy is embedded into $\mathbb{R}^d$ by a text embedding model and stored in a vector index. The stage is fully offline and scales linearly with corpus size.

Retrieval. A query q is passed through the same proxy-generation and embedding pipeline to obtain e_q. We perform nearest-neighbor search against the index and return the top-k items in their original modalities; the schematized proxies provide a transparent rationale for every match.

Domain adaptation. Adapting AMUSE to a new domain requires only editing the dataset description and schema (e.g., adding fields such as `instrument` or `material`). No retraining, no labeled data, and no encoder changes are needed – a property that supports rapid deployment to emerging modalities.

4 Experiments

Scope and models. We evaluate AMUSE on three modality families with single-modality corpora: image retrieval (CUB-200-2011 (Wah et al., 2011), Cars196 (Krause et al., 2013), SOP (Song et al., 2016)), text retrieval (FiQA (Maia et al., 2018; Thakur et al., 2021), LoCo v1 (Saad-Falcon et al., 2024)), and text-to-audio retrieval (DCASE 2024 Task 8, DCASE 2025 Task 6). All proxy-generation and embedding calls use Azure OpenAI Batch APIs. We test three MLLMs spanning the cost–capability range (GPT-4o-mini, GPT-4o (OpenAI, 2023), GPT-5 (OpenAI, 2025)) and two text-embedding models (text-embedding-3-small / -large). For audio, we use Gemini 2.5 / 3 Flash for proxy generation due to its native audio understanding. The framework is model-agnostic; these choices reflect availability rather than design constraints.

4.1 Image Retrieval

Setup. On CUB and Cars196 we report Recall@K with $K \in \{1, 2, 4, 8\}$; on SOP we report $K \in \{1, 10, 10^2, 10^3\}$, following Wang et al. (2025): R@K is the fraction of queries with at least one relevant item in the top K under cosine distance. We use the standard test/test protocol (test queries against the test corpus, relevance by class membership) with no train-set information used at any stage. Baseline numbers in Table 1 are taken directly from Wang et al. (2025).

Results. Table 1 compares AMUSE to recent deep metric learning and retrieval baselines. Crucially, all baselines except IRT_O require task-specific training on each dataset, whereas AMUSE is entirely training-free. On Cars196, AMUSE establishes a new state-of-the-art, surpassing the previous best trained method (PLEor) by 1.9 pp in R@1 and outperforming VCE by 2.7 pp. On

CUB-200-2011, AMUSE exceeds several trained baselines (CFGA, ProxyNCA++, Proxy Anchor) and matches PLEor's R@1, trailing VCE by 7 pp, a competitive gap given that VCE benefits from task-specific training. Compared to IRT_O, the only other training-free baseline, AMUSE delivers 16 pp higher R@1 on CUB, underscoring the advantage of schematized proxies over direct image-text embedding. On SOP, AMUSE lags behind trained methods; this dataset's 22,634 fine-grained product categories present a challenging setting where task-specific supervision provides significant benefit.

Foundation-model embedders. To address the concern that comparing a foundation-model-powered pipeline against trained metric-learning baselines is unfair, we also benchmark two direct MLLM-as-embedder baselines that, like AMUSE, exploit pretrained knowledge: zero-shot CLIP ViT-L/14 (Radford et al., 2021) (428M parameters, 768-dim embedding) and the dedicated multimodal embedder Qwen3-VL-Embedding 8B (Qwen Team, 2025) (8B parameters, 4096-dim embedding), both run in fp16. Each is evaluated in two modes: *image*, embedding the image alone, and *image+text*, fusing the image embedding with the embedding of a GPT-generated description of it. In the *image* mode, AMUSE outperforms both on Cars196 (96.3 vs. 87.7 R@1 for CLIP and 63.6 for Qwen) and matches the stronger one on CUB (74.8 vs. CLIP's 74.4); notably, the 8B fine-tuned embedder trails even zero-shot CLIP on these fine-grained tasks. The *image+text* fusion is highly asymmetric: on Cars196 it lifts Qwen by 18 pp (63.6→81.5), narrowing the gap to CLIP from 24 to 6 pp, while *degrading* CLIP (87.7→84.8); on the finer-grained CUB it fails to rescue Qwen (38.0→35.9) and collapses CLIP (74.4→58.3). The CUB result indicates that the free-form descriptions used for fusion lack the species-level discriminative cues that fine-grained bird retrieval demands – a manifestation of the same canonicalization gap quantified in Section 4.2. That a textual description helps a weaker visual embedder on a coarse task, hurts a strong one, and fails outright on a fine-grained one reinforces our central claim: the gains come from schematized *text*, not merely from invoking a foundation model. The cost asymmetry is also large: CLIP indexes both corpora at ∼300 images/s on a single GPU, whereas Qwen-8B runs at 4–11 images/s – a 50–80× throughput gap that further weighs against the dense multimodal embedder for large-scale indexing.

4.2 Schema vs. Free-Form Caption Ablation

Setup. A central claim of AMUSE is that structured schematized proxies outperform unstructured free-form descriptions. To isolate this effect, we ran a controlled ablation on the two fine-grained image benchmarks where structure should matter most: CUB-200-2011 and Cars196. Holding the MLLM (GPT-5) and embedding model (text-embedding-3-large) fixed, we compared three proxy-generation strategies: **Schema (AMUSE)**, the standard AMUSE proxy against the dataset's structured schema; **Free-form caption**, a generic "describe this image in thorough detail" prompt with no schema fields and no task-specific instructions; and **Free-form w/ instruction**, the same prompt augmented with a short task-oriented instruction that emphasizes discriminative attributes but still produces unstructured prose with no field structure. All three conditions share MLLM, embedding model, splits, and evaluation protocol; the only variable is the prompt.

Results. Table 2 reports Recall@K for the three conditions. The schema condition wins R@1 on both datasets, beating the plain free-form baseline by 19.3 pp on CUB and 14.9 pp on Cars196, confirming that schematized proxies are substantially more effective than unstructured captions. The more revealing comparison is against instruction-guided free-form text: on Cars196 the gap collapses to 0.8 pp R@1 and free-form w/ instruction is even marginally better at R@4 and R@8, while on the more fine-grained CUB the schema retains a 6.8 pp R@1 lead. This decomposes the schema's contribution into two distinct effects: (i) *task awareness* – the schema implicitly encodes which attributes matter, and accounts for most of the gap on Cars196 where coarse cues such as make, model, and color suffice; and (ii) *structural canonicalization* – consistent attribute terminology across the corpus, which remains decisive on CUB where species discrimination depends on standardized vocabulary for plumage, beak shape, and silhouette. The takeaway is that schemas are most valuable precisely in the regime where AMUSE targets fine-grained, attribute-dense modalities; for coarser categorization, a well-instructed free-form prompt approaches the schema's effectiveness.

Table 1: Recall@k (%) on CUB-200-2011, Cars196, and SOP. Trained-method numbers cited from Wang et al. (2025). Foundation-model embedders (CLIP, Qwen3-VL-Embedding) and AMUSE use no dataset-specific training; CLIP/Qwen are reported in two modes – *image* (image alone) and *image+text* (image fused with a GPT-generated description) – and SOP was not run for them. **Bold** marks the best result per column *within each group*.

Method	CUB-200-2011				Cars196				SOP			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@10	R@10^2	R@10^3
Trained Methods												
DCML (Zheng et al., 2021a)	68.4	77.9	86.1	91.7	85.2	91.8	96.0	98.0	79.8	90.8	95.8	–
DRML (Zheng et al., 2021b)	68.7	78.6	86.3	91.6	86.9	92.1	95.2	97.4	79.9	90.7	96.1	–
DAS (Liu et al., 2022)	69.2	79.3	87.1	92.6	87.8	93.2	96.0	97.9	80.6	91.8	96.7	99.0
CEP (Boudiaf et al., 2020)	69.2	79.2	86.9	91.6	89.3	93.6	96.6	98.4	81.1	91.7	96.3	98.8
IBC (Seidenschwarz et al., 2021)	70.3	80.3	87.6	92.7	88.1	93.3	96.2	98.2	81.4	91.3	95.6	–
Proxy Anchor (Kim et al., 2020)	71.1	80.4	87.4	92.5	88.3	93.1	95.7	97.5	80.3	91.4	96.4	98.7
HIST (Lim et al., 2022)	71.4	81.1	88.1	–	89.6	93.9	96.4	–	81.4	92.0	96.7	–
ProxyNCA++ (Teh et al., 2020)	72.2	82.0	89.2	93.5	90.1	94.5	97.0	98.4	81.4	92.4	96.6	99.0
IRT_R (El-Nouby et al., 2021)	76.6	85.0	91.1	94.3	–	–	–	–	84.2	93.7	97.3	99.1
DFML (Wang et al., 2023a)	79.1	86.8	–	–	89.5	93.9	–	–	–	–	–	–
DFML-PA (Wang et al., 2023a)	71.9	81.3	88.0	92.8	–	–	–	–	84.2	93.8	–	–
PLEor (Wang et al., 2023b)	74.8	84.5	91.3	94.9	**94.4**	**96.9**	**98.3**	**98.9**	–	–	–	–
CFGA (Yu et al., 2024)	72.2	81.2	87.9	92.7	87.5	92.2	95.1	97.1	–	–	–	–
FN-NET (Ye et al., 2025)	77.4	85.2	90.5	92.1	88.4	91.3	94.7	97.0	80.6	90.2	96.4	97.3
VCE (Wang et al., 2025)	**81.8**	**89.1**	**93.6**	**96.3**	93.6	96.2	98.0	98.7	**85.7**	**94.9**	**98.0**	**99.4**
Training-Free Methods												
IRT_O (El-Nouby et al., 2021)	58.5	–	–	–	–	–	–	–	**52.8**	–	–	–
CLIP ViT-L/14 (Radford et al., 2021) (image)	74.4	**84.0**	**90.9**	**95.3**	87.7	94.0	97.3	98.9	–	–	–	–
(image+text)	58.3	71.9	82.6	90.2	84.8	92.1	96.5	98.8	–	–	–	–
Qwen3-VL-Emb. (Qwen Team, 2025) (image)	38.0	49.5	61.5	72.8	63.6	75.1	84.5	91.8	–	–	–	–
(image+text)	35.9	46.9	58.3	70.2	81.5	90.2	95.6	98.1	–	–	–	–
AMUSE (Ours)	**74.8**	82.4	87.2	90.9	**96.3**	**97.9**	**98.7**	**99.0**	50.1	**73.4**	**88.5**	**96.6**

Table 2: Schema vs. free-form caption ablation on fine-grained image retrieval (%). All conditions use GPT-5 for proxy generation and text-embedding-3-large for embedding. Best per column in bold; ties bolded together.

Method	CUB-200-2011				Cars196			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
Free-form caption	55.50	65.99	75.74	83.86	81.40	88.66	93.90	97.07
Free-form w/ instruction	67.99	78.63	86.11	**91.81**	95.46	**97.92**	**98.78**	**99.20**
Schema (AMUSE)	**74.75**	**82.44**	**87.24**	90.92	**96.26**	**97.92**	98.67	99.03

4.3 Field-Level Schema Ablation

Setup. Section 4.2 establishes that structured schemas outperform free-form captions, but does not isolate *which* schema fields are doing the work. To answer this question, we run a field-level ablation on Cars196 – the benchmark where schemas yield the strongest results in Table 1. We partition the auto-generated Cars196 schema (Table 11) into five field groups: *Vehicle Identification* (make, model, year, color); *Visual Composition* (orientation, background, composition style, dominant features); *Image Quality* (texture, lighting, production quality); *Affective Response* (first impression, emotional response, psychological impact); and *Context & Uniqueness* (intended audience, narrative context, unique aspect, interaction experience, subtext). We perform two complementary ablations on the test set: (i) a *drop-one* sweep that removes a single group while keeping the other four; and (ii) a *cumulative add-one* sweep that starts from *Vehicle Identification* alone and grows the schema by one group at a time until reaching the full five-group configuration. All runs use GPT-5 + text-embedding-3-large with the same retrieval protocol as Table 1.

Results. Table 3 reports the cumulative add-one trajectory. *Vehicle Identification* alone already reaches 95.73 R@1; adding *Visual Composition* produces an 8-field schema at 96.46 R@1, slightly exceeding the full 18-field schema; and each subsequent group (Image Quality, Affective Response, Context & Uniqueness) holds R@1 within $\pm$0.20 pp of this two-group peak. A complementary drop-one sweep (not tabulated) corroborates the same asymmetry from the other direction: removing *Vehicle Identification* from the full schema collapses R@1 by 9.86 pp (96.26 $\rightarrow$ 86.40), while dropping any of the other four groups changes R@1 by less than 0.1 pp. Taken together, the schema's contribution on fine-grained Cars196 retrieval is

concentrated in a small set of domain-core identity attributes; affective, contextual, and quality descriptors function largely as filler.

4.4 Text and Document Retrieval

Setup. FiQA (Maia et al., 2018; Thakur et al., 2021) (short financial Q&A, 57,638 docs, 648 test queries) and LoCo v1 (Saad-Falcon et al., 2024) (long-context, 4k–32k tokens) together cover the short- and long-document regimes. Both queries and documents are converted to schematized proxies before embedding. We sweep the 3×2 grid of MLLMs and embedders and report NDCG@10 alongside Recall@100.

Results. Table 4 compares AMUSE to established retrieval baselines on FiQA. Baseline results are reproduced from the Pyserini BEIR benchmark (Kamalloo et al., 2024). AMUSE with GPT-5 + text-embedding-3-large outperforms BM25 by nearly 21 pp in NDCG@10 and surpasses Contriever by 11.3 pp. AMUSE also exceeds BGE-large by 3.6 pp in NDCG@10 and achieves the highest Recall@100 among all methods. Notably, direct embedding of raw text with text-embedding-3-large yields higher NDCG@10 scores, indicating that schematization is more beneficial when bridging modalities than when operating within text. The ablation study in Table 5 shows that both the MLLM and embedding model contribute to performance, with GPT-5 providing 8–12 pp gains over GPT-4o.

Table 4: Comparison with retrieval baselines on the FiQA benchmark (%).

Method	NDCG@10	R@100
BM25 (Thakur et al., 2021)	23.60	53.90
Contriever (Izacard et al., 2022)	32.90	65.60
SPLADE++ (Formal et al., 2022)	34.70	63.10
BGE-large (Xiao et al., 2024)	40.60	74.20
text-embedding-3-small[1]	44.91	–
text-embedding-3-large[1]	**55.00**	–
AMUSE (Ours)	**44.22**	**78.17**

[1]MTEB Leaderboard (https://huggingface.co/spaces/mteb/leaderboard), accessed 2026-01-21.

Table 5: AMUSE ablation on FiQA: effect of MLLM and embedding model (%).

MLLM	Embedding	NDCG@10	R@100
GPT-4o-mini	text-emb-3-small	29.19	66.14
GPT-4o-mini	text-emb-3-large	30.91	66.78
GPT-4o	text-emb-3-small	30.72	66.07
GPT-4o	text-emb-3-large	33.89	68.85
GPT-5	text-emb-3-small	41.41	77.43
GPT-5	text-emb-3-large	**44.22**	**78.17**

Long-context retrieval baselines. Table 6 presents performance of various retrieval models on LoCo v1 (Saad-Falcon et al., 2024), a benchmark designed to evaluate long-context retrieval capabilities. LoCo v1 tests retrieval over documents ranging from 4k to 32k tokens, making it particularly relevant for assessing models that can handle extended sequences without chunking.

Table 6: Retrieval model performance on LoCo v1 (NDCG@10, %). Baseline results from Saad-Falcon et al. (2024).

Method	NDCG@10
Trained Methods (Finetuned on LoCo v1)	
BGE-Large Finetuned	64.80
M2-BERT-128	70.30
M2-BERT-2k	82.30
M2-BERT-8k	86.90
M2-BERT-32k	**94.70**
Training-Free Methods (Zero-Shot)	
ColBERTv2	53.60
BGE-Large Zeroshot	56.50
OpenAI Ada	63.20
Jina Embeds.	67.20
E5-Mistral	71.40
BM25	**79.90**
AMUSE (Ours)	72.29

Among training-free methods, AMUSE (GPT-5 + text-embedding-3-large, 72.29 NDCG@10) is the strongest neural embedding approach, surpassing E5-Mistral by 0.89 pp, Jina Embeddings by 5.09 pp, and even the LoCo-finetuned BGE-Large (64.80). It trails BM25 (79.90) and the LoCo-finetuned M2-BERT family (up to 94.70 at 32k context), suggesting that schematized proxies compress long-document information differently than lexical matching or architectures with explicit long-context support.

4.5 Text-to-Audio Retrieval

Setup. We retrieve audio items given text queries on DCASE 2024 Task 8 (single-positive, 1k items) and DCASE 2025 Task 6 (multi-positive, 1k items), both built on the Clotho audio-captioning dataset (Drossos et al., 2020). Proxies are generated by Gemini 2.5 / 3 Flash and embedded with

Table 3: Cumulative add-one field-level schema ablation on Cars196 (%). All runs use GPT-5 + text-embedding-3-large. Field groups are added one at a time starting from *Vehicle Identification*; the full schema corresponds to all five groups defined in Table 11. ΔR@8 is the change in R@8 relative to the full schema; **bold** marks the best entry in each Recall column.

Configuration	R@1	R@2	R@4	R@8	ΔR@8
Full schema (baseline)	96.26	**97.92**	**98.67**	**99.03**	–
Vehicle Identification (1 group)	95.73	97.44	98.32	98.84	−0.19
+ Visual Composition (2 groups)	**96.46**	97.90	98.51	98.82	−0.21
+ Image Quality (3 groups)	96.26	97.88	98.47	98.81	−0.22
+ Affective Response (4 groups)	96.34	**97.92**	98.59	98.93	−0.10
Adding the final group (Context & Uniqueness) yields the full schema; see baseline row.					

text-embedding-3-large. We compare four schema configurations: a manually curated *Rich* schema (27 fields), a 5-field *Simple* schema, a *Simple + Metadata* variant that augments Simple with computed audio features (duration, tempo, loudness), and an *Auto-Generated* schema produced by the MLLM from a corpus sample. Gemini 3 Flash was released late in this study and is reported in the Auto-Generated row and on the 2025 benchmark only.

Results (DCASE 2024). Tables 7 and 8 summarize the experiments on the 2024 benchmark. We report mAP@10 as the primary metric. Our results indicate that upgrading the underlying MLLM to Gemini 3 Flash with an auto-generated schema yields a significant boost, reaching 17.75% mAP@10. Notably, adding computed low-level metadata degraded performance (14.60% mAP), suggesting that mixing raw signal features with semantic descriptions may introduce noise into the embedding space. While these results trail the state-of-the-art ensemble systems (Primus and Widmer, 2024) which achieve up to 41.9% mAP@10 through extensive fine-tuning on large-scale paired datasets, AMUSE demonstrates competitive zero-shot performance against the standard baseline without any task-specific training.

Table 7: AMUSE schema ablation on DCASE 2024 Task 8 (%). mAP@10 is the official metric. G2.5 / G3 denote Gemini 2.5 Flash / Gemini 3 Flash.

Configuration	R@1	R@5	R@10	mAP@10
Rich-27 (G2.5)	9.90	23.00	31.70	15.60
Simple-5 (G2.5)	9.60	23.10	32.40	15.50
Simple+Meta (G2.5)	9.40	21.50	29.40	14.60
Auto-Gen (G2.5)	9.90	23.00	31.70	15.60
Auto-Gen (G3)	**10.80**	**25.90**	**36.90**	**17.75**

Table 8: Comparison with baselines on DCASE 2024 Task 8 (%). mAP@10 is the official metric. SOTA refers to Primus and Widmer (2024).

Method	R@1	R@5	R@10	mAP@10
SOTA	**30.70**	**56.30**	**68.60**	**41.60**
DCASE baseline	13.00	34.30	48.00	22.20
AMUSE (Ours)	10.80	25.90	36.90	17.75

Results (DCASE 2025). Table 9 reports results on the multi-positive 2025 evaluation set, whose official metric is mAP@16. The best AMUSE configuration reaches 19.7 mAP@16, against a baseline of 33.0 and a SOTA of 42.1 (Primus and Widmer, 2024). The training-free approach still captures meaningful semantic signal in the harder multi-positive setting, but the gap to fine-tuned ensembles widens.

Table 9: Comparison with baselines on DCASE 2025 Task 6 (%). mAP@16 is the official metric for the multi-positive setting. AMUSE uses the same Auto-Gen schema with Gemini 3 Flash reported in Table 7.

Method	mAP@16
SOTA (Primus and Widmer, 2024)	**42.10**
DCASE 2025 baseline	33.00
AMUSE (Ours)	19.70

Baseline context. The DCASE Task 6B/8 baseline lineage (Xie et al., 2023) pairs a PANNs (Kong et al., 2020) audio encoder with a frozen Sentence-BERT (Reimers and Gurevych, 2019) text encoder; top-ranked 2024 systems (Primus and Widmer, 2024; Kulik et al., 2024) ensemble specialized audio encoders trained on AudioCaps/WavCaps with strong text embeddings.

5 Discussion

Schemas are a primary lever, but not always the dominant one. On fine-grained image retrieval the schema dominates: Cars196 reaches 96.3 R@1 without task-specific training, and Sections 4.2–4.3 show that schema choice moves R@1

more than the underlying MLLM choice. Elsewhere the picture is more mixed. On DCASE audio (Table 7), the 5-field and 27-field schemas land within ±0.70 pp R@10 of each other, while a Gemini 2.5 Flash → Gemini 3 Flash upgrade adds 2.15 pp mAP@10. On FiQA (Table 5), stepping from GPT-4o to GPT-5 with the schema fixed lifts NDCG@10 by 10.33 pp – a larger swing than any single schema intervention we measure. Schemas dominate when the retrieval signal is attribute-dense and the raw input is structurally impoverished (fine-grained images); the MLLM matters as much or more when the input is text or longer audio.

Why and where schemas help. The free-form ablation (Section 4.2) decomposes the schema's advantage into *task awareness* (what attributes to look at) and *structural canonicalization* (consistent vocabulary across the corpus). On CUB the schema retains a 6.8 pp R@1 lead over instruction-guided free-form text, indicating that canonicalization carries the load when small lexical drift between near-identical items dominates nearest-neighbor ranking. The field-level ablation (Section 4.3) localizes that signal: on Cars196 the four *Vehicle Identification* attributes account for nearly all of R@1, and a pruned 8-field schema slightly *exceeds* the full 18-field one. Schemas work because they canonicalize the *right* attributes, not because they impose structure per se.

Interpretability, adaptability, and cost. Decoupling modality understanding from retrieval gives AMUSE two underappreciated properties: every retrieval decision is traceable to a small set of human-readable attributes, and adapting to a new domain or modality requires only schema-level edits rather than data curation and retraining. These benefits come at the cost of one MLLM call and one embedding call per item; indexing Cars196 (16,185 images) via Azure's Batch API costs ∼$15–20 and runs in 4–6 hours, with query-time latency adding 1–2 s.

When schemas hurt. The FiQA result exposes a clear boundary: when inputs are already well-structured text, direct embedding with text-embedding-3-large reaches 55.00 NDCG@10 while AMUSE's schematized variant lags at 41.41 – a 13.6 pp gap. Schematization is most useful when it *adds* structure to a structurally impoverished input (images, audio), not when it re-paraphrases content that already has structure; we therefore recommend AMUSE for non-text modalities, noisy text, and interpretable-retrieval applications, and direct embedding for clean, well-formed text.

6 Conclusion

AMUSE translates inputs of heterogeneous modality into schema-driven textual proxies and embeds them with off-the-shelf text encoders, requiring no task-specific training; it sets a new state of the art on Cars196 (96.3 R@1) and reaches viable zero-shot performance on DCASE audio. Field-level ablation shows that on fine-grained image retrieval the schema is the dominant lever and concentrates on a small set of domain-core attributes, while on text and audio the underlying MLLM also matters; automated schema optimization and mixed-modality corpora are the most promising directions for future work.

Limitations

Several limitations warrant discussion. First, our experiments use only commercial MLLMs (GPT-4o family); evaluating open-source alternatives such as LLaVA (Liu et al., 2024), Qwen-VL (Bai et al., 2023), or InternVL (Chen et al., 2024) remains future work. These models may enable lower-cost or fully local deployments, though potentially with different quality trade-offs. Second, the current evaluation focuses on single-modality corpora; extending to truly mixed-modality collections where queries and documents span different modalities remains an open challenge. Third, while schematized proxies enhance interpretability, they may lose fine-grained details that continuous embeddings retain; for instance, subtle visual textures or timbral nuances may be difficult to verbalize. Fourth, the reliance on commercial MLLMs introduces cost and latency considerations for large-scale indexing, though this can be mitigated with caching, batching, or distillation to smaller models. Finally, the schema generation step, while largely automated, still benefits from domain expertise to ensure that the most discriminative attributes are captured.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025.

Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. 2020. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In *Computer Vision – ECCV 2020*, pages 548–564, Cham. Springer International Publishing.

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

Xin Cong, Bowen Yu, Mengcheng Fang, Tingwen Liu, Haiyang Yu, Zhongkai Hu, Fei Huang, Yongbin Li, and Bin Wang. 2023. Universal information extraction with meta-pretrained self-retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4084–4100, Toronto, Canada. Association for Computational Linguistics.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Nazarii Drushchak, Nataliya Polyakovska, Maryna Bautina, Taras Semenchenko, Jakub Koscielecki, Wojciech Sykala, and Michal Wegrzynowski. 2025. Multimodal retrieval-augmented generation: Unified information processing across text, image, table, and video modalities. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 59–64, Vienna, Austria. Association for Computational Linguistics.

Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. 2021. Training vision transformers for image retrieval. *Preprint*, arXiv:2102.05644.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2353–2359, Madrid, Spain. ACM.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, Vancouver, BC, Canada. IEEE.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980, Virtual and Singapore. IEEE.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 26574–26585, Seattle, WA, USA. IEEE.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4904–4916, Virtual Event. PMLR.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations (ICLR)*.

Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2024. Resources for brewing BEIR: Reproducible reference models and statistical analyses. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2948–2952, Washington, DC, USA. ACM.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. 2020. Proxy anchor loss for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3235–3244, Seattle, WA, USA. Computer Vision Foundation / IEEE.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating images with multimodal language models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy

Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348, Virtual. PMLR.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Victoria W., Fuzheng Zhang, and Guorui Zhou. 2025. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *Preprint*, arXiv:2505.19650.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, Sydney, Australia. IEEE Computer Society.

Jan Kulik, Bartlomiej Zgorzynski, Juliusz Kruk, Ivan Ryzhankow, Anna Ples, and Theodore Lamort de Gail. 2024. Take it for granted: Improving language-based audio retrieval with large language models. Technical report, DCASE 2024 Challenge.

Jungsoo Lee, Janghoon Cho, Hyojin Park, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. 2025. Generalized contrastive learning for universal multimodal retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jongin Lim, Sangdoo Yun, Seulki Park, and Jin Young Choi. 2022. Hypergraph-induced semantic tuplet loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 212–222, New Orleans, LA, USA. IEEE.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations (ICLR)*, Singapore. OpenReview.net.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, New Orleans, LA, USA. Curran Associates, Inc.

Lizhao Liu, Shangxin Huang, Zhuangwei Zhuang, Ran Yang, Mingkui Tan, and Yaowei Wang. 2022. Das: Densely-anchored sampling for deep metric learning. In *European Conference on Computer Vision*, pages 399–417, Cham. Springer.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2023. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations (ICLR)*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion of the The Web Conference 2018*, pages 1725–1734, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. 2024. Ret-llm: Towards a general read-write memory for large language models. *Preprint*, arXiv:2305.14322.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. Label-free concept bottleneck models. In *International Conference on Learning Representations*, Kigali, Rwanda. OpenReview.net.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025. GPT-5 system card. Accessed: 2025-01-17.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations (ICLR)*.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2023. Concept-based explainable artificial intelligence: A survey.

Paul Primus and Gerhard Widmer. 2024. A knowledge distillation approach to improving language-based audio retrieval models. Technical report, DCASE 2024 Challenge.

Qwen Team. 2025. Qwen3-VL-Embedding: Multimodal embedding models. `https://huggingface.co/Qwen`. Qwen3-VL-Embedding 8B model release.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, Virtual Event. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha, and Christopher Ré. 2024. Benchmarking and building long-context retrieval models with LoCo and M2-BERT. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42918–42946. PMLR.

Jenny Denise Seidenschwarz, Ismail Elezi, and Laura Leal-Taixé. 2021. Learning intra-batch connections for deep metric learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9410–9421, Virtual Event. PMLR.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629, New Orleans, LA, USA. IEEE.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, Las Vegas, NV, USA. IEEE Computer Society.

Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. 2020. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *Computer Vision – ECCV 2020*, pages 448–464, Cham. Springer International Publishing.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, Virtual Event. NeurIPS.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Chengkun Wang, Wenzhao Zheng, Junlong Li, Jie Zhou, and Jiwen Lu. 2023a. Deep factorized metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7672–7682, Vancouver, BC, Canada. IEEE.

Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. 2023b. Open-set fine-grained retrieval via prompting vision-language evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19381–19391, Vancouver, BC, Canada. IEEE.

Yuetian Wang, Shuo Ye, Wenjin Hou, and Shuhuang Chen. 2025. Vce: Visual concept embedding for open-set fine-grained image retrieval. *Knowledge-Based Systems*, 329:114311.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2024. C-Pack: Packaged resources to advance general chinese embedding. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–650, Washington, DC, USA. ACM.

Huang Xie, Samuel Lipping, and Tuomas Virtanen. 2023. Baseline system for language-based audio retrieval (task 6b) in dcase 2023 challenge. `https://github.com/xieh97/dcase2023-audio-retrieval`.

Shuo Ye, Qinmu Peng, Yiu ming Cheung, Yu Wang, Ziqian Zou, and Xinge You. 2025. Fn-net: Adaptive data augmentation network for fine-grained visual categorization. *Pattern Recognition*, 165:111618.

Han Yu, Huibin Lu, Min Zhao, Zhuoyi Li, and Guanghua Gu. 2024. Gradient aggregation based fine-grained image retrieval: A unified viewpoint for cnn and transformer. *Pattern Recognition*, 149:110248.

Mert Yuksekgonul, Maggie Wang, and James Zou. 2022. Post-hoc concept bottleneck models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, Online. OpenReview.net.

Delong Zeng, Yuexiang Xie, Yaliang Li, and Ying Shen. 2025. Enhancing multimodal retrieval via complementary information extraction and alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22092–22105, Vienna, Austria. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9274–9285, Nashville, TN, USA. Computer Vision Foundation / IEEE.

Wenzhao Zheng, Chengkun Wang, Jiwen Lu, and Jie Zhou. 2021a. Deep compositional metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9320–9329, Virtual Event. Computer Vision Foundation / IEEE.

Wenzhao Zheng, Borui Zhang, Jiwen Lu, and Jie Zhou. 2021b. Deep relational metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12045–12054, Montreal, QC, Canada. IEEE.

A Schema Definitions

Tables 10 and 11 present schemas auto-generated by the AMUSE pipeline (Figure 1) for text and image retrieval, respectively. The schema generation

phase produces domain-appropriate fields based on the dataset description and representative samples.

B Prompts and Validation Pipeline

This appendix details the prompts and validation pipeline used by AMUSE: the one-shot schema-generation step, the proxy-generation prompts compared in the schema vs. free-form ablation (Section 4.2), and how returned outputs are validated. All proxy-generation conditions share an "expert content analyst" framing and the same model (GPT-5); they differ only in whether the prompt requests structured output and whether dataset-level task context is provided. Full templates are released with the code.

Schema generation. A schema is produced once per dataset and then frozen. The MLLM is fed (a) a fixed meta-prompt that asks it to emit a Pydantic-style schema with named fields, primitive types, and one-line field descriptions; (b) a short user-provided dataset description (name, modality, purpose; see Section 3); and (c) 3–5 representative items each from the corpus and query splits, hand-selected to span typical variation (e.g., for Cars196: different makes, orientations, and backgrounds). The selected sample paths are recorded in the experiment configuration so the same schema is reproducible across runs. The resulting schema is committed as a Pydantic class (Tables 10 and 11) and reused unchanged for every subsequent proxy-generation call; we did not iteratively refine it across runs.

Schema (AMUSE). The system message names the modality, points to the target schema, and instructs the model to be specific, factual, and comprehensive when filling every field. The user message issues an image-analysis request asking for a JSON response matching the schema in Table 11:

> *Please analyze this image (a CORPUS item) and provide a comprehensive structured description according to the provided schema. Extract detailed information for each schema field based on your analysis . . . enabling semantic similarity matching for information retrieval.*

Free-form caption. Same expert framing, but the system message asks for "detailed, free-form captions" rather than a schema, and the user message carries no dataset-specific context:

> *Please provide a comprehensive caption for this content, capturing all relevant details and characteristics in free-form prose . . . enabling semantic similarity matching for information retrieval.*

Free-form w/ instruction. Identical to the free-form caption with one addition: the dataset name, purpose, and description are appended to the user message. For Cars196 this appendix reads:

> *Dataset context: You are generating captions for Stanford Cars (Cars196) Dataset (fine-grained vehicle similarity and retrieval). [The standard Cars196 description: ∼16K images of 196 classes defined by Make, Model, and Year; the goal is to learn embeddings that group same-model images while separating visually similar but different models or year-variants.]*

This single block of dataset context is the only difference between the two free-form conditions, and is what we call "task instruction" in Section 4.2.

Validation and retry. Proxy generation runs in OpenAI's JSON mode (`response_format={type:"json_object"}`) for GPT models and the equivalent JSON output mode of Gemini for audio. Both guarantee well-formed JSON but not schema conformance, so every returned proxy is validated against the frozen Pydantic schema before embedding; items that fail validation are re-submitted as single-call requests outside the batch (the batch API itself additionally retries failed individual requests up to three times by default). We do not perform any post-hoc filtering, normalization, or rewriting of validated proxies before embedding.

C Sample Proxies

To make the qualitative difference between the three proxy-generation conditions concrete, Table 12 reproduces the proxies generated for a single Cars196 corpus image (ID 75, a front-on view of a 2012 Ferrari California convertible parked on an urban street). Texts are verbatim GPT-5 outputs, lightly truncated where indicated.

All three proxies correctly identify the make and model. The free-form variants differ subtly in confidence—the unguided caption hedges ("likely a California") while the instructed version commits without hedging—reflecting how dataset context (Appendix B) primes the MLLM toward task-relevant cues. The schema variant exposes the identifying attributes as discrete fields (`make`, `model`, `year`, `color`), which keeps them dominant in the embedded vector and limits dilution by background and affective descriptors. This is consistent with the field-level ablation in Table 3, where Vehicle Identification alone preserves nearly all of the retrieval signal on Cars196.

Table 10: Auto-generated schema for LoCo v1 text retrieval. All fields are strings or string lists; *optional.

Field	Description
Content & First Impressions	
title*, summary*	Document title and concise overview
first_impression	Initial emotional response or dominant features
dominant_features	Key characteristics (tone, style, phrases)
Affective & Sensory	
emotional_response	Emotional journey, mood, atmosphere
sensory_experience	Textures, rhythms, tones perceived
psychological_impact	Effects on reader, memories triggered
Structure & Quality	
structure	Composition, flow, clarity
quality_indicators	Coherence, depth, terminology use
style_choices	Language style, rhetorical devices
Context & Purpose	
cultural_context	Cultural/situational context
intended_audience	Target audience
purpose	Primary goal (inform, persuade, question)
subtext*	Hidden meanings or implications
Retrieval-Specific	
interaction_experience	Reader engagement, memorability
uniqueness	Distinguishing characteristics
retrieval_relevance	Importance for ranking
aspect_discussed	Specific topics/entities covered
sentiment_polarity	Positive, negative, or neutral
opinion_holder*	Entity expressing the opinion

Table 11: Auto-generated schema for Cars196 image retrieval. All fields are strings unless noted; *optional.

Field	Description
Vehicle Identification	
make, model	Car manufacturer and model name
year (int)	Model year
color	Dominant color of the vehicle
Visual Composition	
orientation	Car orientation (side, front, rear, etc.)
background	Description of the background setting
composition_style	Artistic style (symmetrical, dynamic, etc.)
dominant_features	Key visual features that stand out
Image Quality	
texture_quality	Surface texture (smooth, metallic, matte)
lighting_quality	Lighting type (soft, harsh, natural)
production_quality	Overall image production quality
Affective Response	
first_impression	Initial emotional response to image
emotional_response	Mood or atmosphere evoked
psychological_impact	Feelings or memories triggered
Context & Uniqueness	
intended_audience	Target audience for the image
narrative_context	Story or context behind the image
unique_aspect	What makes this image memorable
interaction_experience	How viewers engage with content
subtext*	Hidden meanings in the image

Table 12: Proxies generated for Cars196 corpus image ID 75 under the three conditions of Section 4.2. Each cell is the verbatim model output (light truncation marked "…"). The schema row lists a representative subset of fields; the remaining 11 affective/context fields are omitted for space.

Condition	**Generated proxy (Cars196 ID 75)**
Schema (AMUSE)	`make`: Ferrari. `model`: California (Convertible). `year`: 2012. `color`: bright red (Rosso-like). `orientation`: front, head-on. `dominant_features`: [Prancing horse emblem centered in a black, oval/rounded grille; smooth, glossy red paint with strong reflections; long, sculpted hood with a central intake/scoop; rounded, teardrop headlights; curvy front fenders and a low, wide stance; convertible with the top down and tan/beige interior visible; …]. `background`: Urban street scene with a wrought-iron fence and greenery; sidewalk curbs and bollards/fire hydrant nearby; muted buildings beyond; orange directional signs further back; a pedestrian partially visible at the left edge …. `first_impression`: A vivid red Ferrari commands the street, low and poised, projecting confidence and luxury. *(11 affective/context fields omitted.)*
Free-form caption	A bright red Ferrari convertible—likely a California—fills the frame in a front-on, street-level view, its low, sculpted nose and central hood vent catching soft daylight. The glossy paint throws smooth reflections of trees and buildings, while teardrop-shaped headlights and a wide oval black grille with the prancing horse emblem give the car a poised, predatory look. The top is down, revealing a tan leather interior behind the raked windshield. … A passerby blurs along the left edge, underscoring motion and scale. The overall mood is upscale and cinematic: an exotic Italian sports car momentarily at rest amid everyday city textures.
Free-form w/ instruction	Low, wide, and unmistakably exotic, a bright red Ferrari California convertible fills the frame in a straight-on, eye-level view. The soft daylight glints across its glossy paint, highlighting the car's rounded, muscular front fenders and the clean, sculpted lines of the hood with a prominent central scoop near the leading edge. … With the top down, the tan leather interior is visible through the raked windshield, adding a warm contrast to the vivid red exterior. The car is parked at the curb on an urban street … a wrought-iron fence, greenery, and yellow directional road signs hint at a city intersection.