MULTIMODAL DATA CURATION THROUGH RANKED RETRIEVAL

Pratyush Muthukumar, Harshil Kotamreddy, Sarah Amiraslani, Tomo Kanazawa, Ramani Akkati, Shaan Jain, & Andrew Mathau
NVIDIA
{pmuthukumar, hkotamreddy, samiraslani, tkanazawa, rakkati, shaanj, amathau}@nvidia.com

ABSTRACT

Shared embedding spaces are widely used for multimodal search and data curation. In practice, two problems often limit how well this works. First, embeddings can reflect modality more than meaning, so examples cluster by input type even when the underlying content matches. Second, the paired supervision used to train these spaces is often noisy. When we blend many heterogeneous, human-labeled datasets, these issues reinforce each other and degrade cross-modal retrieval. We present a framework that improves alignment by acting on both the training pairs and the embedding model. Symmetric Nucleus Subsampling (SNS) refines training pairs by trimming raw inputs and annotations to the portions that best support each other. Expert Embedding Engine (EEE) combines complementary embedding experts using a learned projection network, together with a bias-aware objective that reduces modality-driven separation in the embedding space. We demonstrate that this approach collapses the modality gap by over 90% on average vs base embedding experts and is a strong data curator, with datablends from our method outperforming stratified sampling and traditional curation baselines in downstream model performance.

1 INTRODUCTION

Consider a user searching a large archive that mixes images, clips, transcripts, and audio recordings. An ideal state would be a single interface where a short query reliably surfaces the right content, regardless of whether the answer lives in a frame, a caption, or a few seconds of sound. This expectation is a big reason shared-embedding approaches (popularized by contrastive vision–language training) have become the default backbone for multimodal search, recommendation, and dataset analytics. Increasingly, the same similarity-search loop is also used upstream to curate the training mixtures that produce these models.

This shift toward embedding-first systems also mirrors a broader trend in modern model training. For large language models and multimodal models alike, data quality and data composition increasingly determine what the model learns. When we scale up training, we rarely rely on a single clean dataset. We blend many sources, each with its own annotation style, noise patterns, and coverage. In that setting, curation becomes a first-class problem: which examples should we include, which should we downweight, and how do we combine datasets without washing out rare but valuable signal? Search-based retrieval is a practical way to curate across many corpora because it can surface semantically related examples, make mixtures easy to audit, and let us iterate quickly. At the same time, this raises the stakes for the embedding space itself, since retrieval will only be reliable if distance actually tracks meaning. That dependence on the embedding space hides a fragile assumption: that distance primarily reflects semantic similarity. In real deployments, the geometry often reflects something else just as strongly - modality identity. Text tends to cluster with text, images with images, and audio with audio, even when examples are describing the same underlying concept (Figure 1). This directly affects cross-modal retrieval, where nearest neighbors may share modality rather than meaning. It also complicates fusion strategies that rely on consistent geometry and reduces the reliability of downstream embedding workflows like clustering, deduplication, and active data selection.

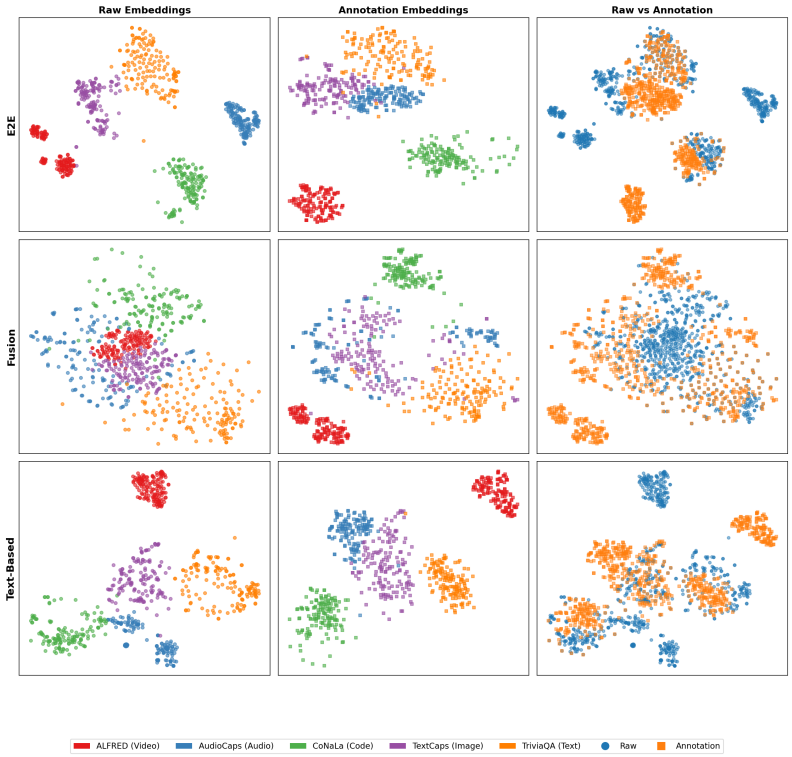


Figure 1: A 2D t-SNE visualization of paired data embeddings by common multimodal embedding expert implementations (text-based, fusion, end-to-end), illustrating modality-dependent clustering.

A second, independent issue is that the training signal used to create these embedding spaces is rarely clean. Paired datasets often exhibit raw–annotation misalignment: captions omit important objects or actions, transcripts include irrelevant context, and annotations can refer to content that is off-screen or inaudible. When supervision is only partially grounded, the model is effectively asked to align a raw sample with an annotation that shares limited information. In response, training can reward shortcuts - features that correlate with the dataset or modality rather than the described event - which in turn reinforces modality bias instead of correcting it.

Most existing solutions tackle one side of this problem at a time. Representation-level methods aim to reduce modality separation through calibration or alignment strategies, but typically accept the supervision signal as given (Long et al., 2024). Data-centric filtering and reweighting strategies try to improve pair quality, but often assume a fixed encoder family and do not explicitly address systematic modality-driven geometry (Zha et al., 2025). In settings where we want to merge datasets and learn a unified embedding space that supports robust search and analysis, these issues are difficult to decouple: noisy pairs make geometry worse, and biased geometry makes it harder to detect which pairs are noisy.

We therefore take a coupled approach that intervenes on both the examples and the embedding model. On the data side, we introduce Symmetric Nucleus Subsampling (SNS). Rather than treating each paired sample as atomic, SNS identifies high-information nuclei on both sides of the pair by selecting the most relevant portions of the raw input with respect to the annotation, and the most grounded portions of the annotation with respect to the raw input, then retains content that remains consistent under a symmetric similarity check. The goal is not to "clean" the dataset into a different distribution, but to make each training pair carry a clearer alignment signal while preserving a safe fallback to the original pair when trimming does not improve alignment.

On the model side, we introduce an Expert Embedding Engine (EEE), a mixture-of-experts embedding architecture that combines embedding spaces through a learned projection mechanism. The experts are designed to capture different strengths (e.g., text-centric representations which are useful when strong text encoders are available), fusion-based representations (useful when jointly modeling

modalities), and modality-unified encoders so that the system can adapt to the input and annotation characteristics of each example. Training includes a bias-aware objective that discourages modality-driven separation.

We study this framework in practical retrieval and clustering settings over image–text, audio-text, video-text, and text–text collections. Our evaluation focuses not only on downstream metrics but also on embedding-space behavior, including modality-wise separation and clustering structure, and on how data-side subsampling and model-side projection interact.

Our contributions are as follows:

- We analyze how modality bias and raw–annotation misalignment jointly affect multimodal embedding quality, particularly in mixed, human-labeled corpora.
- We propose SNS, a symmetric, annotation-aware subsampling method that improves pair consistency by trimming irrelevant content on either side of a multimodal pair.
- We propose EEE, a mixture-of-experts embedding engine with a learned projection layer applying a bias-aware objective to reduce modality-driven separation while preserving semantic neighborhood structure.
- We design a robust evaluation study using a human-selected task set for balanced comparisons of downstream model performance impact by datablends curated through variants of our approach and industry standard baselines.

2 Preliminaries and Background

2.1 Paired multi-modal datasets

We define paired multi-modal datasets as a collection of paired samples $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}$ is a *raw data* view in one high-bandwidth modality (e.g., image, video, long text) and $y_i \in \mathcal{Y}$ is a supervisory description and/or label *annotation data* view in another modality (typically text). We embed both raw and annotation data as ℓ_2-normalized vectors for clustering, retrieval, and further manipulations.

2.2 Contrastive Learning (CLIP) and Mutual Information

Central to modern retrieval and transfer learning is the concept of dual-encoder setups, wherein encoders f_x, f_y map paired data (x_i, y_i) to embeddings $z\{x_i\}$, $z\{y_i\}$ in a shared embedding space with a training objective to increase similarity of matched pairs (x_i^+, y_i^+) and decreasing similarity of *in-batch negative* mismatched pairs (x_i^-, y_i^-) in shared space; a temperature parameter scales logits/similarities in this contrastive objective (Radford et al. (2021). Key to this formulation of contrastive learning is the assumption that paired samples (x_i, y_i) share more semantic content than mismatched pairs, so 'positive' (semantic) similarities should be systematically higher than 'negative' similarities (shared modality).

Conceptually, contrastive objectives are often framed as an attempt to maximize *mutual information* (MI) between paired data views (van den Oord et al., 2019). This relies on certain sampling assumptions that can degrade in high-dimensional spaces/with large MI (Poole et al., 2019), so we treat MI as an intuition and rely on similarity-based proxies rather than attempting to estimate it directly.

2.3 Failure modes: Modality gap and data misalignment

As alluded to above, multi-modal embedding models face a fundamental challenge: individual embedding functions can develop geometric separation, in which embeddings cluster by modality identity rather than by shared semantic content. This modality gap can harm semantic classification and retrieval, even when paired samples have high semantic similarity (Liang et al., 2022).

In addition to the modality gap, paired data also commonly contain other types of misalignment:

- **Extraneous/off-target annotation:** annotation contains irrelevant context not observable in or supported by the raw data view (e.g., off-topic tangents in image description).
- **Incomplete/under-descriptive annotation:** annotation incompletely spans salient raw data content; this can be spatial or temporal (e.g., in a 30-minute automotive dash camera video, only describe street signs or only describe the last 5 minutes of activity, respectively)

Misaligned sample pairs violate the assumption that matched pairs share key semantic content, thereby amplifying geometric artifacts such as modality gaps and further complicating cross-modal retrieval.

3 Related Work

As a result, data selection and filtering are widely recognized as critical in large-scale representation learning, particularly for contrastive pretraining where web-scale corpora contain noise, duplication, and weakly aligned pairs. Recent work studies practical filtering strategies and training refinements that improve contrastive vision language pretraining under noisy data (Radenovic et al., 2023). Dataset curation benchmarks and systematic comparisons of filtering, reweighting, and sampling strategies for training multimodal models highlight that quality and mixture choices can dominate downstream outcomes (Gadre et al., 2023). Additional discussion of related work can be found in Appendix A.7.

4 Methods

To rank relevant data samples across several human-labeled datasets and modalities, we devise a three-step approach:

- First, we use SNS to reduce misalignment within raw data and annotation pairs by removing irrelevant portions of the raw data and/or the annotation.
- Second, we use EEE to reduce modality-specific bias from any specific embedding model by using several embedding models with different multi-modal approaches.
- Finally, in order to rank embeddings generated by the EEE, we combine the embedding spaces of each of the experts using a projection network.

While we note the use of specific models in the first two steps, these models can be switched out for any other model that performs the same function. We intend for this method to be a generalizable framework without restriction on specific model usage.

4.1 Reducing Misalignment with Symmetric Nucleus Subsampling

Paired multimodal datasets often have raw data that contain information that is not included in the annotation/label, and vice versa. More concretely, for a paired data sample $(x, y) \in \mathcal{X} \times \mathcal{Y}$, misalignment occurs when $I(x; y) \ll \min(H(x), H(y))$, i.e., the mutual information I is much less than the entropy H of either component. We quantify this using the *information density*, ϕ, of a paired data sample, which is defined as the ratio of mutual information to the total content size:

$$\phi(x, y) = \frac{I(x; y)}{|x| + |y|}. \tag{1}$$

To increase ϕ for any given paired data sample, we propose extracting nuclei from both the raw data and the annotation. This is done through a forward extraction phase and a backward extraction phase. In forward extraction, we extract the most relevant part (nucleus) of the raw data based on the contents of the annotation. In backward extraction, we do the opposite, i.e., we extract the nucleus from the annotation based on the contents of the raw data.

In this work, we use the result of forward extraction (the nucleus of the raw data) when performing backward extraction to maintain semantic consistency between the raw data and the annotation. The overall architecture of the Symmetric Nucleus Subsampler component is visualized in Figure 2.

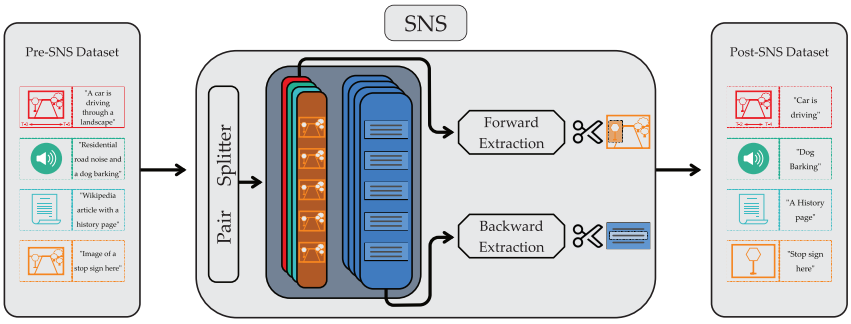


Figure 2: Symmetric Nucleus Subsampler component overview, including Forward Extraction and Backward Extraction modules for effective reduction of misalignment in paired samples

4.1.1 Forward Extraction

In the forward extraction phase, we address data-to-annotation misalignment by identifying portions of raw data that maximally explain the annotation content. Let $\mathcal{N}_\alpha : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ be the forward extraction function. $\mathcal{N}_\alpha$ outputs a subset of the raw data $\tilde{x}$ such that $I(\tilde{x}; y) > \rho \cdot I(x; y)$, where ρ is the mutual information ratio. If no such subset exists, $\mathcal{N}_\alpha$ simply outputs the original raw data.

In practice, $\mathcal{N}_\alpha$ is implemented using several modality-specific segmentation methods and mutual information is approximated using embedding similarity. We use Omni-Embed-Nemotron-3B (Xu et al., 2025b) to compute embeddings across all modalities. While this still induces modality-specific bias, it serves as a noisy proxy to mutual information (see Figure 10). The segmentation methods for each modality are as follows:

Text: We use Omni-Embed-Nemotron-3B (Xu et al., 2025b) to embed the annotation and each sentence of the corresponding raw data. We then calculate the cosine similarity between each sentence and the annotation and keep a sentence if the similarity is higher than τ_α.

Image: We use the Grounding-DINO (Liu et al., 2024) image segmentation model to create bounding boxes for the most relevant parts of the image based on the annotation. We keep the bounding boxes with confidence scores higher than τ_α. If there are multiple bounding boxes, we get the minimum spanning bounding box. We then extract the part of the image within the bounding box.

Video and Audio: We use moment detection models (CG-DETR Moon et al. (2023) for video and AM-DETR Munakata et al. (2025) for audio) to extract the most relevant time span within the video/audio data to the annotation. We keep the time spans with confidence scores higher than τ_α. If there are multiple time spans within the video that are relevant to the annotation, we remove all data between the extracted time spans and splice together the relevant portions.

In practice, τ_α can have different values for each modality; however, we present it as a single value for simplicity. Once a subset $\tilde{x}$ of the raw data x is extracted, we check that $I(\tilde{x}; y) > \rho \cdot I(x; y)$ before replacing x with $\tilde{x}$.

4.1.2 Backward Extraction

In the backward extraction phase, we address annotation-to-data misalignment by identifying portions of the annotation that maximally explain the content of the raw data. Let $\mathcal{N}_\beta : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ be the backward extraction function. $\mathcal{N}_\beta$ outputs a subset of the annotation $\tilde{y}$ such that $I(x; \tilde{y}) > \rho \cdot I(x; y)$ where ρ is the mutual information ratio. If no such subset exists, $\mathcal{N}_\beta$ simply outputs the original annotation. In practice, we implement $\mathcal{N}_\beta$ using several modality-to-text models by first converting the raw data to a description and then using embedding similarity to extract the most relevant sentences from the annotation. The models we use to describe the raw data are as follows:

Text: We simply use the text raw data with no further processing.

Image: We use the Nemotron-Nano-12B-v2-VL model (NVIDIA et al., 2025) to generate a description of the image.

Video: We use the Cosmos-Reason2-2B model (NVIDIA, 2025a) to generate a description of the video.

Audio: We use the Phi-4-multimodal-instruct model (Abouelenin et al., 2025)to generate a description of the audio.

To compute embeddings, we once again use Omni-Embed-Nemotron-3B (Xu et al., 2025b). We keep sentences from the annotation if their cosine similarity with the description generated is greater than some threshold τ_β. Finally, we replace the original annotation y with the extracted annotation $\tilde{y}$ if $I(x;\tilde{y}) > \rho \cdot I(x;y)$. We ablate values of τ_α, τ_β, and ρ in Appendix A.1.

4.2 Reducing Modality-Specific Bias with the Expert Embedding Engine

Once data pairs pass through the SNS process, they are more closely aligned. We now discuss embedding each of these data pairs in order to retrieve the most relevant data pairs to some query q. While we can simply use a single multi-modal embedding model, we find that these models tend to have modality gaps, i.e., data points of a certain modality tend to be close to each other (Figure 1).

To address modality gaps, we propose using several embedding models (experts) that compute embeddings differently. Each expert will tend to exhibit a different modality bias, giving us more information about the semantic meaning of each data pair. Specifically, we use the following embedding models:

End-to-End Expert: We use the Omni-Embed-Nemotron-3B multimodal embedding model (Xu et al., 2025b), which uses an end-to-end architecture to compute embeddings for all modalities.

Fusion Expert: We use the ImageBind fusion model (Girdhar et al., 2023), which uses an embedding fusion architecture to combine several embedding spaces while using images as an anchor.

Text Expert: We first convert data of all modalities to text using the same modality-to-text models used in Section 4.1.2. Then, we use the Llama-Nemotron-Embed-1B text embedding model (Moreira et al., 2024) to embed the text descriptions.

However, we now run into a different problem: how do we rank multiple distinct embedding spaces?

4.3 Combining Multiple Embedding Spaces With a Projection Network

We train a lightweight neural network to adaptively combine embeddings from all three expert embedding engines into a unified representation. Without this network, each embedding space returns results separately at query time, making it difficult to rank relevance across different experts.

The network takes the concatenated embeddings from all K experts as input (dimension $K \times d$) and outputs a single embedding vector of dimension d:

$$\mathbf{e}_{\text{fused}} = f_\theta([\mathbf{e}_1; \mathbf{e}_2; \ldots; \mathbf{e}_K]) \tag{2}$$

where f_θ is a multi-layer neural network and $[\cdot;\cdot]$ denotes concatenation.

There is one exception to this fusion process: we need a grounded set of embeddings that act as anchors, of which the network will learn the geometry of the embedding space and guide the data sample embeddings to their respective locations accordingly. We treat the *annotations* in the paired representation of data samples *(annotation, raw data)* as these anchors. Thus, raw data samples of varied modalities are passed through the projection layer to arrive at $\mathbf{e}_{\text{fused}}$, but annotation embeddings are left unchanged. In practice, since our dataset consisted of all text annotations, we opted to simply use the embeddings generated from the text-based expert as these anchor embeddings, as depicted in Figure 3.

4.3.1 Loss Function Formulation

The projection network is trained with two objectives: (1) preserving semantic similarity between paired data and annotations, and (2) minimizing the modality gap in the fused embedding space.

The total loss has three terms:

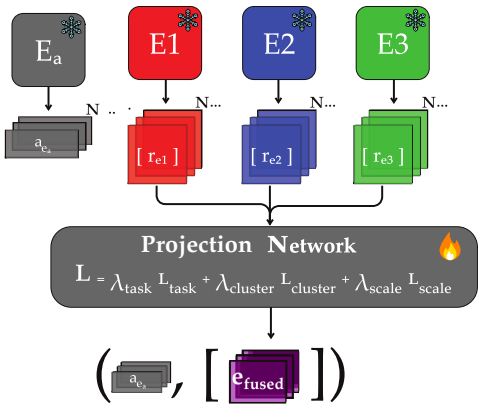


Figure 3: Expert Embedding Engine (EEE) & Projection Network components. The learned embedding space of paired samples can be used to curate datablends using embedding similarity to the query vector.

Task Loss: An InfoNCE-style contrastive loss that encourages positive pairs (data and its annotation) to be close while pushing negative pairs apart:

$$L_{\text{task}} = -\log \frac{\exp(\text{sim}(\mathbf{x}, \mathbf{x}^+)/\tau)}{\sum_{\mathbf{x}_i \in \mathcal{B} \setminus \{\mathbf{x}\}} \exp(\text{sim}(\mathbf{x}, \mathbf{x}_i)/\tau)} \tag{3}$$

where τ is the temperature parameter and the denominator sums over all in-batch samples except the anchor.

Cluster Bias: Penalizes the distance between each modality centroid and the overall centroid, encouraging different modalities to overlap in the embedding space:

$$L_{\text{cluster}} = \sum_{m \in \mathcal{M}} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}\|_2^2 \tag{4}$$

where $\boldsymbol{\mu}_m$ is the mean embedding of modality m, and $\boldsymbol{\mu}$ is the overall mean across all modalities $\mathcal{M}$.

Scale Bias: Penalizes differences in cluster spread between each modality and the overall distribution:

$$L_{\text{scale}} = \sum_{m \in \mathcal{M}} |\sigma_m - \sigma| \tag{5}$$

where σ_m is the average distance of modality m's embeddings from $\boldsymbol{\mu}_m$, and σ is the corresponding overall spread.

The total loss is:

$$L_{\text{total}} = \lambda_{\text{task}} L_{\text{task}} + \lambda_{\text{cluster}} L_{\text{cluster}} + \lambda_{\text{scale}} L_{\text{scale}} \tag{6}$$

We run a comprehensive ablation study to arrive at final $\lambda_{\text{task}}, \lambda_{\text{cluster}}$, and λ_{scale} terms as well as projection network architecture, details described in Appendix A.3.

Once embeddings are passed through the projection network, they are part of a unified embedding space and can now be used for retrieval. During query time, queries are passed through the text-based expert in a similar manner to the annotations used for grounding.

5 Results

5.1 Datasets

The candidate data pools comprises five human-annotated multimodal datasets spanning four modalities: **AudioCaps** (Kim et al., 2019) (audio captioning), **TextCaps** (Sidorov et al., 2020) (image captioning), **ALFRED** (Shridhar et al., 2020) (video instruction following), **CoNaLa** (Yin et al., 2018) (code captioning), and **TriviaQA** (Joshi et al., 2017) (text-text uestion answering). Each

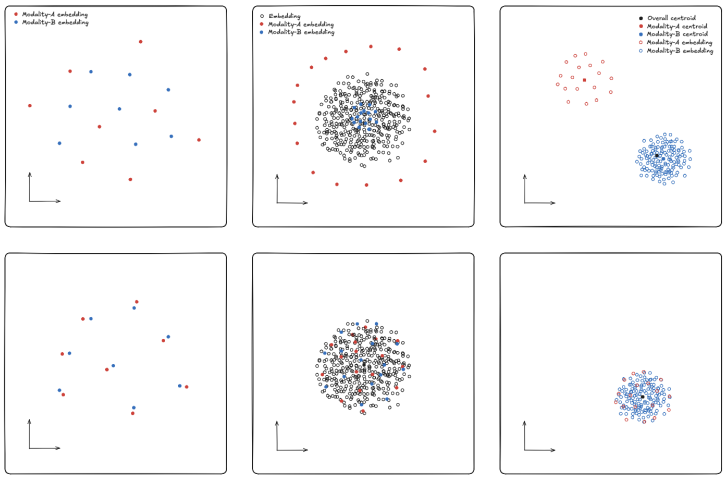


Figure 4: Effect of bias loss terms on modality gap reduction. Top row: Before training, embeddings from different modalities (A: red, B: blue) occupy distinct regions with separated centroids and varying cluster sizes. Bottom row: After training with cluster center bias ($L_{cluster}$) and scale bias (L_{scale}), modality embeddings overlap in a shared region, centroids align, and cluster spreads become uniform—enabling fair cross-modal retrieval. Appendix A.3.1 describes how this proposed effect is demonstrated in our empirical data.

dataset stores paired *(raw data, annotation)* entries, where the raw modality varies (audio, image, video, or text) and the annotation is fixed as natural-language text. From this combined pool of 10k candidates (2k per pool), each curation strategy selects a fixed blend of n=5,000 samples for downstream fine-tuning.

5.2 Downstream Evaluation Study

Each curated blend of 5,000 samples is converted into an instruction-tuning dataset by wrapping every *(raw data, annotation)* pair into a prompt–completion format. The model receives the raw media alongside a natural-language instruction (e.g., "*Describe the media content in detail*") and is trained to generate the human-written annotation as the target response. This transforms the curation problem into a direct measure of multimodal instruction-following quality—better-curated blends should yield more informative supervision and, consequently, lower validation perplexity.

We fine-tune **Qwen2.5-Omni-3B** (Xu et al., 2025a), a multimodal causal language model, using parameter-efficient LoRA adaptation (Hu et al., 2022) on the curated blends. Each configuration is trained for 750 steps and replicated three times to compute 95% confidence intervals. Validation perplexity serves as the primary metric, providing a task-agnostic signal of how well the fine-tuned model absorbs the curated data distribution without over-fitting to any single downstream benchmark.

Validation perplexity is computed on a static, human-curated held-out set of 500 pairs for all variants and baselines. To construct this set, we draw from an unseen partition of the same five data pools using a blind random shuffle, focusing on "natural, real-world scenes containing objects, landscapes, subjects, or people" – which is the text query we use to curate blends for the curation task. Two of the authors of this study independently reviewed the shuffled candidates, and each selected 250 pairs they judged to best exemplify the query, yielding a combined evaluation set of 500 pairs (223 audio-text, 204 image-text, 53 video-text, 20 text-text).

We evaluate the following variants of our proposed architecture in this downstream evaluation study:

- *EEE + Projection Only*: Expert Embedding Engine with learned projection; no SNS.
- *SNS Forward + EEE + Projection*: Adds forward Symmetric Nucleus Subsampling, which filters raw data conditioned on the annotation.
- *SNS Backward + EEE + Projection*: Uses backward SNS instead of forward SNS, extracting annotation-relevant substructures from the raw data.

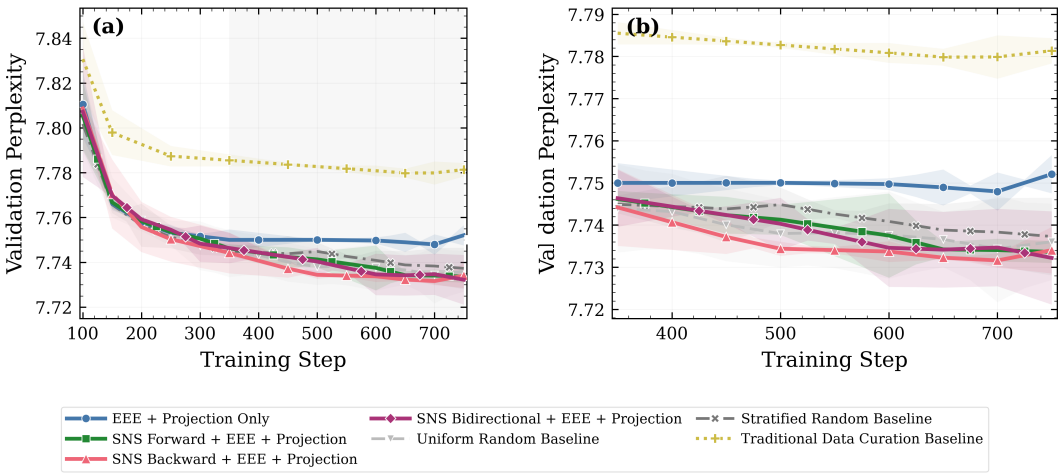


Figure 5: Validation perplexity (mean ± 95% CI, n=3) across SNS & EEE configuration variants vs baselines. **(a).** Full validation perplexity curves, **(b).** Last epoch validation perplexity curves.

- *SNS Bidirectional + EEE + Projection*: Combines forward and backward SNS to capture both directions of the data–annotation relationship, along with EEE + projection network.

5.3 Baselines

- **"Traditional" Data Curation Pipeline** To evaluate our method against common industry practices for curating datablends, we implement a baseline that combines heuristic filtering with semantic ranking. Further details on this approach are available in Appendix A.6.1.
- **Uniform Random Sampling.** We randomly sample $k = 5{,}000$ pairs uniformly from the complete candidate pool without replacement.
- **Stratified Random Sampling.** To control for potential pool-level imbalance, we evaluate a stratified random sample that ensures equal representation across data sources. We sample 1,000 pairs from each of the five pools for a total of $k = 5{,}000$ pairs.

5.4 Downstream Performance Experiments

Figure 5 and Table 1 present the validation perplexity trajectories for all seven data curation strategies, each averaged over three independent training runs with 95% confidence intervals. At step 750, the lowest perplexity of 7.732 is achieved by both *SNS Bidirectional + EEE + Projection* and *SNS Backward + EEE + Projection*.

A clear three-tier pattern emerges among the methods. First, the *Traditional Data Curation Baseline* is consistently and significantly worse than all other methods, converging to 7.781 ± 0.003 at step 750—well above the rest of the field. Second, the *EEE + Projection Only* variant and the two random

Table 1: Validation perplexity (mean ± 95% CI, n=3) at selected training steps. **Bold** indicates the lowest (best) perplexity per column. The *Best PPL* column indicates the best perplexity over all steps.

Method	Step 250	Step 500	Step 750	Best PPL
EEE + Projection Only	7.753 ± 0.004	7.750 ± 0.001	7.752 ± 0.005	7.748 ± 0.005
SNS Forward + EEE + Projection	7.754 ± 0.001	7.741 ± 0.004	7.734 ± 0.003	7.734 ± 0.003
SNS Backward + EEE + Projection	$\mathbf{7.750 \pm 0.010}$	$\mathbf{7.734 \pm 0.002}$	7.734 ± 0.005	$\mathbf{7.732 \pm 0.002}$
SNS Bidirectional + EEE + Projection	7.755 ± 0.001	7.740 ± 0.006	$\mathbf{7.732 \pm 0.011}$	$\mathbf{7.732 \pm 0.011}$
Uniform Random Baseline	7.753 ± 0.006	7.738 ± 0.008	7.736 ± 0.009	7.734 ± 0.013
Stratified Random Baseline	7.751 ± 0.005	7.745 ± 0.004	7.737 ± 0.009	7.737 ± 0.009
Traditional Data Curation Baseline	7.787 ± 0.005	7.783 ± 0.001	7.781 ± 0.003	7.780 ± 0.002

Table 2: Modality gap (ℓ_2 distance between per-modality embedding centroids) for each expert encoder and the learned projection network. The projection network reduces the average modality gap by over *90%* relative to the base experts. Figure 21 depicts the embedding space geometry.

	Gap				Δ vs. Projection (%)			
Space	**Video**	**Audio**	**Image**	**Text**	**Video**	**Audio**	**Image**	**Text**
E2E Expert	44.29	44.79	27.59	18.77	−99.8	−99.8	−99.7	−99.6
Fusion Expert	30.72	31.21	30.80	46.13	−99.7	−99.7	−99.7	−99.8
Text Expert	0.440	0.292	0.263	0.254	−81.6	−66.6	−66.8	−71.6
EEE + Projection (Ours)	**0.081**	**0.098**	**0.087**	**0.072**	—	—	—	—

baselines form a middle tier, converging to perplexities between 7.736 and 7.752. Third, the three SNS-augmented variants achieve the lowest perplexities, with *SNS Backward + EEE + Projection* and *SNS Bidirectional + EEE + Projection* reaching 7.734 and 7.732 respectively at step 750.

Table 2 illustrates the modality gap collapse effect from our approach (*EEE + Projection*) compared to base experts, reducing ℓ_2 distance between per-modality centroids by over 90% on average.

6 Conclusion

Ranked retrieval is increasingly used to curate training mixtures across multimodal, multidomain datasets. However, our study shows that small geometric quirks and label noise can compound into systematic errors when retrieval is treated as standard infrastructure. We demonstrate that standard retrieval based data selection can actively degrade training mixtures when the embedding space exhibits uncorrected modality bias.

We presented a framework that addresses this by acting jointly on training pairs and embeddings through two complementary components: Symmetric Nucleus Subsampling (SNS) and Expert Embedding Engine (EEE). The SNS reduces raw data to supervision misalignment by trimming the raw input and the annotation under a symmetric similarity gate. The EEE trains a mixture of embedding experts with a learned projection using a bias-aware objective.

In our multimodal instruction-following evaluation on a five-pool mixture spanning audio-text, image-text, video-text, and text-text pairs, the *SNS Bidirectional + EEE* configuration achieved the lowest validation perplexity score of **7.732** outperforming stratified random sampling and the traditional data curation pipeline. This performance is underpinned by a fundamental geometric shift: our projection network collapses the modality gap, reducing the ℓ_2 separation between modality centroids by an average of over **90%** compared to base experts (Table 2), ensuring that retrieval is driven by semantic alignment rather than input format.

Despite these gains, our framework has limitations. First, our multimodal instruction-following evaluation is limited in scale. Secondly, our traditional curation baseline is limited as heuristic quality filters and single encoder rankings are applied to text-based supervisions only, which may underestimate what a stronger multimodal curation pipeline operating on both the raw data and supervision could achieve.

A natural next step is to make both trimming and projection uncertainty-aware, using signals like expert disagreement, augmentation instability, or description inconsistency to decide when to trim, when to fall back, and which experts to trust. Another direction is to decouple proposal from verification: use fast embeddings to retrieve candidates, then apply a stronger cross-modal verifier for reranking to reduce self-reinforcing selection effects. Finally, we plan to evaluate curation more directly and at larger scales, with metrics that track mixture diversity and tail coverage over iterations.

ACKNOWLEDGMENTS

LLM Usage (manuscript wide). In accordance with ICLR 2026 policy, we disclose that we used a large language model during manuscript preparation across multiple sections of the paper. The model was used to assist with drafting and editing text, improving clarity and organization, and suggesting alternative phrasing. All technical content, claims, experimental results, and citations were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs, 2025.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What Made You Do This? Understanding Black-Box Decisions with Sufficient Input Subsets. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1810.03805.

Will Epperson, Arpit Mathur, Adam Perer, and Dominik Moritz. Texture: Structured Exploration of Text Datasets. *arXiv preprint arXiv:2504.16898*, 2025. URL https://arxiv.org/abs/2504.16898.

Samir Yitzhak Gadre et al. DataComp: In Search of the Next Generation of Multimodal Datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.14108.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space to Bind Them All. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Florian Grötschla, Luca A. Lanzendörfer, Marco Calzavara, and Roger Wattenhofer. AEye: A Visualization Tool for Image Datasets. *arXiv preprint arXiv:2408.04072*, 2024. URL https://arxiv.org/abs/2408.04072. Accepted at IEEE VIS 2024.

Lukas Heine, Fabian Hörst, Jana Fragemann, Gijs Luijten, Jan Egger, Fin Bahnsen, M. Saquib Sarfraz, Jens Kleesiek, and Constantin Seibold. Spacewalker: Traversing Representation Spaces for Fast Interactive Exploration and Annotation of Unstructured Data. *arXiv preprint arXiv:2409.16793*, 2024. URL https://arxiv.org/abs/2409.16793.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Shrimai Prabhumoye, Ayush Dattagupta, Vibhu Jawa, Jiwei Liu, Ryan Wolf, Sarah Yurick, Varun Singh, Dong Hyuk Chang, Ao Tang, Lawrence Lane, Charlie Truong, Huy Vu, Abhinav Garg, Praateek Mahajan, Nikolay Karpov, and Oliver König. NeMo-Curator: A Toolkit for Data Curation. https://github.com/NVIDIA-NeMo/Curator, 2024.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1601–1611, 2017.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in the Wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 119–132, 2019.

Mario Leopold, Farzad Tashtarian, and Klaus Schoeffmann. diveXplore: An Open-Source Software for Modern Video Retrieval with Image/Text Embeddings. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM) Open Source Software Track*, 2025. doi: 10.1145/3746027.3756877. URL https://dl.acm.org/doi/10.1145/3746027.3756877.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the Gap: Understanding the Modality Gap in Multi-Modal Contrastive Representation Learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, 2024. URL https://arxiv.org/abs/2303.05499.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. *arXiv preprint arXiv:2406.15126*, 2024.

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-Guided Query-Dependency Calibration for Video Temporal Grounding. *arXiv preprint arXiv:2311.08835*, 2023.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. NV-Retriever: Improving Text Embedding Models with Effective Hard-Negative Mining. *arXiv preprint arXiv:2407.15831*, 2024.

Hokuto Munakata, Taichi Nishimura, Shota Nakada, and Tatsuya Komatsu. Language-Based Audio Moment Retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

NVIDIA. Cosmos-Reason2-2B. https://huggingface.co/nvidia/Cosmos-Reason2-2B, 2025a. Reasoning vision-language model for physical AI and robotics.

NVIDIA. llama-3.2-nv-embedqa-1b-v2 Model by NVIDIA, 2025b. URL https://build.nvidia.com/nvidia/llama-3_2-nv-embedqa-1b-v2/modelcard.

NVIDIA, :, Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Guo Chen, Karan Sapra, Zhiding Yu, Adi Renduchintala, Charles Wang, Peter Jin, Arushi Goel, Mike Ranzinger, Lukas Voegtle, Philipp Fischer, Timo Roman, Wei Ping, Boxin Wang, Zhuolin Yang, Nayeon Lee, Shaokun Zhang, Fuxiao Liu, Zhiqi Li, Di Zhang, Greg Heinrich, Hongxu Yin, Song Han, Pavlo Molchanov, Parth Mannan, Yao Xu, Jane Polak Scowcroft, Tom Balough, Subhashree Radhakrishnan, Paris Zhang, Sean Cha, Ratnesh Kumar, Zaid Pervaiz Bhat, Jian Zhang, Darragh Hanley, Pritam Biswas, Jesse Oliver, Kevin Vasques, Roger Waleffe, Duncan Riach, Oluwatobi Olabiyi, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Pritam Gundecha, Khanh Nguyen, Alexandre Milesi, Eugene Khvedchenia, Ran Zilberstein, Ofri Masad, Natan Bagrov, Nave Assaf, Tomer Asida, Daniel Afrimi, Amit Zuker, Netanel Haber, Zhiyu Cheng, Jingyu Xin, Di Wu, Nik Spirin, Maryam Moosaei, Roman Ageev, Vanshil Atul Shah, Yuting Wu, Daniel Korzekwa, Unnikrishnan Kizhakkemadam Sreekumar, Wanli Jiang, Padmavathy Subramanian, Alejandra Rico, Sandip Bhaskar, Saeid Motiian, Kedi Wu, Annie Surla, Chia-Chih Chen, Hayden Wolff, Matthew Feinberg, Melissa Corpuz, Marek Wawrzos, Eileen Long, Aastha Jhunjhunwala, Paul Hendricks, Farzan Memarian, Benika Hall, Xin-Yu Wang, David Mosallanezhad, Soumye Singhal, Luis Vega, Katherine Cheung, Krzysztof Pawelec, Michael Evans, Katherine Luna, Jie Lou, Erick Galinkin, Akshay Hazare, Kaustubh Purandare, Ann Guan, Anna Warno, Chen Cui, Yoshi Suhara, Shibani Likhite, Seph Mard, Meredith Price, Laya Sleiman, Saori Kaji, Udi Karpas, Kari Briski, Joey Conway, Michael Lightstone, Jan Kautz, Mohammad Shoeybi, Mostofa Patwary, Jonathen Cohen, Oleksii Kuchaiev, Andrew Tao, and Bryan Catanzaro. NVIDIA Nemotron Nano V2 VL, 2025. URL https://arxiv.org/abs/2511.03929.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On Variational Bounds of Mutual Information. In *International conference on machine learning*, pp. 5171–5180. PMLR, 2019.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2301.02280.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10740–10749, 2020.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *European Conference on Computer Vision (ECCV)*, pp. 742–758, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, 2019. URL https://arxiv.org/abs/1807.03748.

Jin Xu, Zhifang Yan, Zitian Liu, Yihao Zhang, Chunfeng Zhu, Junyang Ye, Jianwei Song, Jiaxi Lu, Sicheng Yang, Shuai Yan, et al. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025a.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-Embed-Nemotron: A Unified Multimodal Retrieval Model for Text, Image, Audio, and Video, 2025b. URL https://arxiv.org/abs/2510.03458.

Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. Learning to Mine Aligned Code and Natural Language Pairs from Stack Overflow. In *Proceedings of the 15th International Conference on Mining Software Repositories (MSR)*, pp. 476–486, 2018.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-Centric Artificial Intelligence: A Survey. *ACM Computing Surveys*, 57(5):1–46, 2025. doi: 10.1145/3700100.

A APPENDIX

A.1 SNS ABLATIONS

In the following, we ablate the Symmetric Nucleus Subsampler (SNS) component across various hyperparameter ranges and configurations.

A.1.1 SNS DEFINITION

Forward extraction. The forward extraction function $\mathcal{N}_\alpha$ identifies raw components most relevant to the annotation:

$$\tilde{x} = \mathcal{N}_\alpha(x, y) = \{x_i : I(x_i; y) > \tau_\alpha\}, \tag{7}$$

where x_i denotes components such as text spans, image regions, or video clips, and τ_α controls extraction granularity.

Backward extraction. Symmetrically, $\mathcal{N}_b$ retains only annotation components grounded in observable data:

$$\tilde{y} = \mathcal{N}_b(x, y) = \{y_j : I(x; y_j) > \tau_b\}, \tag{8}$$

which removes label or tag elements that are not supported by the raw input.

Information density. The nucleus pair $(\tilde{x}, \tilde{y})$ increases information density:

$$\rho(\tilde{x}, \tilde{y}) = \frac{I(\tilde{x}; \tilde{y})}{|\tilde{x}| + |\tilde{y}|}, \tag{9}$$

since extraction aims to preserve $I(\tilde{x}; \tilde{y}) \approx I(x; y)$ while reducing $|\tilde{x}| + |\tilde{y}|$.

A.1.2 SNS HYPERPARAMETERS

We ablate three SNS knobs.

- **Directionality.** Forward (Eq. 7), backward (Eq. 8), or bidirectional extraction. With reinjection enabled, SNS can overwrite either x or y with $(\tilde{x}, \tilde{y})$ when the gate accepts the variant. We enable reinjection mode for all experiments.
- **MI gate ratio ρ.** We compute $\mathrm{sim}(\tilde{x}, \tilde{y})$ using a unified multimodal encoder and accept variants if
$$\mathrm{sim}(\tilde{x}, \tilde{y}) \geq \rho \cdot \mathrm{sim}(x, y). \tag{10}$$
Larger ρ is stricter and requires variants to match or exceed the original alignment.
- **Thresholds τ_α and τ_b.** These control extraction granularity in the forward and backward directions. Lower thresholds yield larger nuclei, while higher thresholds produce more focused nuclei.

A.1.3 DIRECTION ABLATION

Below are results from varying SNS directionality (`OFF`, `FORWARD`, `BACKWARD`, `BIDIRECTIONAL`) with $\rho = 1.00$ for 350 paired samples for each of the 5 data pools.

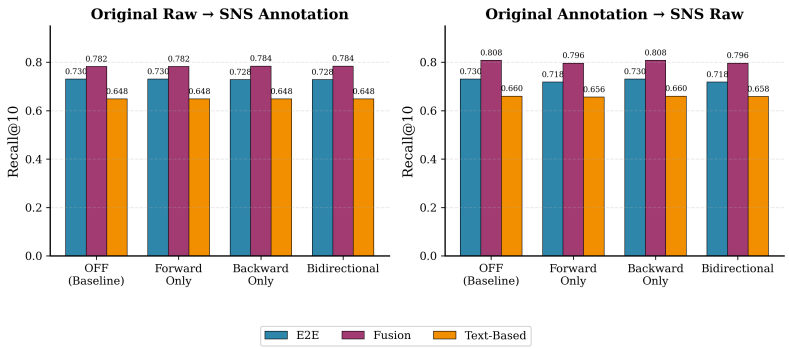


Figure 6: R@10 versus SNS directionality for both retrieval directions: Annotation to Raw Data (A2R); Raw Data to Annotation (R2A).

Table 3: Direction ablation of A2R & R2A R@10 by SNS configuration ($\rho = 1.00$).

Configuration	R→A E2E	R→A Fusion	R→A Text	A→R E2E	A→R Fusion	A→R Text
Baseline (OFF)	**0.7300**	0.7820	0.6480	**0.7300**	0.8080	**0.6600**
Forward Only	**0.7300**	0.7820	0.6480	0.7180	0.7960	0.6560
Backward Only	0.7280	**0.7840**	0.6480	**0.7300**	**0.8080**	**0.6600**
Bidirectional	0.7280	0.7840	0.6480	0.7180	0.7960	0.6580

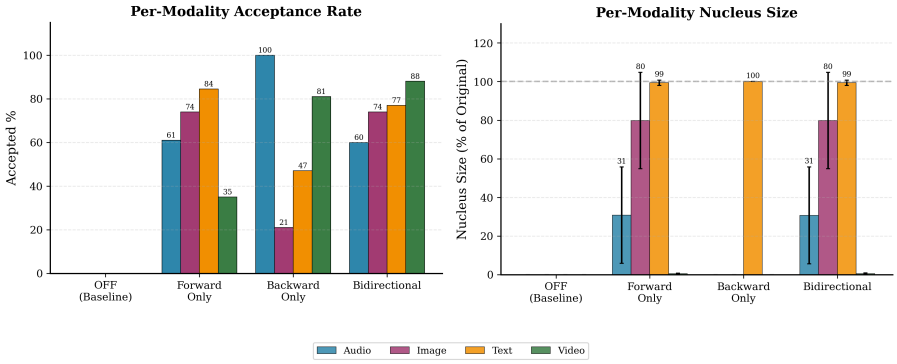


Figure 7: *Left*: fraction of accepted nucleus samples (fixed $\rho = 1.00$). *Right*: accepted nucleus size difference in bytes (*note*: for backwards extraction - all annotations are text).

A.1.4 MI GATE ABLATION

Below are results from varying the MI gate ratio ρ on a fixed sample of 1,000 pairs (200 per pool).

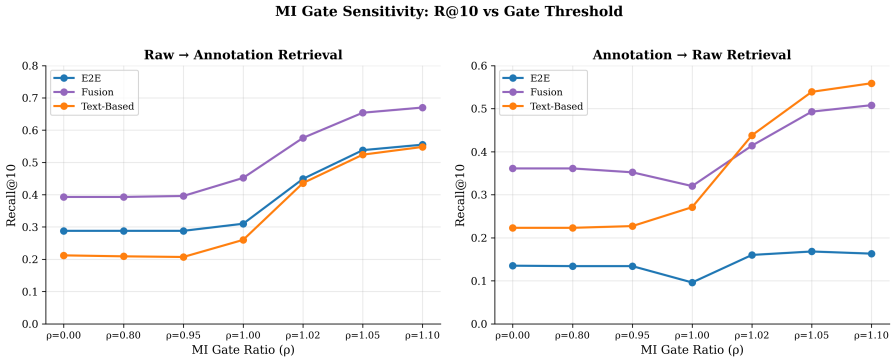


Figure 8: A2R & R2A R@10 versus MI gate ratio ρ.

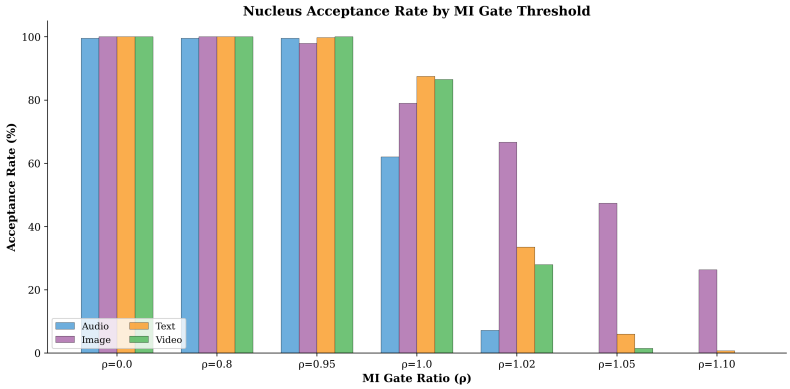


Figure 9: Acceptance rate versus MI gate ratio ρ.

At higher ρ, retrieval improves while the acceptance rate decreases (Figure 8 and Figure 9). This indicates a precision-oriented effect where stricter gating yields fewer but higher alignment pairs. In addition to R@10, a complementary analysis of nucleus compactness and cross-modality neighborhood consistency can better reflect the information density objective of SNS.

Table 4: MI gate ablation: R@10 by expert and retrieval direction.

(a) Raw → Annotation Retrieval

Expert	ρ=0.0	ρ=0.80	ρ=0.95	ρ=1.00	ρ=1.02	ρ=1.05	ρ=1.10
End-to-End	0.29	0.29	0.29	0.31	0.45	0.54	**0.56**
Fusion	0.39	0.39	0.40	0.45	0.58	0.65	**0.67**
Text-Based	0.21	0.21	0.21	0.26	0.44	0.52	**0.55**

(b) Annotation → Raw Retrieval

Expert	ρ=0.0	ρ=0.80	ρ=0.95	ρ=1.00	ρ=1.02	ρ=1.05	ρ=1.10
End-to-End	0.14	0.13	0.13	0.10	0.16	**0.17**	0.16
Fusion	0.36	0.36	0.35	0.32	0.41	0.49	**0.51**
Text-Based	0.22	0.22	0.23	0.27	0.44	0.54	**0.56**

A.1.5 TAU THRESHOLD ABLATION

As referenced in the main article, we ablate SNS with modality-specific τ_α and τ_β. In practice, we have modality-specific τ_α and τ_β, as we learned through the below experiment that multimodal datasets have modality-specific embedding similarity (a proxy to mutual information density) ranges. In the following, we compute the pairwise embedding similarity scores for 1024 pairs for each the 4 distinct modality data pools in our dataset.

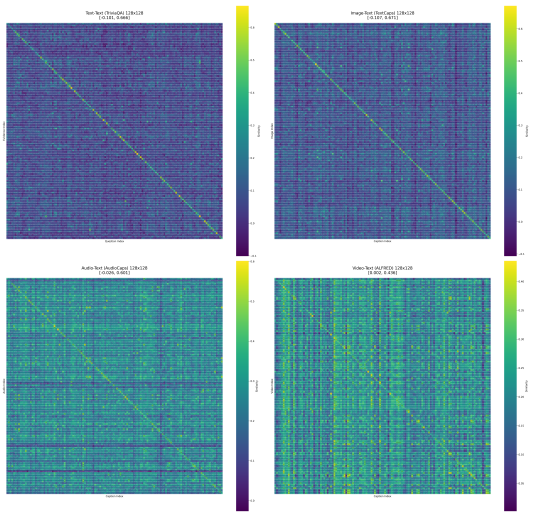


Figure 10: Pairwise similarity across all modalities of paired samples, evaluated on 1024 samples using the Omni-Embed-Nemotron-3B model.

Next, forward thresholds τ_α and backward thresholds τ_b are jointly varied across 250 samples per data pool for the five data pools.

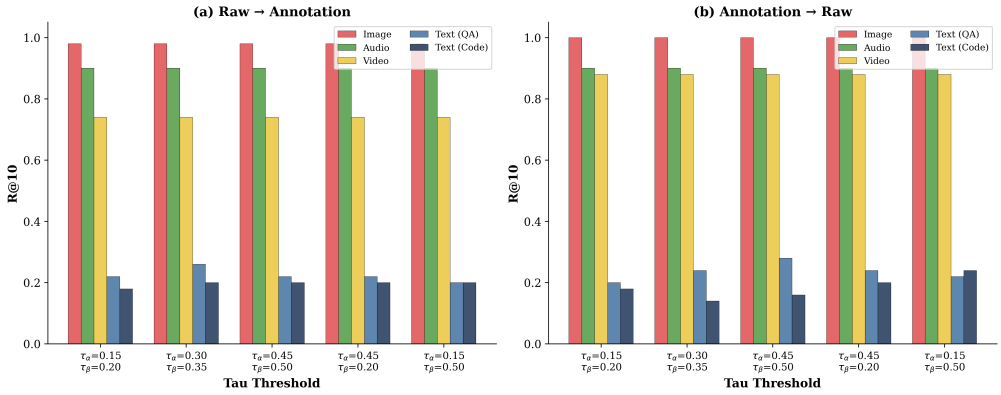


Figure 11: A2R & R2A R@10 versus (τ_α, τ_b), stratified by modality.

Jointly varying τ_α and τ_b has minimal impact on retrieval in this range. This suggests components near the decision boundary contribute less to the gated similarity objective, and future work can explore more aggressive thresholds or adaptive thresholding by modality.

A.2 EEE ABLATIONS

Below, we discuss various ablation experiments applied to the Expert Embedding Engine component.

A.2.1 METRICS

We measure cross-modal retrieval using Recall@K (R@K), defined as the fraction of queries where the correct match appears in the top K results:

$$\text{R@K} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[\text{rank}(x_i, y_i) \leq K\big], \tag{11}$$

where $\text{rank}(x_i, y_i)$ is the position of the true match when retrieving y_i from query x_i. We evaluate both directions:

- **R2A (Raw → Annotation)**: Given raw data, retrieve the matching text annotation.
- **A2R (Annotation → Raw)**: Given text annotation, retrieve the matching raw data.

A.2.2 MODALITY GAP DIAGNOSTIC

Multimodal embeddings can cluster by modality rather than semantic content. For embedding function $f : \bigcup_{m\in\mathcal{M}} \mathcal{X}_m \to \mathbb{R}^d$, modality clustering occurs when:

$$\mathbb{E}\big[\|f(x_i) - f(x_j)\| \mid x_i, x_j \in \mathcal{X}_m\big] < \mathbb{E}\big[\|f(x_i) - f(x_k)\| \mid x_i \in \mathcal{X}_{m_1}, x_k \in \mathcal{X}_{m_2}\big]. \tag{12}$$

A.2.3 MODALITY AND SNS VARIANT ABLATIONS

We evaluate three expert embedding engines (**E2E**, **Fusion**, **Text**) across five pools with four SNS variants (Baseline, Forward, Backward, Bidirectional).

Figure 12: Per pool R@10 heatmaps for A2R and R2A retrieval.

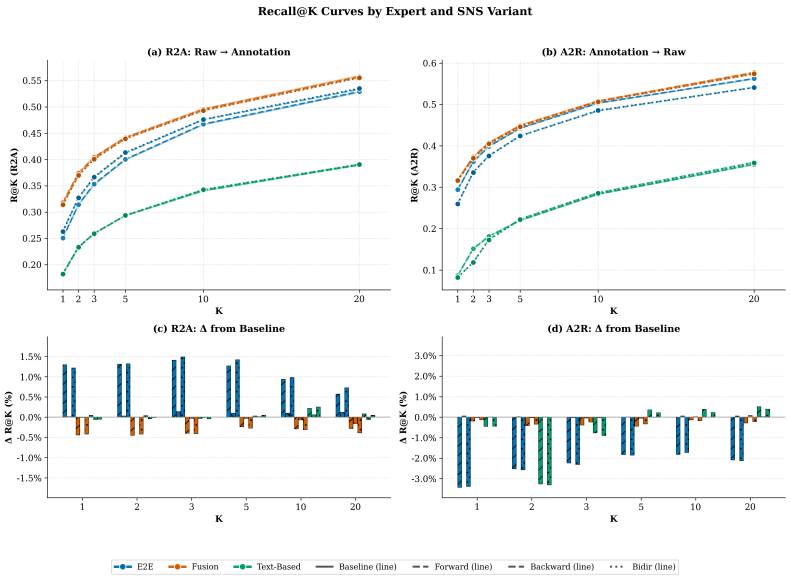


Figure 13: Recall@K curves and change from baseline for each expert and SNS variant.

Table 5: Recall@K by expert and SNS variant. Values shown as R2A / A2R.

Expert	SNS	R@1	R@2	R@3	R@5	R@10	R@20
E2E	Baseline	.251/.294	.314/.361	.352/.399	.399/.442	.466/.503	.528/.562
	Forward	.264/.259	.327/.336	.366/.376	.412/.424	.476/.485	.534/.541
	Backward	.251/.294	.314/.362	.353/.399	.400/.442	.467/.503	.529/.563
	Bidir	.263/.260	.327/.336	.367/.376	.413/.424	.476/.486	.535/.541
Fusion	Baseline	**.318/.318**	**.374/.374**	**.405/.408**	**.442/.450**	.496/.508	.559/.576
	Forward	.314/.316	.370/.370	.401/.404	.440/.445	.493/.507	.556/.574
	Backward	**.318**/.317	**.374**/.373	.404/.407	**.442**/.449	.495/**.508**	.558/**.577**
	Bidir	.314/.316	.370/.370	.401/.405	.439/.446	.493/.507	.555/.574
Text	Baseline	.183/.087	.233/.151	.259/.181	.293/.220	.340/.283	.390/.355
	Forward	.183/.082	.234/.119	.259/.174	.293/.223	.342/.287	.391/.360
	Backward	.182/.087	.233/.151	.259/.181	.293/.220	.341/.283	.389/.355
	Bidir	.182/.082	.233/.118	.259/.172	.294/.222	.343/.285	.391/.359

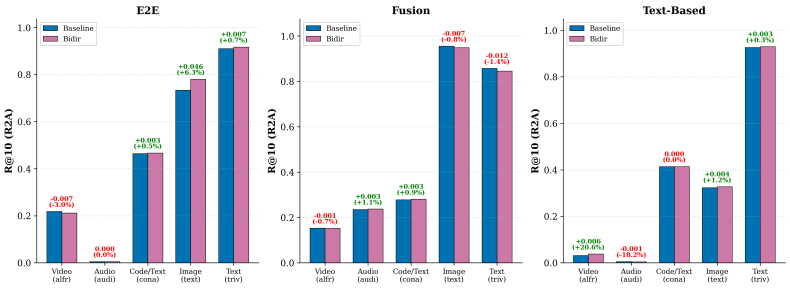
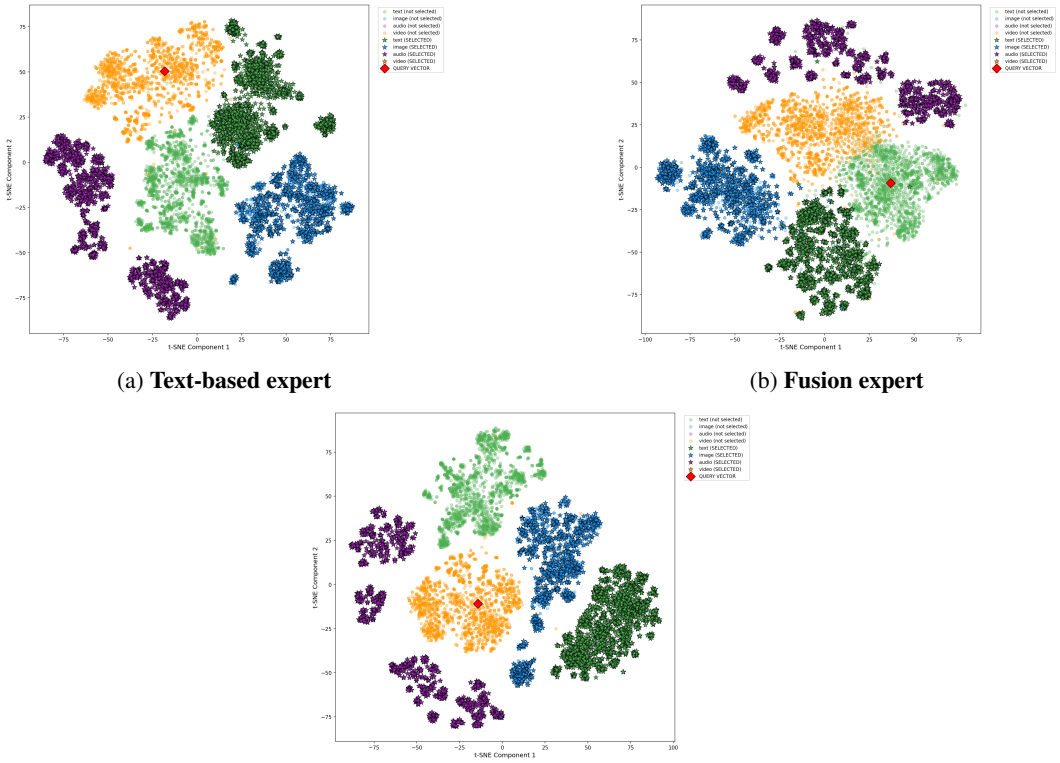


Figure 14: Baseline versus Bidirectional SNS by pool, R2A R@10 retrieval impact

We find that the fusion expert performs well across various R@K values and demonstrates the strongest single-expert correlation between raw data and paired annotation compared to text-based and end-to-end multimodal embedding model implementations.

A.2.4 EXPERT ABLATIONS - DOWNSTREAM EVAL PERFORMANCE

In the following, we share the isolated downstream evaluation results of the individual experts in the EEE. To isolate the experts, we disable SNS & the projection network, and evaluate the individual experts' embedding spaces as datablend curation semantic maps. We find that individual experts curate datablends differently, with vastly different modality compositions, likely due to the geometry of the embeddings in the respective expert's embedding spaces.



(a) **Text-based expert** (b) **Fusion expert**

(c) **End-to-End (E2E) expert**

Figure 15: 2D t-SNE visualization of curated samples (5000 samples of 10000) given query vector "natural, real-world scenes with objects, landscape, subjects, or people" across isolated EEE experts.

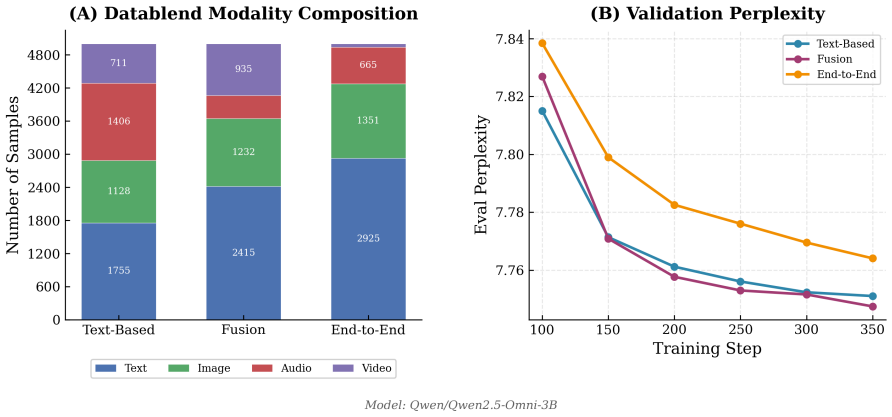


Figure 16: *Left*: Datablend Modality Composition by EEE expert curation, *Right*: Downstream eval validation perplexity curve across 1 epoch fine-tuning Qwen-2.5-Omni-3B for multimodal understanding on curated datablends by EEE expert.

A.3 PROJECTION NETWORK ABLATIONS

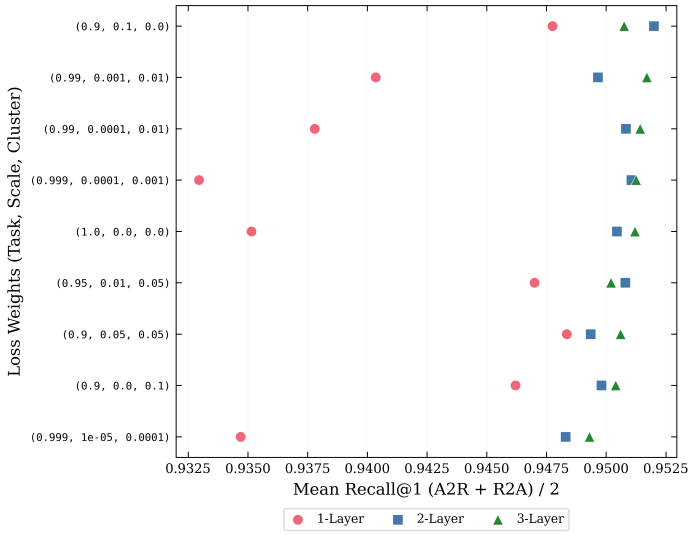


Figure 17: Mean recall (avg. of A2R and R2A) across loss weight configurations and network depths.

Table 6 and Figure 18 summarize the paired-data recall performance of the adaptive projection network across 30 configurations varying in loss term weighting and network depth.

Table 6: Projection network ablation: mean recall (avg. of A2R and R2A) across loss weight configurations and network depths. Bold indicates best per column.

Task	Scale	Cluster	Layers	R@1	R@3	R@5
0.9	0.1	0	2	**0.9520**	0.9919	0.9972
0.99	1e − 3	0.01	3	0.9517	0.9920	0.9970
0.99	1e − 4	0.01	3	0.9514	0.9922	0.9972
0.999	1e − 4	1e − 3	3	0.9512	0.9919	0.9971
1	0	0	3	0.9512	0.9921	0.9973
0.999	1e − 4	1e − 3	2	0.9510	0.9919	0.9973
0.99	1e − 4	0.01	2	0.9508	0.9924	0.9970
0.95	0.01	0.05	2	0.9508	0.9921	0.9971
0.9	0.1	0	3	0.9507	0.9918	0.9972
0.9	0.05	0.05	3	0.9506	0.9917	0.9970
1	0	0	2	0.9505	0.9921	0.9971
0.9	0	0.1	3	0.9504	0.9921	**0.9973**
0.95	0.01	0.05	3	0.9502	0.9919	0.9970
0.9	0	0.1	2	0.9498	0.9920	0.9971
0.99	1e − 3	0.01	2	0.9496	0.9919	0.9969
0.9	0.05	0.05	2	0.9494	0.9921	0.9973
0.999	1e − 5	1e − 4	3	0.9493	0.9925	0.9972
0.9	0.05	0.05	1	0.9484	0.9919	0.9967
0.999	1e − 5	1e − 4	2	0.9483	**0.9926**	0.9972
0.9	0.1	0	1	0.9477	0.9922	0.9966
0.95	0.01	0.05	1	0.9470	0.9924	0.9968
0.9	0	0.1	1	0.9462	0.9916	0.9965
0.99	1e − 3	0.01	1	0.9404	0.9892	0.9962
0.99	1e − 4	0.01	1	0.9378	0.9888	0.9959
1	0	0	1	0.9351	0.9887	0.9955
0.999	1e − 5	1e − 4	1	0.9347	0.9879	0.9950
0.999	1e − 4	1e − 3	1	0.9329	0.9871	0.9946

Network depth has a pronounced effect on retrieval quality. 3-layer projections achieve the highest mean R@1 of 0.9508, while 1-layer networks score 0.9408—a relative gap of 1.07%.

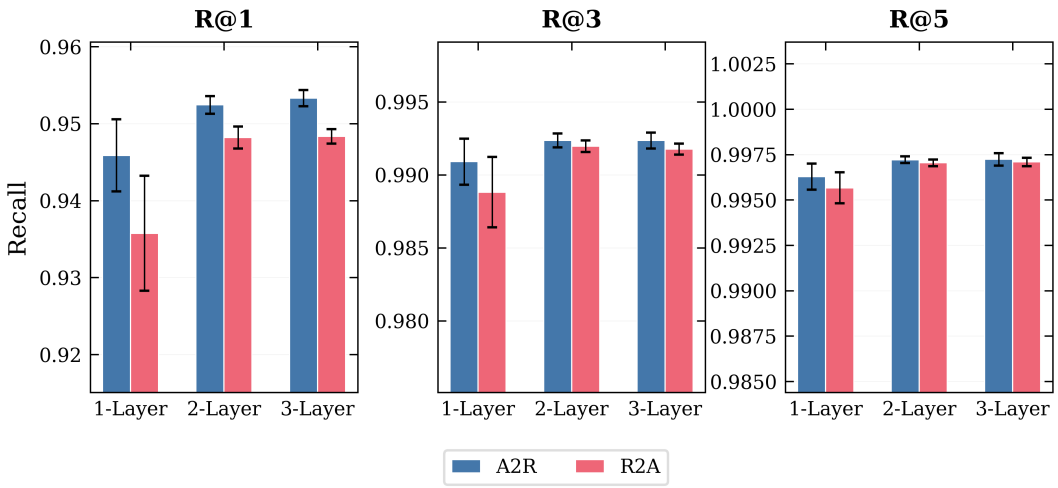


Figure 18: Projection layer depth effect on A2R & R2A R@1,3,5.

The effect of the projection network on modality composition of datablends is pronounced when comparing Figure 19 with Figure 16.

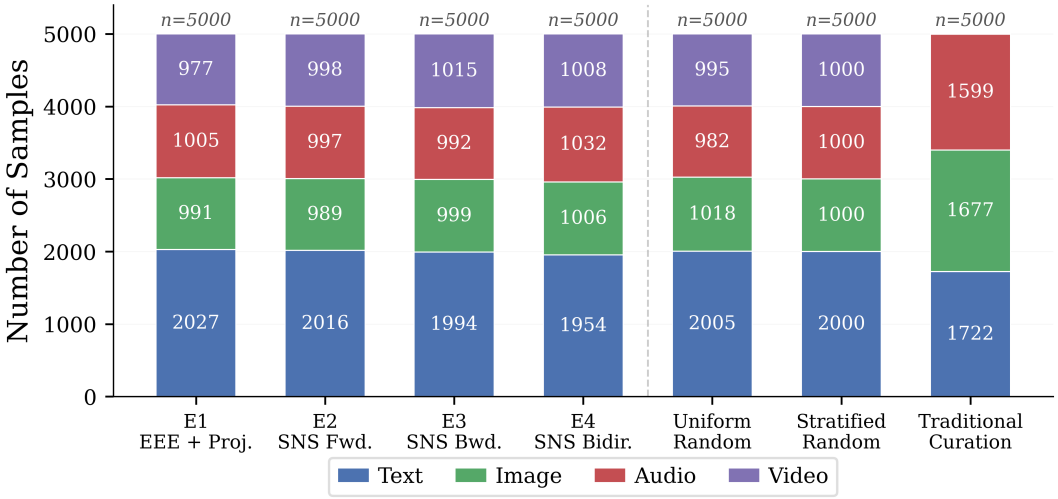
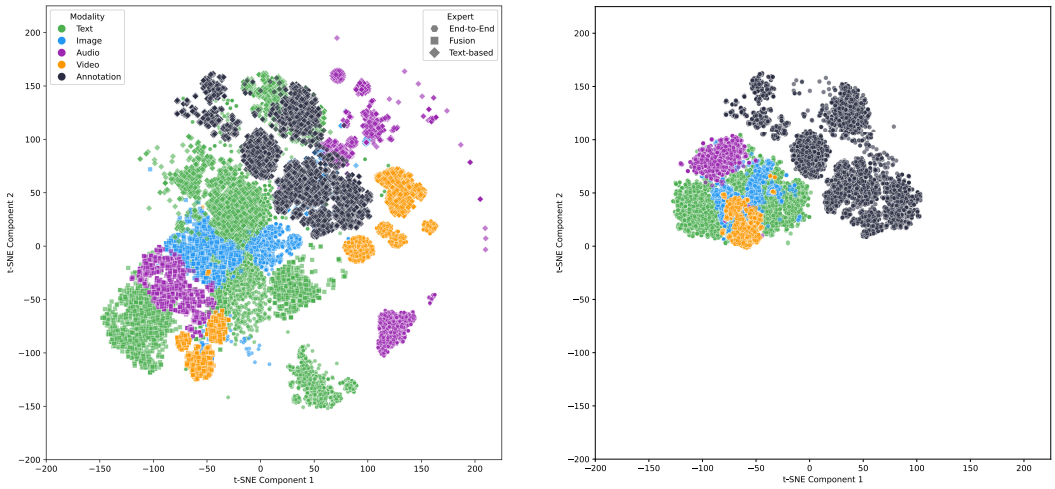


Figure 19: Datablend modality composition compared across downstream evaluation study curation strategies.

A.3.1 PROJECTION NETWORK EMBEDDING SPACE GEOMETRY

We also explore the embedding space geometry closer to understand the impact of the projection network on mitigating modality gap and non-semantic clustering tendencies of base embedding experts.



(a) Isolated base expert embeddings (Fusion, Text-Based, End-to-End), no SNS.

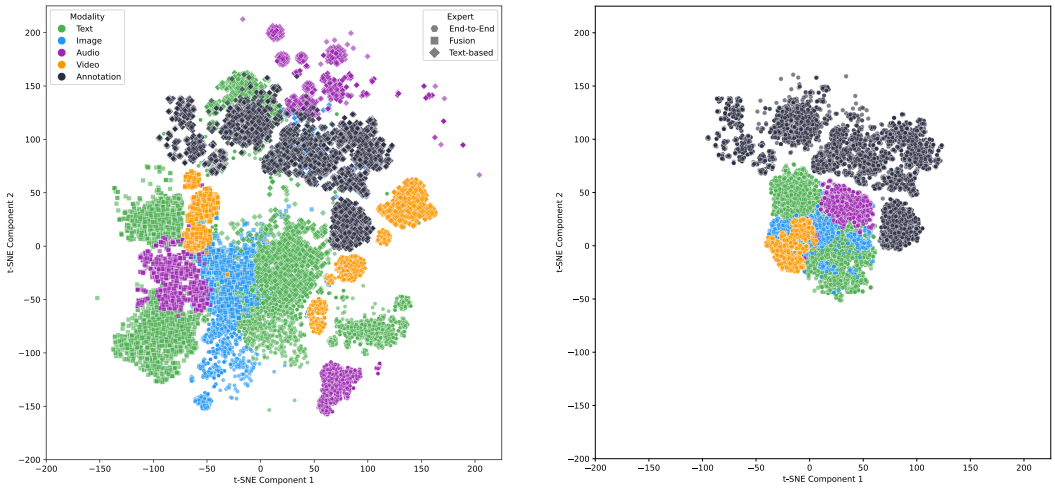
(b) Fused embeddings after Projection Network

Figure 20: 2D t-SNE visualizations of embedding spaces pre- vs post- Projection Network. Modality gap clustering disappears after passing through the projection network. *Note:* the grounded anchor embeddings for the text annotations are also displayed to show learned proximity between raw data embeddings $\left[e_{\text{fused}}\right]$ and the static annotation embeddings $\left[a_{e_a}\right]$.

Table 7: Modality gap (ℓ_2 embedding distance) across embedding spaces.

	Gap				Δ vs. Projection (%)			
Space	**Video**	**Audio**	**Image**	**Text**	**Video**	**Audio**	**Image**	**Text**
E2E Expert	45.73	45.39	29.09	19.50	−99.8	−99.8	−99.7	−99.6
Fusion Expert	30.36	30.82	30.44	45.57	−99.7	−99.7	−99.7	−99.8
Text Expert	0.441	0.294	0.265	0.257	−82.1	−67.3	−67.4	−72.4
EEE + Projection (Ours)	**0.079**	**0.096**	**0.086**	**0.071**	—	—	—	—

A.4 EMBEDDING SPACE GEOMETRY: BASE MULTIMODAL EXPERTS VS EEE + PROJECTION



(a) Base expert embeddings (Fusion, Text-Based, End-to-End)

(b) Fused embeddings after Projection Network

Figure 21: 2D t-SNE visualizations of embedding spaces without **(a).** and with **(b).** the projection network. Modality gap clustering is reduced by over 90% on average vs base experts. All base experts and our approach here apply SNS pre-processing to samples prior to embedding. *Note:* the grounded anchor embeddings for the text annotations are also displayed to show learned proximity between raw data embeddings $\left[e_{\text{fused}}\right]$ and the static annotation embeddings $\left[a_{e_a}\right]$.

A.5 SNS NUCLEUS EXAMPLES

Below, we show a couple of examples of multimodal nucleus extraction artifacts generated by the Symmetric Nucleus Subsampler (SNS) component, along with the Δ in approximate mutual information (MI).

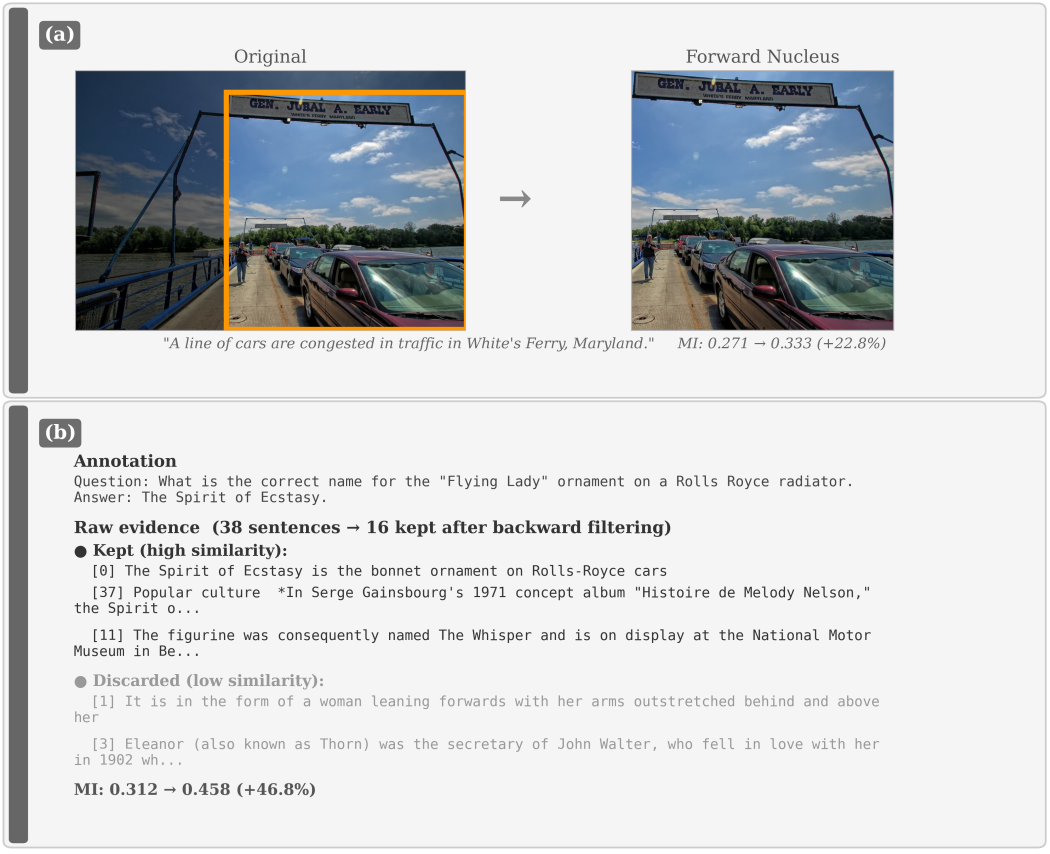


Figure 22: Pre- vs post-SNS examples. (a). Image sample from *TextCaps*, (b). Text sample from *TriviaQA*

A.6 Downstream Evaluation Study

A.6.1 Baselines

"Traditional" Data Curation Baseline We implement a baseline that combines heuristic filtering with semantic ranking on the annotations in our candidate pool of 10,000 samples across five data pools (ALFRED, AudioCaps, CoNaLa, TextCaps, TriviaQA). This pipeline utilizes NeMo Curator Jennings et al. (2024), an open-source toolkit for scalable multimodal data curation, to filter and deduplicate samples, followed by embedding-based selection using a fixed pretrained NeMo retriever model (`llama-3.2-nv-embedqa-1b-v2`) (NVIDIA, 2025b).

Unlike our proposed framework, which addresses raw-annotation misalignment jointly, this baseline operates in three decoupled stages:

1. **Unimodal Quality Filtering**: We first filter the annotations using text-based heuristic rules to remove low-quality supervision. Filters include a maximum non-alphanumeric character ratio (≤ 0.45), a maximum repeated line fraction (≤ 0.7), and boilerplate string removal.
2. **Semantic Deduplication**: We perform exact and semantic deduplication within the annotation pool to prevent over-representation of common topics. We employ sentence-transformers/all-MiniLM-L6-v2 embeddings with K-means clustering ($k = 100$) and apply a cosine similarity threshold ($\epsilon = 0.05$) to prune redundant annotations and their paired raw sample.
3. **Single-Encoder Ranking**: Surviving candidates are ranked by cosine similarity between their annotation embeddings and a fixed target query embedding ("natural, real-world scenes with objects, landscape, subjects, or people"). We select the top $k = 5{,}000$ pairs for the final training mixture.

This baseline represents a "representation-level" selection strategy that assumes the encoder's geometry accurately reflects meaning, without explicitly correcting for modality bias or raw-annotation misalignment.

A.7 Additional related work

A.7.1 Embedding-based dataset exploration and semantic search tools

A complementary body of work uses pretrained embeddings to explore datasets through semantic search and interactive visualization. Spacewalker is closely aligned with our setting because it supports multimodal data including text, images, and video, and provides interactive traversal of representation spaces with configurable embedding backbones and projection methods for inspection and annotation (Heine et al., 2024).

Relation to our work. Spacewalker and related embedding-based exploration systems show that a shared representation space is a practical substrate for browsing neighborhoods, diagnosing clusters, and surfacing outliers across modalities. Our use of embeddings is different in purpose and scale. Rather than focusing on interactive inspection and manual labeling, we use embedding-based search programmatically to curate and merge many human-labeled multimodal datasets into a single training pool. We then apply SNS to denoise and reweight paired examples with respect to their labels or tags, and retrain an expert-based embedding engine on the curated mixture to improve retrieval and reduce modality gaps.

A.7.2 Annotation aware subsampling

Interpretability work such as Sufficient Input Subsets (SIS) identifies minimal subsets of an input that preserve a model's prediction, exposing spurious shortcuts and offering instance-level explanations (Carter et al., 2019). While SIS is not a data selection method, it motivates the broader notion that subset structure can reveal which parts of an example carry signal versus noise.

Relation to our work. SNS is conceptually related to subset-based ideas but is designed for a different goal. SIS constructs sufficient subsets to explain individual model decisions, whereas SNS constructs nucleus variants for paired multimodal inputs and their labels or tags to improve the training distribution. SNS is also symmetric in that it gates informativeness in both directions, from raw modality to annotation and from annotation to raw modality. This symmetry is important when combining multiple labeled datasets with inconsistent taxonomies and variable annotation noise, and it supports our objective of improving retrieval behavior while reducing modality-driven separations in the learned embedding space.

A.7.3 Additional embedding-based exploration tools

Beyond Spacewalker (Heine et al., 2024), several systems instantiate embedding indexed exploration for specific data types. AEye focuses on scalable visualization and navigation for image datasets using embedding indexes and dimensionality reduction (Grötschla et al., 2024). For video collections, diveXplore provides embedding-based search and interactive exploration over videos using image and text embeddings (Leopold et al., 2025). For text corpora, Texture combines structured attribute views with embedding-based overview and neighborhood search to support inspection of data quality and subset construction (Epperson et al., 2025). These systems reinforce the utility of embedding spaces for navigation and inspection across modalities.