Multimodal Data Curation Through Ranked Retrieval

Pratyush Muthukumar [1] Harshil Kotamreddy [1] Sarah Amiraslani [1]
Tomo Kanazawa [1] Ramani Akkati [1] Shaan Jain [1] Andrew Mathau [1]

[1]NVIDIA

Introduction

Embedding-first systems are increasingly useful for modern applications such as multimodal search. However, some challenges are surfaced when multimodal and paired samples are introduced:

The geometry of embedding spaces often reflects modality identity rather than semantic meaning.

Paired samples and annotations often exhibit raw-annotation misalignment, where annotations do not describe all the information present in a sample, and a sample may contain more information than is annotated or vice versa.

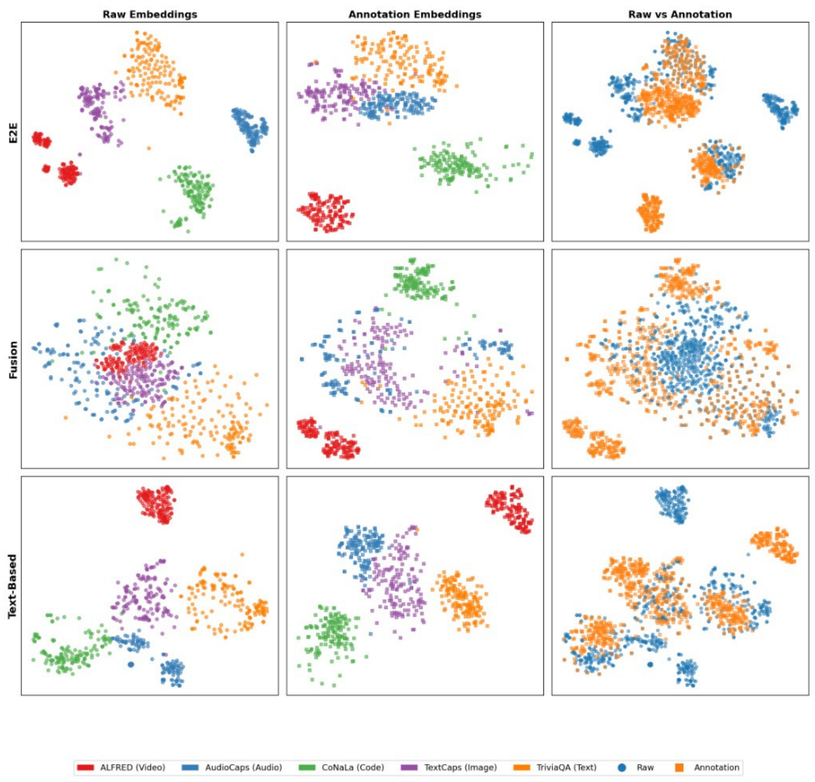


Figure 1. A 2D t-SNE visualization of paired data embeddings by a common multimodal embedding expert implementations (text-based, fusion, end-to-end), illustrating modality-dependent clustering.

Together these challenges reduce the efficacy of cross-modal retrieval by yielding results biased by modality and extraneous information not grounded to the mutual contents of the paired samples. We propose a solution to tackle both of these challenges jointly.

Methods

We use a three-step approach to rank data samples across domains and modalities:

Symmetric Nucleus Subsampler (SNS) reduces misalignment within data-pairs.

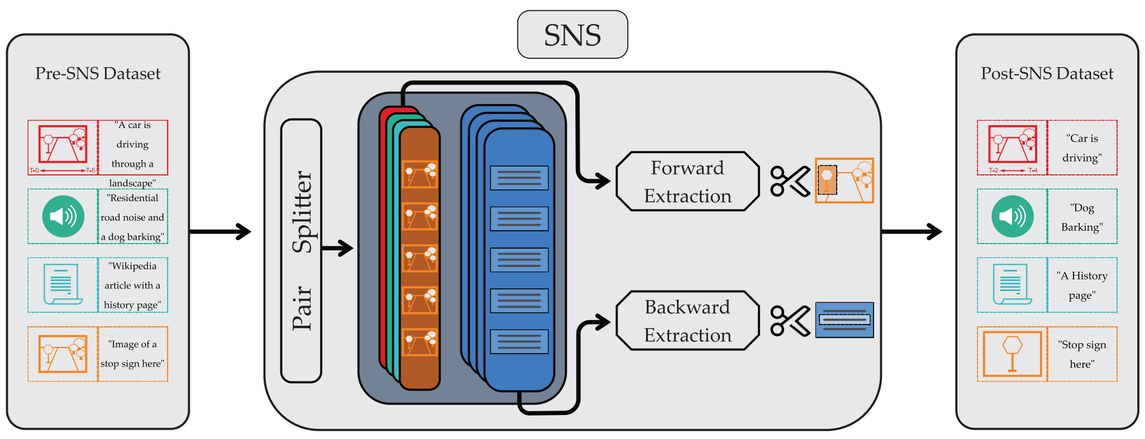


Figure 2. The SNS architecture. We reduce misalignment by extracting portions of the raw data most relevant to the annotation and vice versa. We do this using a suite of multimodal understanding models.

Expert Embedding Engine (EEE) uses several experts to reduce modality-specific bias.

Projection Network combines the embedding spaces of each expert into a new, unified embedding space.

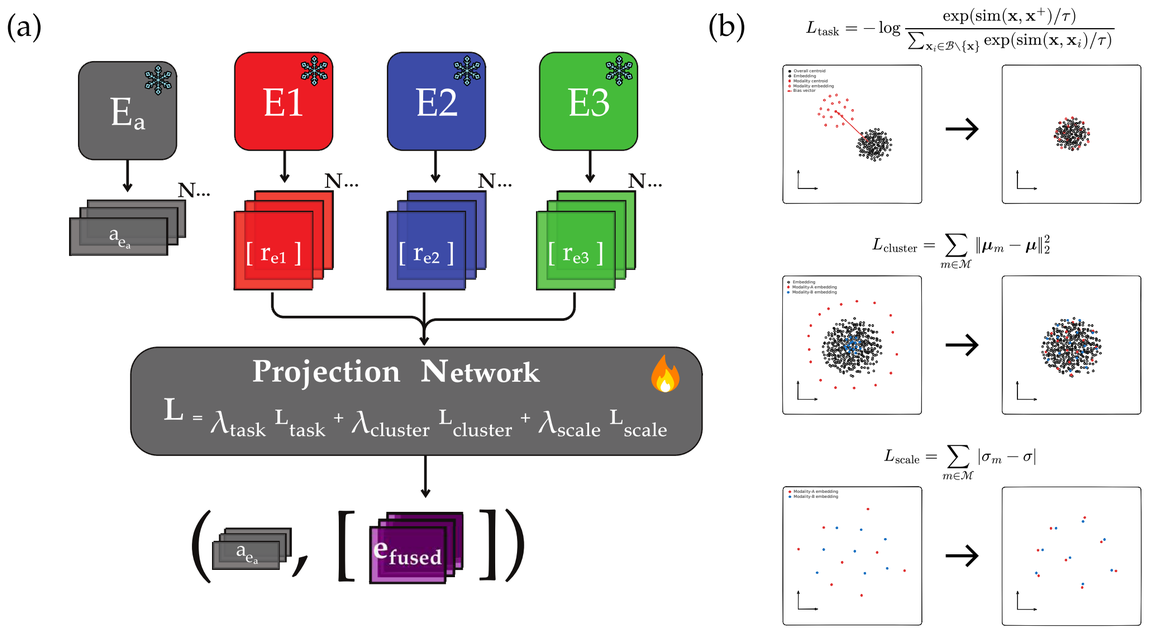


Figure 3. **(a)** The EEE and projection network. The EEE uses several experts to embed data pairs and the projection network combines the embeddings into a single, unified embedding space. **(b)** The loss function components for the projection network. The diagrams show an intuition for the 'before' and 'after' when each component of the loss function is applied to the data pairs.

We formulate a 3-term bias-aware objective to reduce modality-driven separation in the embedding space to learn a fused embedding e_{fused} for the `data` complement anchored by the annotation complement.

Results

We fine-tune **Qwen2.5-Omni-3B** [1] on curated blends from our multimodal curation engine (`SNS + EEE + Projection`) vs ablated variants + curation baselines (`Randomized, Traditional Curator`) [2]. All curation strategies pick from a datapool of 50K sample pairs of `(data, annotation)` with an objective to curate 10K samples for downstream training focusing on "natural, real-world scenes containing objects, landscapes, subjects, or people" [3, 4, 5, 6, 7].

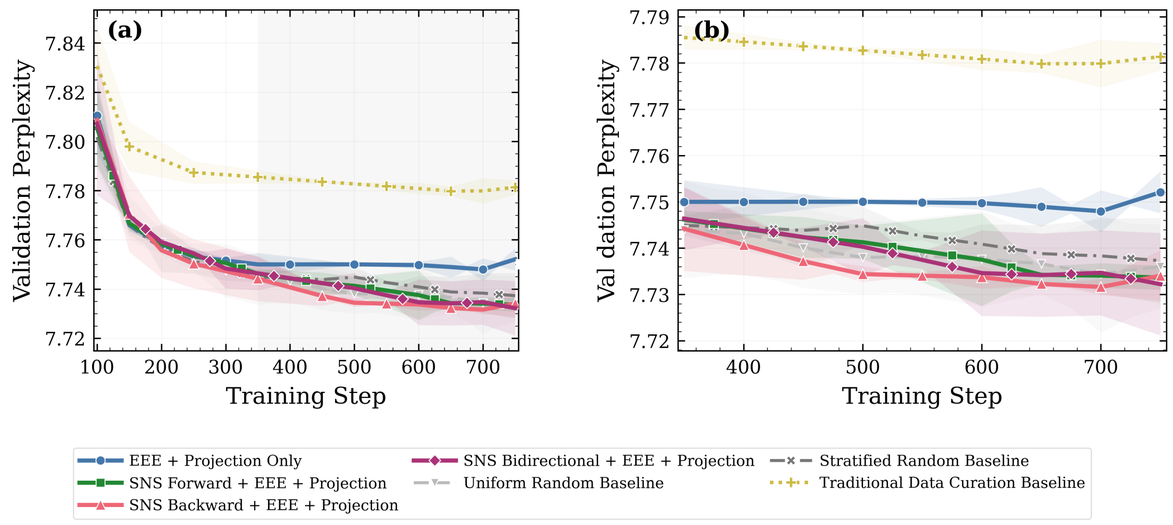


Figure 4. Validation perplexity (mean ± 95% CI, n=3) across SNS & EEE configuration variants vs baselines. **(a).** Full validation perplexity curves, **(b).** Last epoch validation perplexity curves.

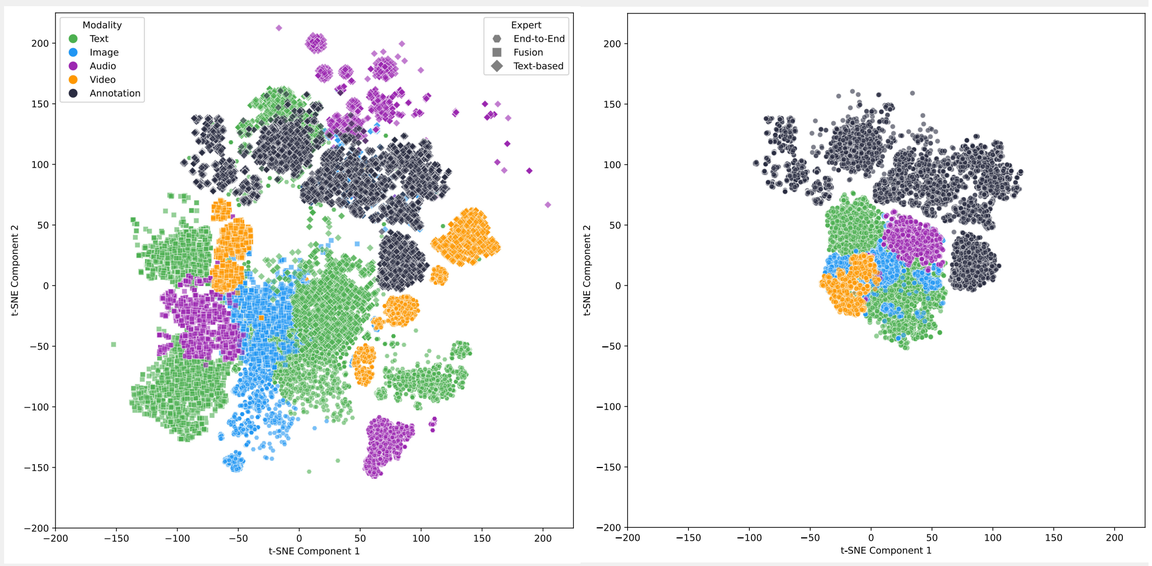
Models trained on the curated datablends are evaluated on a held-out validation set of 500 sample pairs hand-picked in a blind curation setup by humans.

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

Embedding Space Geometry

`SNS (Bidirectional) + EEE + Projection` (ours) collapses the modality gap ($\frac{1}{\mathcal{M}(\mathcal{M}-1)} \sum_{i \neq j}^{\mathcal{M}} \|\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j\|_2$, where $\bar{\mathbf{z}}_i$ = modality centroid for $\mathcal{M}$ modalities) by over 90% compared to any base embedding expert (`Fusion`, `Text-Based`, `End-to-End`).



(a) Base expert embeddings (Fusion, Text-Based, E2E) (b) Fused embeddings after Projection Network

Figure 5. 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 $[e_{\text{fused}}]$ and the static annotation embeddings $[a_{e_a}]$.

Conclusion

Downstream Model Performance Gains Our **SNS Bidirectional + EEE** pipeline achieved the lowest validation perplexity (**7.732**), outperforming both stratified random sampling and traditional curation baselines.

Modality Alignment The projection network successfully collapsed the modality gap, reducing ℓ_2 centroid separation between modalities by **>90%** vs base experts.

Future Work The traditional curation baseline is constrained as filters & rankers are applied to text-based supervisions only. Future iterations will focus on scaling SNS efficiency and developing more robust **multimodal-first baselines**.

References

[1] Jin Xu et al. *Qwen2.5-Omni Technical Report*. 2025.

[2] Joseph Jennings et al. *NeMo-Curator: A Toolkit for Data Curation*. `https://github.com/NVIDIA-NeMo/Curator`. 2024.

[3] Mandar Joshi et al. "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)*. 2017, pp. 1601–1611.

[4] Oleksii Sidorov et al. "TextCaps: A Dataset for Image Captioning with Reading Comprehension". In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 742–758.

[5] Chris Dongjoo Kim et al. "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)*. 2019, pp. 119–132.

[6] Mohit Shridhar et al. "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 10740–10749.

[7] Pengcheng Yin et al. "Learning to Mine Aligned Code and Natural Language Pairs from Stack Overflow". In: *Proceedings of the 15th International Conference on Mining Software Repositories (MSR)*. 2018, pp. 476–486.

Paper Code