DEPTH OVER SPECIALIZATION IN SMALL MULTIMODAL TRANSFORMERS

Jakub Mroz, Henry Ndubuaku
Cactus
{jakub,henry}@cactuscompute.com

ABSTRACT

Shared encoders have proven effective for large-scale multimodal contrastive learning, but it is less clear whether their advantages persist in small, parameter-constrained regimes. We investigate this question through a focused empirical study by training models under strict transformer parameter budgets on a naturally aligned text, image, and speech dataset. Across a range of small model configurations, we observe that allocating transformer parameters to a single shared encoder often yields better retrieval performance than splitting the same capacity across modality-specific encoders. We further find that merging modality-specific encoders into a shared encoder can substantially reduce transformer parameters while preserving comparable performance on several modality pairs. Finally, in trimodal training, we observe an empirical trade-off in which adding a third modality improves weaker modality pairs while degrading stronger ones under fixed capacity. These results suggest that, in tightly constrained settings, parameters allocated to shared representations can be an effective default for parameter-efficient multimodal learning.

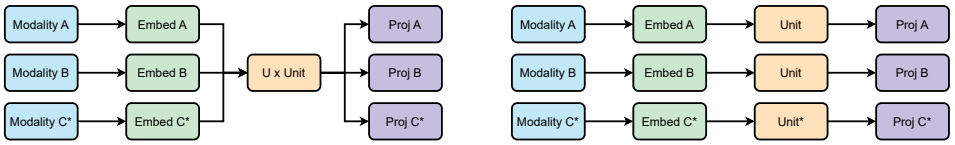
1 INTRODUCTION AND RELATED WORK

Large-scale multimodal contrastive learning models such as CLIP (Radford et al., 2021) and CLAP (Elizalde et al., 2023) typically rely on separate transformer encoders for each modality. While effective at scale, this design introduces substantial parameter overhead. Existing work has shown that shared encoders can improve parameter efficiency and performance by learning representations jointly across modalities (You et al., 2022). Although the benefits of shared encoders are well studied at large scale, it remains unclear whether they persist in small-scale settings. This question is particularly relevant for resource-constrained environments such as mobile devices or edge systems, where models may be limited to only a few million parameters. In small models, practitioners face a practical design choice: should limited transformer capacity be divided across modality-specific encoders or allocated to a single shared encoder? We address this question through a systematic study on the Localized Narratives dataset (Pont-Tuset et al., 2020), which provides naturally aligned text, images, and spoken audio. We train 16 multimodal contrastive models comparing shared and separate encoders across text–image, text–audio, image–audio, and trimodal settings. All models are trained under strict transformer parameter budgets ranging from 2.1M to 6.3M parameters, allowing us to isolate the effect of encoder sharing from overall model size. Our results show that shared encoders outperform separate encoders at matched transformer capacity. We further show that merging modality-specific encoders into a shared encoder can significantly reduce parameter count while retaining much of the retrieval performance. Finally, we analyze trimodal training and identify a trade-off between improving weak modality pairs and preserving performance on strong ones when capacity is fixed.

2 METHODOLOGY

2.1 DATASET

We use the COCO (Lin et al., 2014) subset of Localized Narratives (Pont-Tuset et al., 2020), which provides naturally aligned images, spoken narrations, and text transcriptions. We retain samples with



(a) Shared encoder (b) Separate encoders

Figure 1: Shared and modality-specific encoder architectures. Components marked with * are only present in the trimodal Text–Image–Audio (TIA) setting. In the shared architecture, all modalities are routed through a common encoder stack of U transformer units, whereas in the separate architecture, each modality is processed by an independent encoder.

Architecture	Modalities	Units [U]	Transformer Params [M]
Shared 1U	TI / TA / IA / TIA	1	2.1
Shared 2U	TI / TA / IA / TIA	2	4.2
Shared 3U	TI / TA / IA / TIA	3	6.3
Separate 2U	TI / TA / IA	1+1	4.2
Separate 3U	TIA	1+1+1	6.3

Table 1: Model variants used in our experiments. U denotes a transformer unit consisting of two layers. Reported parameter counts correspond to transformer parameters only.

audio duration shorter than 30 seconds and create a held-out test split from the original validation set. This yields 92,987 training samples (93.9%), 5,088 validation samples (5.1%), and 1,000 test samples (1.0%). The training set spans 82,683 unique images. The 1,000-sample test set spans 874 unique images, with 12.5% of images appearing with annotations from multiple narrators. Each triple is treated as a distinct sample with a single positive match during retrieval evaluation, meaning cross-narrator matches for the same image are not credited. Images are resized to 224×224. Audio is converted to log-mel spectrograms with 64 mel bins, a 16 kHz sampling rate, a hop length of 320, and zero-padding to a maximum duration of 30 seconds (1500 frames). Text is tokenized using the BERT (Devlin et al., 2019) vocabulary with a maximum sequence length of 256 tokens.

2.2 Architecture

Each modality uses modality-specific preprocessing designed to yield comparable sequence lengths across modalities. Text inputs are tokenized using the BERT (Devlin et al., 2019) vocabulary (vocab size 30,522) with a maximum sequence length of 256 tokens. Images are embedded using a ViT-style patch embedding (Dosovitskiy et al., 2020), dividing each image into a 14×14 grid of 16×16 patches, resulting in 196 patch tokens. Audio inputs are converted to log-mel spectrograms with 64 mel bins and are embedded using patch embeddings configured to produce a fixed sequence length of 240 tokens. A learnable [CLS] token is prepended to each input sequence. All models use a transformer encoder with $d_{\text{model}} = 256$, $d_{\text{ff}} = 1024$, and 8 attention heads. We apply dropout with probability 0.2, rotary positional embeddings (RoPE) (Su et al., 2024), RMSNorm (Zhang & Sennrich, 2019), and GELU-gated MLPs (Shazeer, 2020). The [CLS] token output is projected to a shared 512-dimensional embedding space for contrastive learning. In the shared setting, all modalities are routed through the same transformer parameters, differing only in modality-specific preprocessing and projection layers. In the separate setting, each modality is processed by an independent transformer encoder with identical architecture and hyperparameters. By explicitly controlling sequence lengths across modalities, we reduce confounding effects arising from differences in token count or attention cost, allowing comparisons to focus on the effect of encoder sharing under fixed transformer capacity. We compare two architectural designs (Figure 1). Shared models use a single transformer encoder for all modalities, while separate models allocate independent transformer encoders to each modality. Model variants are summarized in Table 1.

2.3 Training

We use a symmetric contrastive loss with a learnable temperature, following CLIP (Radford et al., 2021). For trimodal models, loss is averaged over three pairwise losses: $\mathcal{L}_{\text{TIA}} = (\mathcal{L}_{\text{TI}} + \mathcal{L}_{\text{TA}} + \mathcal{L}_{\text{IA}})/3$.

Model	Text-Image		Text-Audio		Image-Audio	
	Fwd	Bwd	Fwd	Bwd	Fwd	Bwd
Bimodal models						
Shared 1U	11.0±0.9	10.8±0.5	31.5±0.4	32.3±0.8	0.1±0.1	0.3±0.1
Shared 2U	16.4±0.6	15.4±0.5	47.2±3.0	45.6±3.0	0.2±0.1	0.1±0.1
Shared 3U	17.7±0.4	16.9±0.4	51.7±1.4	50.5±1.9	0.1±0.1	0.1±0.0
Separate 2U	11.3±0.3	11.7±0.4	37.7±0.8	36.7±0.9	0.1±0.0	0.1±0.1
Trimodal models						
Shared 1U TIA	6.9±0.5	6.9±0.9	28.5±2.4	25.0±0.5	0.3±0.2	0.4±0.1
Shared 2U TIA	11.9±0.7	11.8±0.3	46.8±1.4	43.4±1.6	0.5±0.2	0.7±0.2
Shared 3U TIA	13.2±0.6	12.4±0.7	52.9±1.4	49.1±1.0	1.0±0.2	0.6±0.1
Separate 3U TIA	7.7±1.1	7.6±1.0	33.6±1.6	30.6±1.6	0.5±0.2	0.3±0.1

Table 2: Recall@1 (%), reported as mean ± std across three seeds. Fwd/Bwd denotes retrieval direction (e.g., text-to-image / image-to-text for Text-Image). Shared encoders outperform separate encoders at matched capacity and retain performance when merging modality-specific encoders.

Model	Text-Image		Text-Audio		Image-Audio	
	Fwd	Bwd	Fwd	Bwd	Fwd	Bwd
Bimodal models						
Shared 1U	12.7±0.5	13.0±0.0	3.0±0.0	2.3±0.5	429.5±3.1	432.3±1.9
Shared 2U	7.7±0.5	7.8±0.2	1.7±0.5	2.0±0.0	450.3±10.4	454.7±2.2
Shared 3U	6.7±0.5	7.0±0.0	1.0±0.0	1.7±0.5	434.7±3.1	429.0±6.2
Separate 2U	13.0±0.8	13.7±0.5	2.0±0.0	2.0±0.0	462.3±10.8	459.0±19.1
Trimodal models						
Shared 1U TIA	20.2±1.3	21.7±0.9	3.0±0.0	3.3±0.5	298.0±10.0	308.5±6.7
Shared 2U TIA	11.7±0.5	12.0±0.0	2.0±0.0	2.0±0.0	198.8±14.1	203.7±12.7
Shared 3U TIA	10.0±0.8	10.0±0.0	1.0±0.0	2.0±0.0	187.7±2.6	191.0±5.1
Separate 3U TIA	19.3±0.5	19.7±1.2	2.3±0.5	2.3±0.5	273.2±20.6	274.3±12.8

Table 3: Median Rank (MedR, lower is better), reported as mean ± std across three seeds. Fwd/Bwd denotes retrieval direction. Adding a third modality improves MedR for the weak image–audio pair at every depth while degrading strong pairs.

All models are trained for 20 epochs using AdamW (Loshchilov & Hutter, 2017) (lr=10^{-3}, linear warmup 10%, cosine decay, weight decay 0.1, batch size 256, gradient clipping 1.0). We save the checkpoint with the lowest validation loss. All 16 models are trained on one A100 GPU sequentially in $\sim$11 hours per seed.

3 Experimentation and Results

We evaluate cross-modal retrieval in both directions (e.g., text-to-image and image-to-text) over a fixed 1,000-item candidate pool. For each query, we rank all candidates by cosine similarity and report Recall@K (K=1,5,10) and Median Rank (MedR). We additionally report Mean Reciprocal Rank (MRR) and NDCG@10 in Tables 7 and 8, both of which are consistent with R@1 and MedR across all configurations. Results are reported as mean ± std across three seeds, and the qualitative ordering of models is consistent across all individual seeds.

3.1 Equal Transformer Parameter Budget

We first compare shared and separate encoders under matched transformer parameter budgets. In the bimodal setting at 2U (4.2M transformer parameters), shared encoders outperform separate encoders across evaluated modality pairs. For text–image retrieval, the shared model achieves 16.4% / 15.4% Recall@1 with median ranks of 7.7 / 7.8, compared to 11.3% / 11.7% Recall@1 and median ranks of 13.0 / 13.7 for the separate model. For text–audio retrieval, shared encoders improve Recall@1 from 37.7% / 36.7% to 47.2% / 45.6%, while median rank improves from 2.0 / 2.0 to 1.7 / 2.0. Image–audio retrieval performance remains low for both architectures, with Recall@1 not exceeding 0.3% and median rank above 400. This trend extends to trimodal models. At 3U (6.3M transformer parameters), the shared TIA model achieves 13.2% / 12.4% Recall@1 with median ranks of 10.0 / 10.0 for text–image retrieval, compared to 7.7% / 7.6% Recall@1 and median ranks of 19.3 / 19.7

for the separate TIA model. Similar relative improvements are observed for text–audio retrieval. Overall, at matched transformer capacity, shared encoders outperform separate encoders across both bimodal and trimodal settings.

3.2 Reducing Parameters via Sharing

We next examine whether modality-specific encoders can be merged into a shared encoder while reducing transformer parameters. Specifically, we compare shared 1U models (2.1M transformer parameters) against separate 2U bimodal models (4.2M transformer parameters total). For text–image retrieval, the shared 1U model achieves 11.0% / 10.8% Recall@1 with median ranks of 12.7 / 13.0, compared to 11.3% / 11.7% Recall@1 and median ranks of 13.0 / 13.7 for the separate 2U model, yielding nearly identical performance with half the transformer parameters. For text–audio retrieval, the shared 1U model reaches 31.5% / 32.3% Recall@1 with a median rank of 3.0 / 2.3, compared to 37.7% / 36.7% Recall@1 and a median rank of 2.0 / 2.0 for the separate model. Image–audio retrieval remains challenging in both cases. In the trimodal setting, the shared 1U TIA model achieves 6.9% / 6.9% Recall@1 with median ranks of 20.2 / 21.7 for text–image retrieval, compared to 7.7% / 7.6% Recall@1 and median ranks of 19.3 / 19.7 for the separate 3U TIA model, despite using substantially fewer transformer parameters. These results show that merging encoders can substantially reduce transformer parameter count while retaining a large fraction of retrieval performance.

3.3 Effect of Adding a Third Modality

Finally, we analyze the effect of adding a third modality under fixed transformer capacity by comparing shared bimodal and trimodal models at the same depth. Across all depths, introducing a third modality improves weak modality pairs while degrading strong ones. For image–audio retrieval, adding text improves median rank at every depth. At 1U, MedR improves from 429.5 / 432.3 to 298.0 / 308.5. At 2U, MedR improves from 450.3 / 454.7 to 198.8 / 203.7. At 3U, MedR improves from 434.7 / 429.0 to 187.7 / 191.0. However, absolute Recall@1 values for image–audio remain at or below 1% in all configurations, reflecting the fundamental difficulty of this pair in Localized Narratives, where audio is a spoken verbal description of the image rather than a co-occurring environmental sound (see Section 4). The text modality acts as a semantic bridge that improves image–audio alignment indirectly. In contrast, strong modality pairs are negatively affected under fixed capacity. At 1U, text–image Recall@1 decreases from 11.0% / 10.8% to 6.9% / 6.9%, with median rank worsening from 12.7 / 13.0 to 20.2 / 21.7. At 2U, Recall@1 drops from 16.4% / 15.4% to 11.9% / 11.8%, with median rank worsening from 7.7 / 7.8 to 11.7 / 12.0. The same pattern holds at 3U. We interpret this as a capacity-allocation effect: adding $\mathcal{L}_{\text{IA}}$ forces the shared encoder to simultaneously optimize three pairwise objectives, and fixed capacity is insufficient to fully satisfy all three.

3.4 Mechanistic Analysis

We conduct three analyses to examine the mechanisms behind the main findings. First, we examine train-to-validation loss gaps (val loss − train loss) at the best checkpoint for each model and seed (Table 9). At matched 2U capacity, shared and separate bimodal models show comparable gaps: 0.81 vs. 0.83 for text–image, 0.46 vs. 0.44 for text–audio, and -0.01 vs. -0.01 for image–audio. In the trimodal setting at 3U, the shared gap (0.59) is 40% larger than the separate gap (0.42), yet the shared model achieves better retrieval across all modality pairs (Table 2). This indicates that the performance advantage of shared encoders is not driven by reduced overfitting, as in the trimodal setting shared encoders generalize worse while still retrieving better, pointing to a representational rather than regularization advantage. Second, we compute per-modality embedding-space diagnostics on the test set (Table 11): cross-modal alignment and per-modality uniformity (Wang & Isola, 2020) and anisotropy (Ethayarajh, 2019). Image–audio models exhibit anisotropy between 0.95 and 0.99, uniformity between -0.04 and -0.16, and alignment distances between 1.58 and 1.85 across all depths, compared to anisotropy between 0.05 and 0.11, uniformity between -3.06 and -3.34, and alignment between 0.85 and 0.89 for text–image models, confirming that image–audio representations remain poorly calibrated regardless of depth or architecture. Third, we measure pairwise cosine similarities between per-loss gradient vectors on the shared transformer parameters (Table 10). $\mathcal{L}_{\text{TI}}$

and $\mathcal{L}_{\text{TA}}$ exhibit mild negative gradient alignment (-0.090 ± 0.016), indicating modest competition for shared capacity. In contrast, $\mathcal{L}_{\text{TI}}$ and $\mathcal{L}_{\text{IA}}$ are positively aligned ($+0.224 \pm 0.013$), as are $\mathcal{L}_{\text{TA}}$ and $\mathcal{L}_{\text{IA}}$ ($+0.166 \pm 0.023$). The degradation of text–image performance when audio is added is therefore not explained by direct gradient conflict between $\mathcal{L}_{\text{TI}}$ and $\mathcal{L}_{\text{IA}}$, which are positively aligned. Instead, the pattern is consistent with a capacity-splitting effect: fixed transformer parameters must simultaneously optimize three objectives, and the mild TI/TA tension is sufficient to degrade the stronger text-anchored pairs under that constraint. The positive $\mathcal{L}_{\text{TI}}/\mathcal{L}_{\text{IA}}$ and $\mathcal{L}_{\text{TA}}/\mathcal{L}_{\text{IA}}$ cosines indicate that these loss pairs do not compete on the shared parameters, which is consistent with the observed image–audio improvement under trimodal training.

4 Limitations

Our findings are subject to several limitations. First, experiments are conducted on a single naturally aligned multimodal dataset, and results may not generalize to other domains or data regimes. Second, we focus exclusively on small transformer models and do not evaluate whether the observed trends persist at larger scales. Third, while we control for transformer parameter counts and explicitly match sequence lengths across modalities, other factors such as optimization dynamics or modality-specific inductive biases may still influence how shared capacity is utilized. In Localized Narratives, audio consists of spoken verbal descriptions of images rather than co-occurring environmental or scene sounds. Image–audio alignment is therefore mediated entirely through semantics, with no direct perceptual overlap between image pixels and audio waveforms. This is structurally more challenging than audio-visual datasets where sounds co-occur with visually related content. Our shared encoder design allocates all transformer layers uniformly to a single shared stack. Whether lower layers capture more transferable representations than higher layers, and which layers benefit most from sharing, has been studied at larger scales (You et al., 2022) and remains an open direction in the small-model regime. We evaluate exclusively in the cross-modal retrieval setting, and whether encoder-sharing advantages extend to zero-shot classification or linear probing on labeled downstream tasks is an important direction for future work.

5 Conclusion

We present a focused empirical study comparing shared and modality-specific encoders for small-scale multimodal contrastive learning under strict transformer parameter budgets. Across a range of bimodal and trimodal configurations, we observe that allocating transformer parameters to a shared encoder often yields improved retrieval performance relative to splitting capacity across modality-specific encoders. Our analysis further suggests that merging modality-specific encoders into a shared encoder can substantially reduce transformer parameters while preserving comparable performance on several modality pairs. In trimodal settings, we observe an empirical trade-off in which adding a third modality improves weaker modality pairs while degrading stronger ones under fixed capacity. Overall, these results suggest that when transformer parameters are the primary bottleneck, allocating them in a shared encoder can be a reasonable and often effective design choice for small multimodal systems.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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)*, pp. 55–65, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pp. 647–664. Springer, 2020.

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.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision*, pp. 69–87. Springer, 2022.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

A Appendix

Architecture	TI	TA	IA	TIA
Shared 1U	10.4	10.3	2.7	10.6
Shared 2U	12.5	12.4	4.8	12.7
Shared 3U	14.6	14.5	6.9	14.8
Separate 2U	12.5	12.4	4.8	–
Separate 3U	–	–	–	14.8

Table 4: Total parameter counts for all model variants (in millions) across bimodal and trimodal configurations. Total parameters include transformer layers, modality-specific preprocessing, and projection heads. See Table 1 for transformer parameter counts.

	Text-Image		Text-Audio		Image-Audio	
Model	Fwd	Bwd	Fwd	Bwd	Fwd	Bwd
Bimodal models						
Shared 1U	32.7±0.6	32.0±0.3	76.9±1.1	77.8±1.4	0.7±0.1	1.0±0.1
Shared 2U	43.1±0.7	41.9±1.3	87.7±1.8	87.6±1.1	0.6±0.2	0.6±0.1
Shared 3U	45.6±0.5	44.7±0.4	89.5±1.4	90.2±0.8	0.7±0.2	0.6±0.3
Separate 2U	32.6±0.3	32.1±0.6	82.0±0.5	81.9±1.2	0.6±0.2	0.6±0.1
Trimodal models						
Shared 1U TIA	22.6±0.7	21.4±1.0	68.1±2.0	67.2±1.2	1.4±0.3	1.4±0.2
Shared 2U TIA	35.0±0.7	33.1±1.8	84.7±0.9	83.7±1.2	3.1±0.1	3.0±0.3
Shared 3U TIA	38.4±1.3	37.4±1.5	89.0±0.1	87.8±0.2	3.8±0.3	2.8±0.3
Separate 3U TIA	25.1±0.4	23.0±0.5	75.0±1.8	74.6±2.5	1.9±0.1	1.8±0.2

Table 5: Recall@5 (%), reported as mean ± std across three seeds. See Tables 2 and 3 for R@1 and MedR.

	Text-Image		Text-Audio		Image-Audio	
Model	Fwd	Bwd	Fwd	Bwd	Fwd	Bwd
Bimodal models						
Shared 1U	45.8±0.0	44.8±0.4	91.3±1.0	91.1±0.4	1.3±0.2	1.7±0.1
Shared 2U	55.3±0.2	56.5±0.9	96.1±0.2	95.5±0.6	1.2±0.2	1.1±0.1
Shared 3U	59.5±0.9	59.0±1.2	96.5±0.6	96.1±0.5	1.1±0.2	1.4±0.3
Separate 2U	45.4±0.8	45.0±0.8	93.3±0.3	92.8±0.5	1.1±0.2	1.2±0.3
Trimodal models						
Shared 1U TIA	35.0±1.9	33.7±0.8	84.2±1.6	84.0±0.1	3.1±0.4	3.1±0.3
Shared 2U TIA	47.3±1.2	47.4±0.8	93.6±0.5	93.7±0.2	5.9±0.3	5.5±0.7
Shared 3U TIA	51.1±1.0	51.7±0.5	96.0±0.3	95.6±0.4	6.5±0.4	5.2±0.4
Separate 3U TIA	37.6±1.2	34.9±1.0	88.5±1.5	88.1±0.6	3.8±0.2	3.1±0.5

Table 6: Recall@10 (%), reported as mean ± std across three seeds. See Tables 2 and 3 for R@1 and MedR.

	Text-Image		Text-Audio		Image-Audio	
Model	Fwd	Bwd	Fwd	Bwd	Fwd	Bwd
Bimodal models						
Shared 1U	0.222±0.007	0.217±0.004	0.504±0.002	0.512±0.007	0.009±0.001	0.012±0.000
Shared 2U	0.291±0.005	0.283±0.007	0.642±0.022	0.632±0.022	0.009±0.001	0.008±0.001
Shared 3U	0.311±0.005	0.303±0.005	0.679±0.013	0.670±0.016	0.009±0.001	0.009±0.001
Separate 2U	0.223±0.001	0.223±0.004	0.561±0.005	0.554±0.007	0.009±0.001	0.008±0.002
Trimodal models						
Shared 1U TIA	0.158±0.006	0.153±0.009	0.460±0.019	0.431±0.008	0.016±0.001	0.018±0.002
Shared 2U TIA	0.234±0.005	0.230±0.007	0.630±0.014	0.606±0.014	0.028±0.001	0.028±0.002
Shared 3U TIA	0.256±0.008	0.248±0.010	0.682±0.010	0.652±0.007	0.034±0.003	0.027±0.001
Separate 3U TIA	0.171±0.007	0.165±0.008	0.512±0.018	0.493±0.019	0.021±0.001	0.018±0.002

Table 7: Mean Reciprocal Rank (MRR), reported as mean ± std across three seeds.

	Text-Image		Text-Audio		Image-Audio	
Model	Fwd	Bwd	Fwd	Bwd	Fwd	Bwd
Bimodal models						
Shared 1U	26.3±0.6	25.8±0.3	59.8±0.2	60.4±0.6	0.6±0.1	0.9±0.0
Shared 2U	34.0±0.4	33.8±0.8	71.8±1.7	70.9±1.8	0.6±0.1	0.5±0.1
Shared 3U	36.7±0.6	35.9±0.6	74.8±1.2	74.0±1.4	0.5±0.1	0.6±0.1
Separate 2U	26.3±0.2	26.2±0.5	64.8±0.3	64.2±0.6	0.5±0.1	0.5±0.2
Trimodal models						
Shared 1U TIA	18.7±1.0	18.0±0.9	54.5±1.9	52.2±0.7	1.4±0.2	1.5±0.2
Shared 2U TIA	27.7±0.6	27.4±0.8	70.2±1.2	68.4±1.1	2.7±0.1	2.6±0.2
Shared 3U TIA	30.3±0.9	29.8±0.9	74.9±0.9	72.5±0.6	3.3±0.3	2.5±0.2
Separate 3U TIA	20.5±0.7	19.2±0.6	59.7±1.8	58.2±1.6	1.8±0.1	1.4±0.2

Table 8: NDCG@10 (%), reported as mean ± std across three seeds.

Modality pair	Shared	Separate
Bimodal at matched capacity (2U, 4.2M)		
Text–Image	0.81 ± 0.07	0.83 ± 0.14
Text–Audio	0.46 ± 0.07	0.44 ± 0.03
Image–Audio	-0.01 ± 0.01	-0.01 ± 0.01
Trimodal at matched capacity (3U, 6.3M)		
Text–Image–Audio	0.59 ± 0.05	0.42 ± 0.07

Table 9: Generalization gap (val loss − train loss) at the best checkpoint, reported as mean ± std across three seeds. In the trimodal setting (3U), the shared gap is 40% larger yet retrieval is better (Table 2), pointing to a representational rather than regularization advantage.

Loss pair	Cosine similarity
$\mathcal{L}_{\text{TI}}$ vs. $\mathcal{L}_{\text{TA}}$	-0.090 ± 0.016
$\mathcal{L}_{\text{TI}}$ vs. $\mathcal{L}_{\text{IA}}$	$+0.224 \pm 0.013$
$\mathcal{L}_{\text{TA}}$ vs. $\mathcal{L}_{\text{IA}}$	$+0.166 \pm 0.023$

Table 10: Pairwise cosine similarity between per-loss gradient vectors on the shared transformer parameters of the SharedTIA 3U model, averaged over 4 test batches per seed and reported as mean ± std across three seeds. $\mathcal{L}_{\text{TI}}$ and $\mathcal{L}_{\text{TA}}$ show mild negative alignment, indicating modest competition for shared capacity. The positive $\mathcal{L}_{\text{TI}}/\mathcal{L}_{\text{IA}}$ and $\mathcal{L}_{\text{TA}}/\mathcal{L}_{\text{IA}}$ cosines rule out direct gradient conflict as the primary cause of text–image degradation, and the pattern is better explained as a capacity-splitting effect where fixed parameters must serve three objectives simultaneously.

Model	Aniso. (image)	Aniso. (text/audio)	Uniform. (image)	Uniform. (text/audio)	Alignment
Text–Image					
Shared 1U	0.068 ± 0.004	0.106 ± 0.005	-3.213 ± 0.011	-3.057 ± 0.025	0.885 ± 0.008
Shared 2U	0.052 ± 0.001	0.089 ± 0.001	-3.320 ± 0.006	-3.162 ± 0.013	0.855 ± 0.003
Shared 3U	0.052 ± 0.004	0.081 ± 0.007	-3.339 ± 0.022	-3.199 ± 0.040	0.850 ± 0.005
Image–Audio					
Shared 1U	0.953 ± 0.008	0.956 ± 0.002	-0.158 ± 0.025	-0.150 ± 0.005	1.584 ± 0.018
Shared 2U	0.961 ± 0.007	0.982 ± 0.008	-0.136 ± 0.019	-0.064 ± 0.027	1.811 ± 0.084
Shared 3U	0.983 ± 0.006	0.988 ± 0.009	-0.066 ± 0.022	-0.044 ± 0.032	1.846 ± 0.028

Table 11: Embedding-space diagnostics on the test set, reported as mean ± std across three seeds. Anisotropy (Ethayarajh, 2019) is near 0 for isotropic embeddings and near 1 for collapsed ones. Uniformity (Wang & Isola, 2020) is more negative for better-spread embeddings. Alignment is the mean squared ℓ_2 distance between matched cross-modal pairs. Image–audio models exhibit near-degenerate anisotropy, near-zero uniformity, and alignment distances 1.8–2.2× those of text–image models.