Project Report : CS 7643
Multimodal Representation Learning for Semantic Image Retrieval

Spencer Uresk **Yasir Salman** **Ebrahim Pichka** **Felipe Oliveira**
Georgia Institute of Technology
{suresk, ysalman6, epichka3, foliveira8}@gatech.edu

Abstract

We conduct a systematic investigation of design choices in training contrastive vision-language models using the MS-COCO dataset. Our study examines backbone architectures (ResNet vs. Vision Transformer), contrastive loss formulations (CLIP loss, semi-hard negative mining, and SigLIP sigmoid loss), training strategies (full fine-tuning vs. frozen backbones), and parameter-efficient fine-tuning through LoRA adapters. We evaluate models on downstream Semantic Image Retrieval task using Recall@K metrics for semantic image retrieval and visualize embedding spaces through t-SNE projections. Our results also reveal that parameter-efficient methods achieve competitive performance while substantially reducing computational requirements, and that loss function selection determines semantic alignment quality but the effect of it on small-scale datasets are minimal.

The codebase for the project can be found at: *github.com/ebrahimpichka/DL-project*

1. Introduction

1.1. Background and Motivation

Multimodal vision-language representation learning learns shared embedding spaces that align visual and textual concepts. Large-scale contrastive models like CLIP [13] and ALIGN [8] have demonstrated remarkable zero-shot capabilities, but their computational demands and the uncertainty about which design decisions most impact performance create barriers for resource-constrained researchers.

In this project, we conduct a systematic investigation of design choices in contrastive vision-language models trained on the MS-COCO dataset [12]. Our dual-encoder architecture follows the CLIP framework (Figure 2), where separate image and text encoders map their inputs into a shared embedding space through contrastive learning. The practical motivation for this study stems from a critical limitation in current image-text retrieval systems: they often fail to generalize beyond the exact wording of training captions. For instance, querying the COCO dataset explorer for "car" returns over 12,000 images, while semantically equivalent terms like "automobile," "vehicle," or "driver" return zero results. This brittleness suggests that many systems memorize caption wording rather than learning true semantic relationships, limiting their practical utility for natural

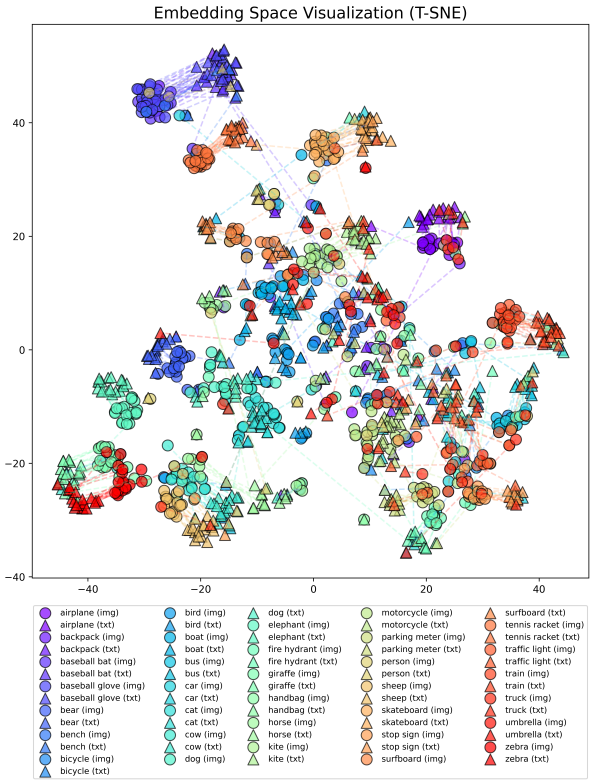

Figure 1: t-SNE plot of our best-performing text-image embedding space on a subset of COCO categories (20 samples per each). Circles/triangles denote image/text embeddings. Dashed lines connect matched image-text pairs. The visualization demonstrates strong semantic grouping where embeddings of the same category group together **regardless of modality**. This model awas trained using `ViT-L/16` and `BERT-base` as the image and text encoder.

language image search.

1.2. Related Work and Research Gap

Current practice in multimodal retrieval is dominated by **large-scale** contrastive models that require massive computational resources and web-scale datasets. The original CLIP [13] trained on **400 mil image-text pairs**, while ALIGN [8] used 1.8 billion noisy web-scraped pairs. Subsequent architectures have explored various improvements: BLIP [11] introduced caption filtering and bootstrapping techniques, CoCa [16] unified contrastive and generative objectives, and BLIP-2 [10] demonstrated efficient training through frozen pretrained encoders. Loss function design has also evolved, with SigLIP [17] showing that sigmoid-based losses enable training with larger batch sizes compared to the standard softmax contrastive loss.

For semantic image retrieval specifically, VSE++ [4] pioneered the use of hard negative mining to improve ranking performance, while cross-attention approaches [9] introduced fine-grained region-word alignment. Recent work has further refined alignment strategies: ALIP [15] leverages synthetic captions with adaptive weighting to mitigate noisy data, HiCLIP [6] incorporates hierarchy-aware attention to capture semantic hierarchies, and SoftCLIP [5] relaxes CLIP's rigid one-to-one constraint for more flexible cross-modal alignment. More recently, RankCLIP [18] extends beyond pairwise matching by introducing list-wise ranking consistency that exploits many-to-many relationships within and across modalities. The underlying contrastive learning framework builds upon self-supervised methods like SimCLR [1] and MoCo [2], which established principles about augmentation strategies and large-batch training.

Despite these advances, a critical research gap remains: while we know that massive data scale improves performance, we lack systematic understanding of which architectural and training decisions most impact embedding quality when working with moderate-scale datasets and computational budgets. Current models either demand web-scale data or fail on synonym-based retrieval when trained on academic-scale datasets. This creates practical challenges for researchers and practitioners who need strong semantic generalization without massive computational resources.

1.3. Research Questions and Contributions

This work addresses the following research questions through controlled experimentation on MS-COCO:

1. **Backbone Architecture:** How do different visual encoders (ResNet vs. Vision Transformer) affect embedding quality and training dynamics?

2. **Loss Function Design:** What is the comparative impact of standard CLIP contrastive loss, semi-hard negative mining (VSE++), and sigmoid-based loss (SigLIP) on retrieval performance?

3. **Training Strategy:** How does full fine-tuning compare to freezing pretrained backbones in terms of both performance and computational efficiency?

4. **Parameter-Efficient Fine-Tuning:** Can LoRA adapters ranks achieve competitive performance while substantially reducing trainable parameters?

Our contributions include: (1) A systematic empirical study comparing architectural and training choices under controlled conditions, (2) quantitative evaluation using standard retrieval metrics (Recall@K) on held-out test data, (3) qualitative analysis through t-SNE visualization of learned embedding spaces, and (4) practical insights into which design decisions matter most for semantic image retrieval when computational resources are constrained. Unlike prior work that focuses on scaling to ever-larger datasets, we investigate how to maximize embedding quality within realistic resource constraints.

1.4. Dataset

We use the MS-COCO (Common Objects in Context) dataset [12] as our primary benchmark. COCO contains over 330,000 images spanning diverse everyday scenes, with each image annotated with five human-written captions. This dataset has become a standard benchmark for vision-language tasks and has been used to evaluate models ranging from VSE++ [4] to CLIP [13] and BLIP [11]. The dataset's scale is large enough to train meaningful representations while remaining computationally tractable for academic research. We split the validation set into separate validation and test subsets to evaluate retrieval performance and identify overfitting during hyperparameter selection.

2. Approach

Our approach follows the dual-encoder paradigm established by CLIP [13], where separate image and text encoders learn to map their respective inputs into a shared embedding space through contrastive learning.

2.1. Model Architecture

Our model consists of an image encoder f_I, a text encoder f_T, and projection heads mapping encoder outputs to a shared embedding space. For the image encoder, we experiment with ResNet-50 [7] and Vision Transformers (`ViT-Base/16`, `ViT-Large/16` [14]). The text encoder uses `BERT-base-uncased` [3] throughout all experiments. Both encoders leverage pretrained weights from their respective domains (ImageNet for vision, large text corpora for language).

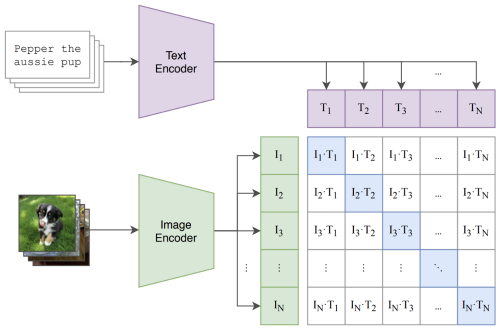

Figure 2: Contrastive pre-training approach. Image and text encoders are jointly trained to maximize similarity for matched pairs (diagonal, blue) while minimizing similarity for unmatched pairs. The learned embeddings enable semantic alignment between visual and linguistic concepts. Image from [13].

Each projection head consists of a two-layer network with ReLU activation, dropout, residual connections, and layer normalization: $\mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{emb}} \to \mathbb{R}^{d_{emb}}$. We set $d_{emb} = 512$ and vary dropout between 0 and 0.2. A learnable temperature parameter τ (initialized to $\log(1/0.07)$) scales similarity scores and is updated during training.

2.2. Contrastive Learning Objectives

Let $I = \{I_1, \ldots, I_N\}$ and $T = \{T_1, \ldots, T_N\}$ denote a batch of N matched image-text pairs. After encoding and projection, we obtain normalized embeddings $\mathbf{v}_i^I$ and $\mathbf{v}_j^T$, and compute similarities $S_{ij} = \mathbf{v}_i^I \cdot \mathbf{v}_j^T / \tau$.

CLIP Contrastive Loss. The standard loss uses symmetric cross-entropy:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{e^{S_{ii}}}{\sum_{j=1}^{N} e^{S_{ij}}} - \log \frac{e^{S_{ii}}}{\sum_{j=1}^{N} e^{S_{ji}}} \right]$$

This treats each batch as multi-class classification where the model identifies correct matches among all pairs.

Semi-Hard Negative Mining. Following VSE++ [4], we focus on challenging negatives within margin m of the positive: $\mathcal{N}(i) = \{j : S_{ij} > S_{ii} - m, j \neq i\}$. The loss becomes:

$$\mathcal{L}_{\text{sh}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{e^{S_{ii}}}{\sum_{j \in \mathcal{N}(i)} e^{S_{ij}}} - \log \frac{e^{S_{ii}}}{\sum_{j \in \mathcal{N}(i)} e^{S_{ji}}} \right]$$

We set $m = 0.2$ and fall back to all negatives when no semi-hard negatives exist.

SigLIP Sigmoid Loss. The SigLIP loss [17] treats each pair as binary classification with $S_{ij} = \mathbf{v}_i^I \cdot \mathbf{v}_j^T \cdot \tau + b$ (bias b initialized to -10.0):

$$\mathcal{L}_{\text{sig}} = -\frac{1}{N^2} \sum_{i,j} \left[y_{ij} \log \sigma(S_{ij}) + (1 - y_{ij}) \log(1 - \sigma(S_{ij})) \right]$$

where $y_{ij} = 1$ if $i = j$ and 0 otherwise. This avoids batch-wide normalization, enabling larger-scale training.

3. Experiments and Results

Our experimental approach systematically investigates key factors affecting multimodal representation learning for semantic image retrieval through five complementary experiments tracked using Weights & Biases (W&B).

The core implementation follows a dual-encoder architecture where separate image and text encoders project inputs into a shared embedding space. The image encoder can be ResNet or Vision Transformer, while the text encoder uses BERT. Each encoder is followed by a projection head: a two-layer network with ReLU activation, dropout, residual connections, and layer normalization. A learnable temperature parameter, initialized to $\log(1/0.07)$, scales similarity scores between embeddings.

For parameter-efficient fine-tuning, we employed LoRA, which inserts trainable low-rank matrices into attention layers rather than updating entire backbones. We tested ranks of 4, 8, 16, and 32. We implemented three loss functions: standard CLIP contrastive loss (symmetric cross-entropy), SigLIP sigmoid loss (independent pair treatment), and semi-hard negative mining (VSE++ style, focusing on challenging negatives).

3.1. Measuring Success

We evaluated models using bidirectional retrieval metrics namely Recall@K and Mean/Median Rank. **Recall@K** (K=1,5,10) measures whether relevant items appear in top K results given a query for both image-to-text (I2T) and text-to-image (T2I). A higher Recall indicates a better performance.

$$\text{Recall@K} = \frac{\text{Number of relevant items in the top K results}}{\text{Total number of relevant items}}$$

Mean/Median Rank indicates the average position of the first relevant item and **Mean Recall** averages Recall@K values per direction. A lower Mean/Median Rank indicates a better performance. We mainly report and focus on the T2I figures as an evaluation metric, as it corresponds to the semantic image retrieval task.

We also qualitatively and visually evaluate the top-4 retrieval results for a group of random text queries. The text queries are designed such that they capture models capability in two senses: (1) being able to generalize over synonym concepts and retrieve same images for different synonym words, and (2) being able to differentiate between images with the same item but within different context.

3.2. Experiment 1: Image Backbone Comparison

This experiment compares ResNet-50 (convolutional) against Vision Transformers (ViT-B/16, ViT-L/16) used

Table 1: Image Backbone Comparison (Full Fine-tuning). All models trained for 5 epochs with batch size 64 on MS-COCO using BERT-base text encoder, 512-dim embeddings, and CLIP loss.

Backbone	T2I R@1	T2I R@5	T2I R@10	Mean Rank
ResNet-50	8.32	31.33	44.30	63.42
ViT-B/16	9.29	33.61	47.45	50.75
ViT-L/16	**11.04**	**38.95**	**53.11**	**43.41**

as pre-trained image encoder backbones to understand whether transformer-based vision models produce embeddings that generalize better for multimodal representations and semantic retrieval. The hypothesis is that ViT may yield richer global embeddings improving zero-shot synonym retrieval, while ResNet offers more stability and efficiency.

We tested three backbones with all other factors fixed: `BERT-base-uncased` encoder, 512-dim embedding space, CLIP contrastive loss, and full fine-tuning and low-rank adaptation. All models trained for 5 epochs with batch size 64, learning rate 10^{-5}, and mixed precision. We measured retrieval performance, computational costs (training time, GPU memory), and training dynamics, selecting the best validation checkpoint for test evaluation.

Table 1 presents retrieval performance across three image encoder architectures. Vision Transformers substantially outperform ResNet-50: ViT-L/16 achieves 11.04% R@1, 38.95% R@5, and 53.11% R@10—improvements of 32.7%, 24.3%, and 19.9% respectively. ViT-B/16 shows intermediate performance, confirming that both architectural paradigm and model capacity contribute to gains.

Figure 3 reveals critical training dynamics. ResNet-50 begins with higher initial loss (4.82 vs. 4.1 for ViTs), suggesting pretrained transformer features provide better initialization. However, all models converge to similar final losses (0.30-0.35) after 50,000 steps. This convergence despite divergent validation performance indicates that equivalent training loss does not guarantee equivalent representation quality, a key finding for model selection.

Validation metrics in Figure 4 show all models performing identically until step 40,000, when clear separation emerges. ViT-L/16 maintains consistent advantage throughout remaining training, with gaps persisting rather than closing. This sustained hierarchy suggests fundamental capacity differences rather than optimization artifacts.

3.3. Experiment 2: Full Fine-tuning vs Frozen Backbone

This experiment assesses the performance-efficiency trade-off between full fine-tuning and frozen-backbone training, where only projection layers is updated. The hypothesis is that full fine-tuning will outperform frozen training but at significantly higher computational cost.

Using ViT-B/16 from Experiment 1, we compared both strategies with identical settings (CLIP loss, batch size 64,

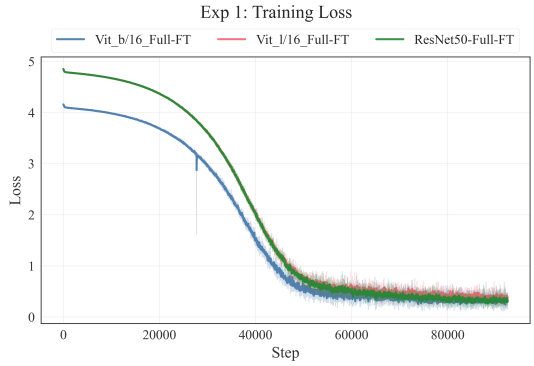

Figure 3: Training loss curves over 5 epochs (∼92,000 steps). All models reach similar final loss despite divergent validation performance, revealing that training loss alone is insufficient for evaluating representation quality.

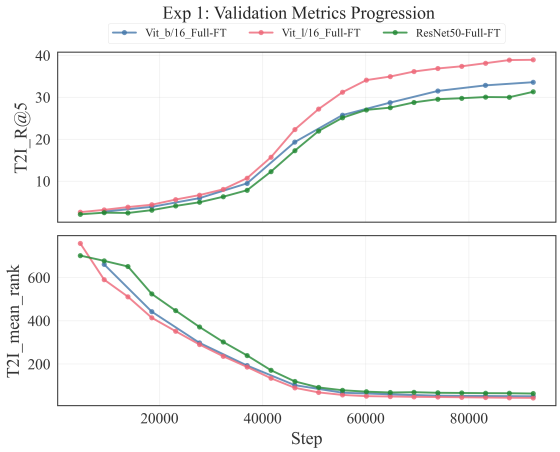

Figure 4: Text-to-image retrieval performance throughout training. Architectural hierarchy emerges after step 40,000 and persists, indicating fundamental capacity differences.

5 epochs) but different learning rates: 10^{-3} for frozen (2-3M trainable params) and 10^{-5} for full fine-tuning (150M params). We measured retrieval performance and computational costs, presenting results as performance-vs-cost curves.

Table 2 reveals frozen backbone training catastrophically fails on MS-COCO. ViT-B/16 frozen achieves near-zero performance (0.01% R@1) due to unstable training and collapsing during training, while ResNet-50 frozen reaches only 2.83% which is 66% worse than full fine-tuning. Full fine-tuning delivers strong results: ViT-B/16 at 9.29% R@1, ResNet-50 at 8.32% R@1.

Frozen models achieve lower training loss (∼0.2 vs. ∼0.35 for fine-tuned) yet dramatically worse validation performance (Figure 5), revealing mismatch between pretrained features and contrastive objectives. Validation progression (Figure 6) shows frozen models stagnate or fail while fine-tuned models steadily improve.

Table 2: Training Strategy Comparison. Frozen trains only projections (2-3M params); full fine-tuning updates all weights (150M params). Both use batch size 64, CLIP loss, 5 epochs.

Strategy	T2I R@1	T2I R@5	T2I R@10	Mean Rank
Frozen Backbones				
ResNet-50	2.83	12.05	19.24	208.00
ViT-B/16	0.01	0.04	0.08	6254.00
Full Fine-tuning				
ResNet-50	8.32	31.33	44.30	63.42
ViT-B/16	**9.29**	**33.61**	**47.45**	**50.75**

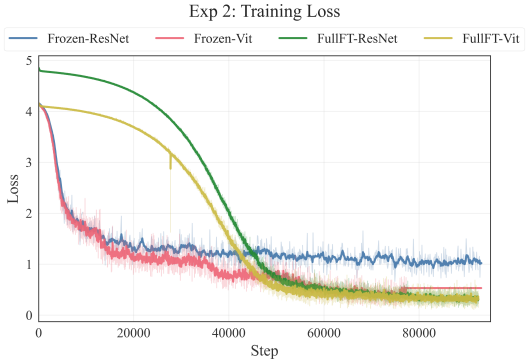

Figure 5: Training loss comparison across models using full fine-tuning or frozen backbones.

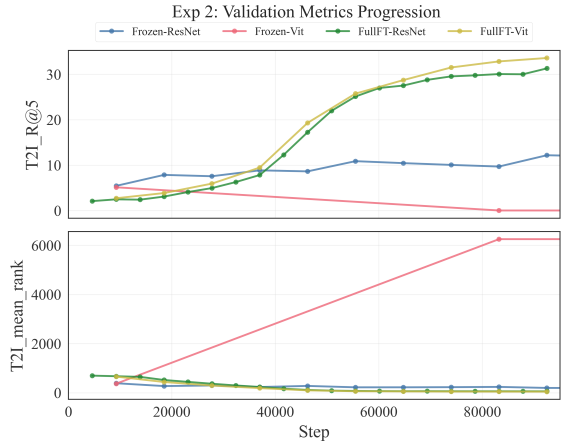

Figure 6: Validation progression showing frozen backbones stagnate while full fine-tuning improves steadily.

3.4. Experiment 3: LoRA Rank Exploration

This experiment investigates whether LoRA at various ranks (4, 8, 16, 32) can approach full fine-tuning performance while maintaining parameter efficiency. The hypothesis is that moderate ranks (8-16) achieve strong performance with dramatic parameter reduction, while very low ranks (4) may underfit and very high ranks (32) provide diminishing returns.

Using ViT-B/16 and BERT, we applied LoRA to attention layers (*Query* and *Value* projections) with alpha equal to rank and 0.2 dropout. All configurations used batch size 64, CLIP loss, and 10 epochs. We measured retrieval metrics, trainable parameters (count and percentage), and computational costs, constructing performance-vs-cost curves and comparing against full fine-tuning and frozen-backbone results.

Table 3: LoRA Rank Ablation on ViT-B/16. All models use batch size 64, CLIP loss, 10 epochs. Performance saturates beyond $r = 8$.

LoRA Rank	T2I R@1	T2I R@5	T2I R@10	Mean Rank
$r = 4$	8.81	32.41	45.91	52.44
$r = 8$	9.29	33.61	47.45	50.75
$r = 16$	**9.38**	**33.75**	**47.89**	**50.04**
$r = 32$	9.22	33.93	47.66	52.28

Table 4: Parameter Efficiency. LoRA achieves full fine-tuning performance with $<2\%$ trainable parameters.

Config	Total (M)	Train. (M)	Train. %	R@1
$r = 4$	197.48	1.61	0.82	8.81
$r = 8$	197.78	1.90	0.96	9.29
$r = 16$	198.37	2.49	1.26	**9.38**
$r = 32$	199.55	3.67	1.84	9.22
Full FT	198.37	198.37	100.0	9.29

Table 3 reveals flat performance across LoRA ranks 4-32, with all configurations achieving 8.81-9.38% R@1. The optimal $r = 16$ (9.38% R@1) matches full fine-tuning (9.29%) while using only 1.26% trainable parameters (Table 4).

3.5. Experiment 4: Loss Function Comparison

This experiment compares three contrastive objectives: standard CLIP loss (symmetric cross-entropy), SigLIP (sigmoid-based, independent pairs), and semi-hard negative mining (VSE++, focusing on challenging negatives). The hypothesis is that CLIP loss achieves highest overall performance, semi-hard mining improves fine-grained ranking, and SigLIP provides most stable optimization.

To isolate loss effects, we used ResNet-50 with frozen backbone across all three conditions, training only projection layers. All experiments used batch size 64, 5 epochs, and learning rate 10^{-3}. For semi-hard loss, we set margin to 0.2; for SigLIP, bias initialized to -10.0. We evaluated Recall@K, mean rank, and training stability.

Table 5 compares three contrastive objectives using frozen ResNet-50 to isolate loss effects. All perform similarly poor (2.3-2.8% R@1), with CLIP loss marginally best.

Training dynamics (Figure 7) show SigLIP converging to higher loss ($\sim$1.9) despite comparable validation performance, again demonstrating training-validation disconnect with frozen encoders. Validation progression (Figure 8) reveals highly unstable learning across all losses, with erratic oscillations throughout training. All converge

Table 5: Loss Function Comparison with frozen ResNet-50. All losses fail similarly due to frozen backbone bottleneck. Batch size 64, 5 epochs.

Loss Function	T2I R@1	T2I R@5	T2I R@10	Mean Rank
SigLIP [17]	2.39	10.27	16.58	245.02
Hard Negative [4]	2.71	11.93	19.56	207.31
Contrastive [13]	**2.83**	**12.05**	**19.24**	**208.00**

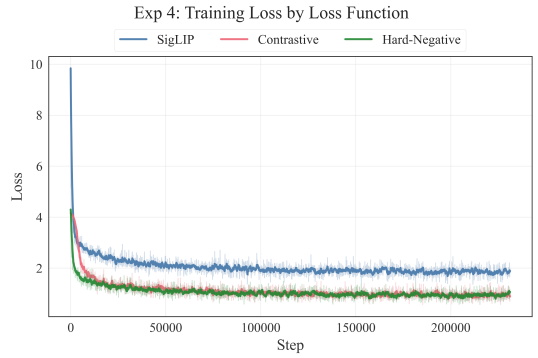

Figure 7: Training loss trajectories. SigLIP reaches higher loss but comparable validation performance.

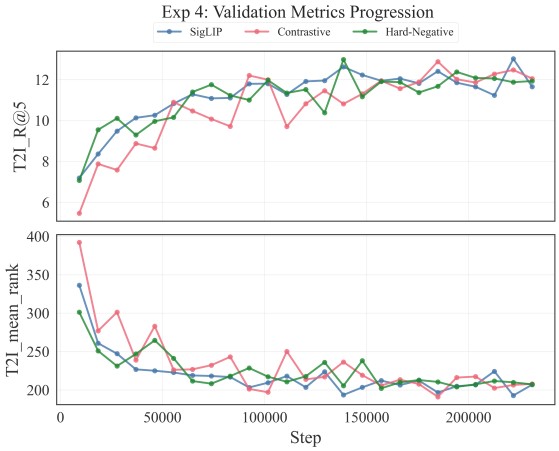

Figure 8: Validation progression showing rather unstable but increasingly improving training. Instability indicates frozen encoders prevent stable learning.

to ∼12% R@5 but maintain significant noise, contrasting sharply with smooth curves from fine-tuned experiments.

3.6. Qualitative Retrieval Analysis

We conducted qualitative evaluation of our best model's (`ViT-L/16+BERT`) retrieval through *synonym matching* and *concept differentiation* tests. For synonym matching across five concept groups (Aircraft, Cat, Dog, Television, Vehicle), the model achieves strong semantic clustering for common synonyms (0.80+ similarity) with consistent retrieval, but struggles with low-frequency terms and regional colloquialisms. For concept differentiation across four attribute types (Size, Texture, Action, Style), the model successfully differentiates actions through salient visual features but shows limited capability for fine-grained attribute distinctions like material texture or relative size. **Detailed analysis with retrieval visualizations and similarity matrices is provided in Appendices A and B.**

4. Discussion

We conducted this study to understand which design decisions matter most when training multimodal image-text retrieval models under realistic limited resource constraints. Using the MS-COCO dataset, we trained CLIP-style dual encoders and systematically tested the effect of hyperparameters such as image backbone, fine-tuning strategy, and loss function to isolate their impact on semantic retrieval. Our approach allowed us to evaluate how each choice affects both performance and computational efficiency, providing practical guidance for building models that generalize well without requiring large-scale computation.

Backbone architecture was the single largest contributor to performance differences. Vision Transformers significantly outperformed ResNet-50 on every retrieval metric when trained under identical conditions. Their global self-attention provides stronger semantic representations and better alignment with text. All models reached nearly identical training losses, meaning that validation metrics are most important for predicting representation quality.

Training strategy had the second-largest impact. **Full fine-tuning** of encoders was essential as frozen backboned had extremely poor performance for ViT and weak results for ResNet, regardless of loss function. Training only the projection layer could not adapt pretrained features to the contrastive objective.

LoRA fine-tuning provided the most effective middle ground. Low-rank adaptation allowed the model to update only 1–2% of parameters while matching or slightly exceeding the performance of full fine-tuning. Ranks between 8 and 16 captured nearly all variation, meaning that the trainable parameters are a low-dimensional subspace.

Loss function choice mattered far less than training strategy. With frozen encoders, CLIP loss, SigLIP, and semi-hard negative mining all performed similarly poorly because of the encoder bottleneck. While CLIP loss remained slightly better overall, the results show that loss refinements may only help once the model has enough capacity to learn meaningful representations.

Putting all together, the most effective model configuration is one that uses a ViT image encoder, LoRA fine-tuning of Rank 16, and CLIP contrastive loss. This model produced strong semantic grouping where embeddings of the same category group together regardless of modality (Figure 1). These findings help clarify which design decisions matter most when building well-performing multimodal retrieval models under resource-constrained conditions.

Student Name	Contributed Aspects	Details
Spencer Uresk	Model Architecture	Built the foundation for image and text encoders and projections to the shared space. Allowed for backbone and hyperparameter swapping for experimentation.
Yasir Salman	Data Loading and Preprocessing	Pulled training and validation images from COCO dataset and split into Train/Val/Test Dataloaders. Normalized and augmented images for stable training.
Ebrahim Pichka	Training and Evaluation	Created the training script for the model. Implemented the loss types and metrics that were used for evaluation.
Felipe Oliveira	Experimentation and Results	Built the evaluation procedure to compare training and validation performances. Compiled the performance outputs from the hyperparameter tuning into interpretable results.

Table 6: Contributions of team members.

5. Work Division

Our team divided the project into broad roles to keep the workflow organized. Spencer focused on the model architecture, Yasir handled data loading and preprocessing, Ebrahim worked on training and evaluation, and Felipe managed experimentation and results. Even though we each had a main area of focus, our different computational limitations meant the roles weren't completely rigid. Whenever someone finished their part or couldn't run certain experiments, they helped teammates with theirs, whether that meant debugging code, running additional tests, or reviewing outputs. This flexible setup allowed us to keep all parts of the project moving in parallel and made the final outcome feel like a collaborative effort.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 2

[4] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2018. 2, 3, 6

[5] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: softer cross-modal alignment makes clip stronger. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. 2

[6] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[9] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 2

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[12] 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*, pages 740–755. Springer, 2014. 1, 2

[13] 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*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6

[14] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 2

[15] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023. 2

[16] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. In *Transactions on Machine Learning Research*, 2022. 2

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, pages 11975–11986, 2023. 2, 3, 6

[18] Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zhili Feng, Zenghui Ding, and Yining Sun. Rankclip: Ranking-consistent language-image pretraining, 2025. 2

A. Qualitative Analysis: Synonym Matching Capability

To evaluate whether our trained model captures true semantic relationships rather than memorizing exact caption wording, we designed a synonym matching test. This evaluation addresses the practical limitation highlighted in our introduction: conventional retrieval systems fail when query terms differ from training captions, even when semantically equivalent. We constructed five concept groups where each contains 3-4 synonym queries referring to the same underlying concept, then measured both text embedding similarity and retrieval consistency.

A.1. Text Embedding Analysis

Figures 9 through 13 display cosine similarity matrices between synonym text embeddings for five concept groups: Aircraft, Cat, Dog, Television, and Vehicle. The matrices reveal strong semantic clustering with similarity scores consistently above 0.80 for within-concept comparisons. For the Aircraft group (Figure 9), queries "A plane," "An airplane," "An aircraft," and "A jet" achieve pairwise similarities ranging from 0.91 to 0.98, indicating the text encoder successfully maps these lexically distinct but semantically equivalent terms into nearby regions of embedding space.

The Cat group (Figure 10) exhibits particularly interesting patterns. While "cat," "kitty," and "kitten" form a tight cluster with similarities of 0.84-0.88, the term "feline" shows lower similarity (0.56-0.64) to other variants. This

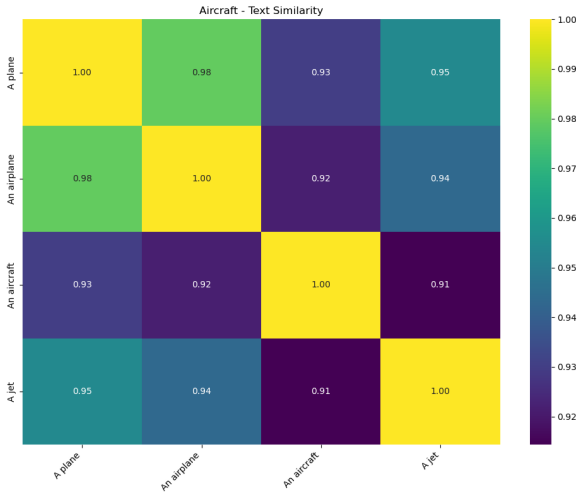

Figure 9: Text similarity matrix for Aircraft synonyms. High similarity scores (0.91-0.98) demonstrate semantic clustering of lexically distinct but conceptually equivalent terms.

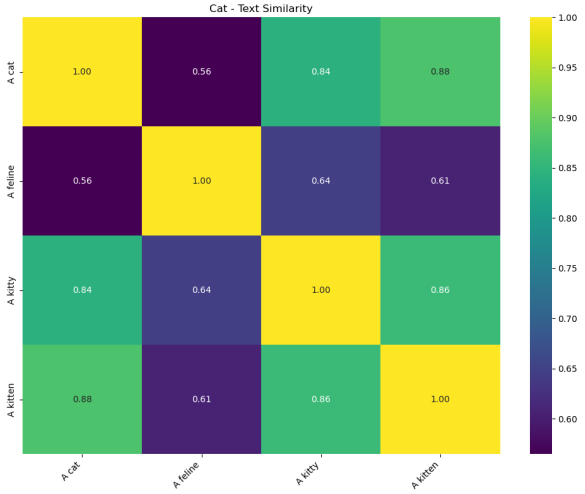

Figure 10: Text similarity matrix for Cat synonyms showing strong alignment between "cat," "feline," "kitty," and "kitten" (0.56-0.88).

suggests the model has learned stronger associations between colloquial terms ("cat," "kitty") that likely co-occur more frequently in natural language captions, while the more formal term "feline" occupies a slightly different semantic neighborhood. Despite this variation, all terms remain sufficiently close to enable effective retrieval.

The Dog group (Figure 11) mirrors this pattern with "dog," "pup," and "puppy" forming a coherent cluster (0.74-0.88), while "canine" maintains moderate distance (0.55-0.84). For Television (Figure 12), "tv" and "television" achieve near-perfect alignment (0.99), reflecting their frequent co-occurrence as abbreviation and full form. How-

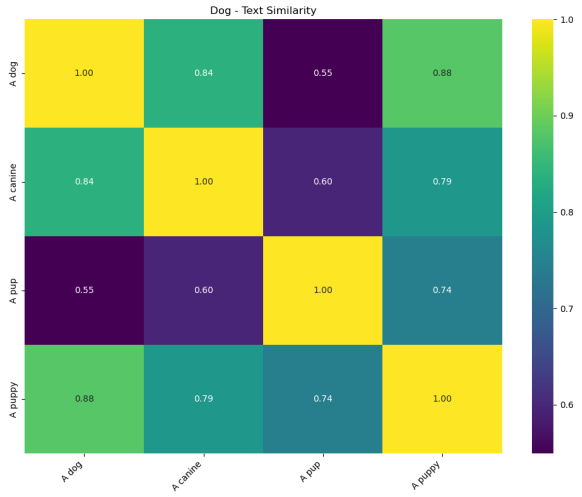

Figure 11: Text similarity matrix for Dog synonyms. The model learns strong associations between informal terms ("dog," "puppy": 0.88) while formal term "canine" shows moderate separation (0.55-0.84).

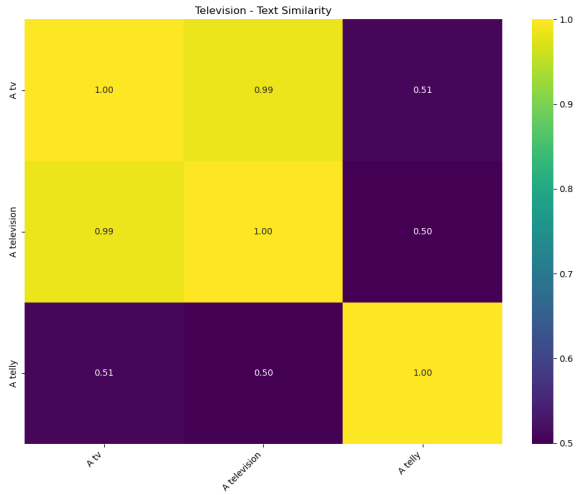

Figure 12: Text similarity matrix for Television synonyms demonstrating near-perfect alignment between "tv" and "television" (0.99) while colloquial "telly" shows distinct separation (0.50-0.51).

ever, the British colloquialism "telly" shows marked separation (0.50-0.51), likely due to its geographic and cultural specificity reducing training data occurrence.

The Vehicle group (Figure 13) demonstrates hierarchical semantic structure. Generic terms "car," "automobile," and "vehicle" cluster tightly (0.87-0.88), while the more specific "motorcar" exhibits lower similarity (0.56-0.81). This hierarchy suggests the model has learned conceptual abstraction levels, where broader category terms occupy overlapping regions while specialized variants maintain slight separation.

Figure 13: Text similarity matrix for Vehicle synonyms. Generic terms ("car," "automobile," "vehicle") cluster tightly (0.87-0.88) while specific term "motorcar" shows lower similarity (0.56-0.81).

A.2. Retrieval Consistency Analysis

Figures 14 through 18 present top-4 retrieval results for each synonym query, revealing whether embedding similarity translates to consistent image retrieval. The Aircraft queries (Figure 14) demonstrate strong retrieval consistency: all four synonym terms retrieve highly overlapping sets of aircraft images with similarity scores ranging from 0.525 to 0.587. Notably, retrieved images span diverse aircraft types (commercial jets, military fighters, propeller planes, cargo aircraft), indicating the model generalizes beyond specific visual attributes to capture the abstract aircraft concept.

For the Cat group (Figure 15), queries "cat," "kitty," and "kitten" reliably retrieve domestic cat images with scores of 0.511-0.565. However, "feline" produces mixed results, including giraffe images among top retrievals. This aligns with the text similarity analysis showing "feline" as an outlier, confirming that the biological taxonomy term activates a broader semantic region encompassing various feline-like animals rather than specifically domestic cats. This illustrates a known challenge in contrastive learning: overgeneralization to superordinate categories when training data lacks sufficient specific examples.

The Dog retrievals (Figure 16) show excellent consistency across all synonym queries, with scores ranging from 0.490 to 0.625. Retrieved images span various dog breeds, sizes, and contexts (indoor/outdoor), demonstrating the model's ability to abstract beyond superficial visual features. Even the formal term "canine," despite lower text similarity scores, successfully retrieves relevant dog im-

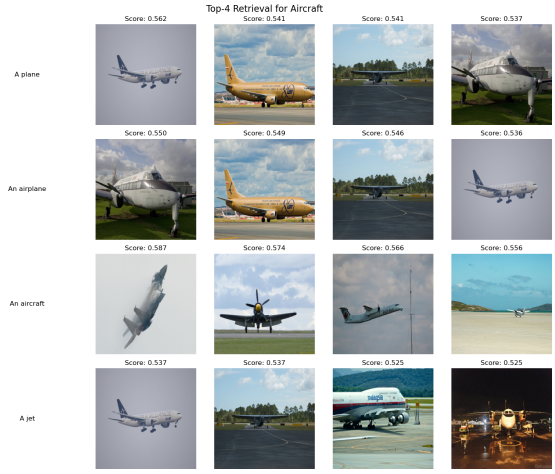

Figure 14: Top-4 image retrieval for Aircraft synonyms. Consistent retrieval across all queries demonstrates robust semantic generalization beyond exact wording.

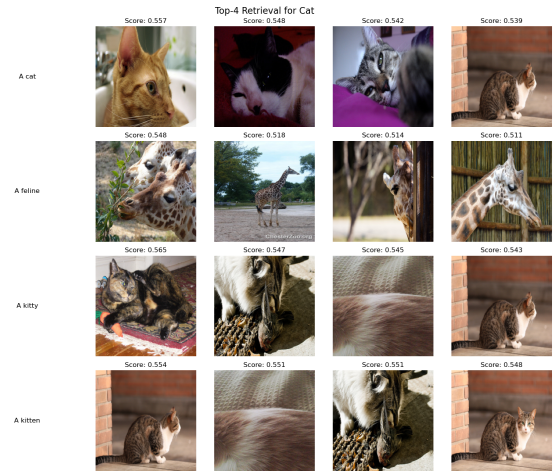

Figure 15: Top-4 image retrieval for Cat synonyms showing consistent cat image retrieval despite query "feline" returning some giraffe images, revealing partial semantic confusion.

ages, suggesting the retrieval robustness benefits from the broader semantic neighborhood learned during training.

Television queries (Figure 17) reveal both success and failure modes. "Tv" and "television" produce nearly identical retrievals (scores 0.622-0.666) with consistent television images across different settings. However, "telly" fails significantly, retrieving microwave ovens instead of televisions. This failure correlates with the low text similarity (0.50) and likely stems from insufficient training examples containing the colloquial term, causing the model to associate it with visually similar appliances rather than semantic equivalents.

The Vehicle group (Figure 18) demonstrates semantic

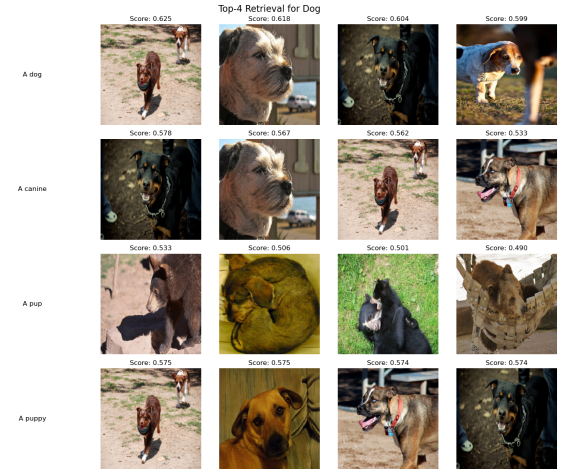

Figure 16: Top-4 image retrieval for Dog synonyms. All queries successfully retrieve dog images (0.490-0.625) with diverse breeds and contexts.

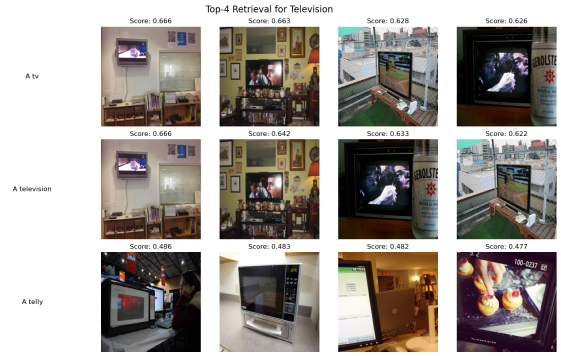

Figure 17: Top-4 image retrieval for Television synonyms. "Tv" and "television" retrieve identical results (0.622-0.666) while "telly" shows confusion with microwave images.

hierarchy in action. "Car," "automobile," and "vehicle" consistently retrieve automobile images with high scores (0.553-0.617), spanning sedans, trucks, and service vehicles. However, "motorcar" produces motorcycle images instead, revealing fine-grained semantic confusion between compositionally similar terms. This suggests the model has learned morphological associations ("motor + car") that occasionally override distributional semantics when training data is sparse for archaic terminology.

A.3. Findings and Implications

The synonym matching analysis reveals both the capabilities and limitations of our contrastive learning approach. The model successfully learns semantic equivalence for common synonym groups, achieving high embedding similarity (0.80+) and consistent retrieval across lexically distinct queries. This demonstrates that contrastive training

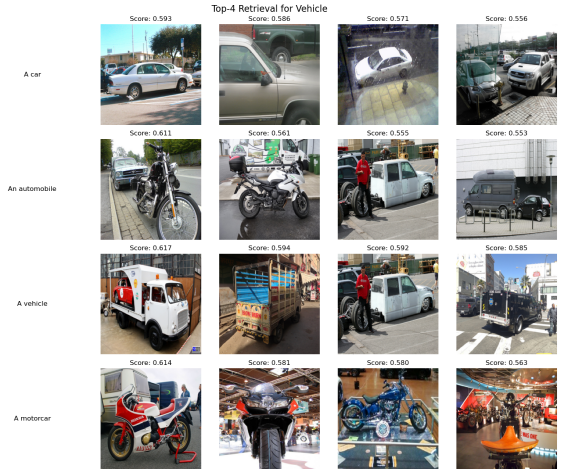

Figure 18: Top-4 image retrieval for Vehicle synonyms showing strong retrieval for "car" queries (0.553-0.617) but "motorcar" retrieves motorcycles, revealing fine-grained semantic confusion.

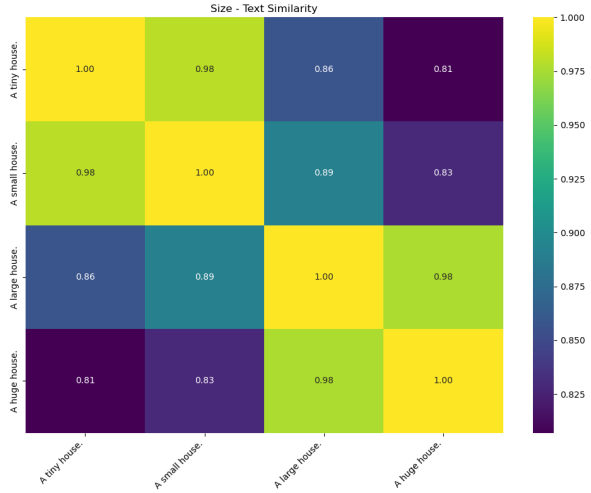

Figure 19: Text similarity matrix for Size descriptors showing gradient structure with high within-cluster similarity and moderate between-cluster separation.

on MS-COCO's caption diversity enables generalization beyond exact wording to conceptual understanding.

However, several failure patterns emerge. Terms with lower frequency in natural language ("feline," "telly," "motorcar") show reduced semantic alignment and retrieval accuracy. This indicates that contrastive learning's effectiveness depends critically on term co-occurrence patterns in training data rather than learning abstract lexical relationships. The model struggles with formal/technical terms, regional colloquialisms, and archaic vocabulary that appear less frequently in modern web-sourced captions.

Additionally, the model occasionally overgeneralizes to superordinate categories ("feline" retrieving giraffes) or confuses compositionally similar terms ("motorcar" retrieving motorcycles). These errors suggest limitations in handling hierarchical taxonomies and morphological reasoning, areas where contrastive learning without explicit linguistic supervision remains weak.

For practical deployment, these findings suggest that synonym-robust retrieval requires either massive training data covering diverse vocabulary, or hybrid approaches incorporating explicit synonym expansion at query time. The current model handles common informal vocabulary well but requires query preprocessing (mapping "telly" $\rightarrow$ "television") for robust handling of specialized, regional, or archaic terms.

B. Qualitative Analysis: Concept Differentiation Capability

While synonym matching evaluates semantic clustering, concept differentiation assesses whether the model learns fine-grained distinctions between related but distinct concepts. We designed four differentiation tests examining attributes (Size, Texture), actions (Action), and styles (Style), where lexical similarity is high but semantic distinctions are critical for accurate retrieval.

B.1. Size Differentiation

Figure 19 presents the text similarity matrix for size descriptors applied to houses: "tiny house," "small house," "large house," and "huge house." The matrix reveals a gradient structure with high similarity between adjacent size terms (0.83-0.98) and moderate similarity between extreme pairs (0.81-0.86). This gradient indicates the model has learned relative rather than absolute size representations, where "tiny" and "small" occupy nearby regions, as do "large" and "huge," while the two clusters maintain separation.

The retrieval results (Figure 20) demonstrate partial success in translating this gradient to visual differentiation. "Tiny house" retrieves compact interior spaces and small residential structures (scores 0.565-0.593), successfully capturing the compact scale concept. "Small house" maintains this trend with single-story residences and modest buildings (0.564-0.581). However, "large house" and "huge house" show considerable overlap with "small house" retrievals rather than consistently retrieving mansions or large structures (0.594-0.631).

This mixed performance reveals a fundamental challenge in grounding relative attributes through contrastive learning. While the model captures ordinal relationships in text space, translating these to visual scale requires understanding perspective, context, and comparative reference frames. Training captions rarely provide explicit size com-

Figure 20: Top-4 image retrieval for Size descriptors. Model successfully differentiates "tiny" but struggles to consistently retrieve progressively larger structures for "large" and "huge."'

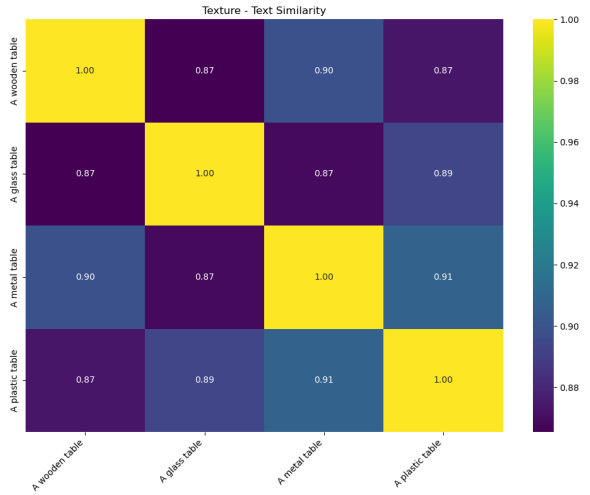

Figure 21: Text similarity matrix for Texture descriptors showing strong clustering around shared "table" concept (0.87-0.91) with limited material differentiation.

parisons ("this house is larger than that one"), forcing the model to learn size concepts from distributional patterns alone, which proves insufficient for consistent differentiation.

B.2. Texture Differentiation

Figure 21 shows text similarity for material descriptors: "wooden table," "glass table," "metal table," and "plastic table." The matrix exhibits moderate similarity (0.87-0.91) among all terms, suggesting the model groups them primarily by the shared "table" concept rather than material properties. This clustering indicates that compositional structure ("[material] + [object]") creates high baseline similarity, potentially masking fine-grained material distinctions.

Retrieval results (Figure 22) confirm this limitation. All texture queries successfully retrieve table images (scores 0.562-0.646), demonstrating robust object recognition. However, material consistency is limited: "wooden table" retrievals include various furniture pieces, "glass table" returns mixed dining and decorative tables, "metal table" shows office furniture and equipment, and "plastic table" retrieves diverse table types without clear material specificity.

This pattern reveals that contrastive learning prioritizes object category ("table") over attributes ("wooden") when both appear in composite descriptions. During training, images with captions like "wooden table" are contrasted primarily against images of different objects ("chair," "lamp"), with insufficient negative examples of "glass table" or "metal table" in the same batch. Consequently, the model learns strong category boundaries but weak within-category attribute differentiation, a known limitation of batch-based contrastive learning without hard negative mining.

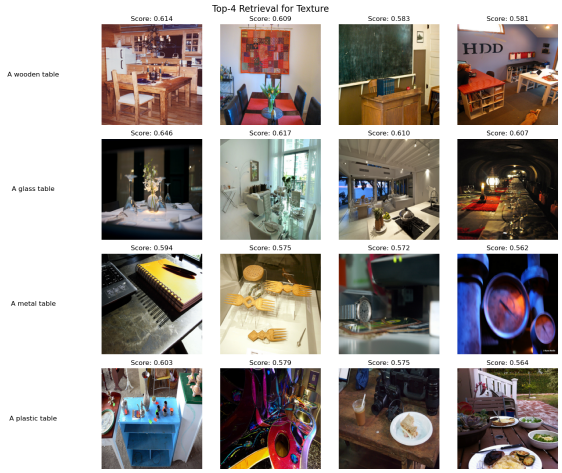

Figure 22: Top-4 image retrieval for Texture descriptors. Model successfully identifies table objects but shows limited material specificity in top retrievals.

B.3. Action Differentiation

Figure 23 presents similarity for action descriptors in identical contexts: "person running in a field," "person sitting in a field," "person jumping in a field," and "person walking in a field." The matrix shows clear action-based clustering: "running" and "walking" are highly similar (0.92), "sitting" groups separately (0.67-0.88), and "jumping" occupies an intermediate position (0.76-0.86).

Retrieval analysis (Figure 24) demonstrates excellent action differentiation. "Person running in a field" retrieves images showing running motion with high scores (0.600-0.682). "Person sitting in a field" consistently returns seated

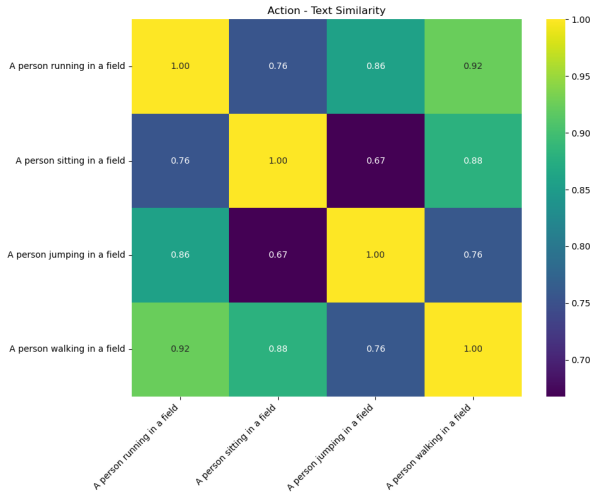

Figure 23: Text similarity matrix for Action descriptors showing clear clustering based on action dynamics: locomotion (running/walking: 0.92) vs. static posture (sitting).

Figure 24: Top-4 image retrieval for Action descriptors. Model successfully differentiates action types, retrieving contextually appropriate human poses and motions.

poses in outdoor settings (0.608-0.625). "Person jumping in a field" successfully captures mid-air poses (0.591-0.604), and "person walking in a field" retrieves walking motion frames (0.642-0.697).

This success indicates that action verbs create strong semantic gradients in the learned embedding space, likely because actions manifest as distinctive visual features (body pose, motion blur, spatial positioning) that contrastive learning can exploit. Unlike abstract attributes (size, material), actions produce salient visual differences that align naturally with caption distinctions, enabling the model to learn robust action differentiation from the MS-COCO training distribution.

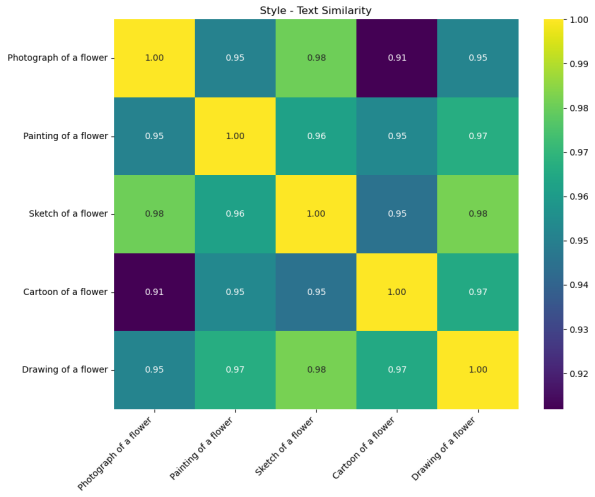

Figure 25: Text similarity matrix for Style descriptors showing very high similarity (0.91-0.98) with limited differentiation between artistic mediums.

B.4. Style Differentiation

Figure 25 shows text similarity for artistic styles applied to flowers: "photograph of a flower," "painting of a flower," "sketch of a flower," "cartoon of a flower," and "drawing of a flower." The matrix reveals high similarity across all style terms (0.91-0.98), suggesting the model primarily groups them by content ("flower") rather than artistic medium. The distinction between drawing-related terms ("sketch," "drawing": 0.98) and rendered styles ("painting," "cartoon": 0.95-0.97) is subtle.

Retrieval results (Figure 26) show mixed style differentiation. "Photograph of a flower" successfully retrieves photographic flower images with consistent realism (scores 0.600-0.623), indicating the model distinguishes photographic rendering from other media. However, "painting," "sketch," "cartoon," and "drawing" show significant overlap in retrievals (0.552-0.617), with most returning photographic or naturalistic flower images rather than distinctively stylized content.

This limitation reflects MS-COCO's inherent bias toward photographic images. The dataset contains primarily real-world photographs with few artistic renderings, stylized graphics, or illustrations. Without sufficient training examples of paintings, sketches, or cartoons, the model cannot learn distinctive visual features associated with these style terms. Consequently, style descriptors behave primarily as semantic null operators, leaving the content term ("flower") as the dominant retrieval signal.

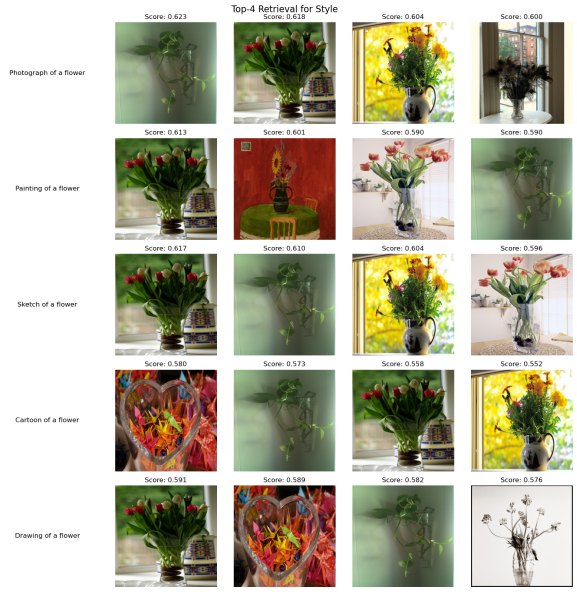

Figure 26: Top-4 image retrieval for Style descriptors. Model shows partial success: "photograph" reliably retrieves realistic images (0.600-0.623) while other styles show mixed media types.

B.5. Findings and Implications

The concept differentiation analysis reveals domain-specific strengths and weaknesses in the model's ability to learn fine-grained distinctions. Action differentiation succeeds because actions manifest as salient visual features (body pose, motion) that align naturally with linguistic distinctions. The model learns robust mappings between action verbs and corresponding visual patterns, achieving high retrieval accuracy.

In contrast, attribute differentiation (size, material, style) proves challenging. Size requires understanding relative scale and perspective, materials demand recognizing subtle visual textures, and styles presuppose familiarity with artistic conventions—all requiring richer inductive biases than contrastive learning provides. The model learns object categories robustly but treats compositional attributes as secondary, often ignored signals.

Dataset composition critically constrains differentiation capability. MS-COCO's photographic bias prevents learning style distinctions, its diverse scales complicate size learning, and material co-occurrence patterns are insufficient for texture differentiation. These limitations suggest that achieving human-level fine-grained understanding requires either massive-scale training with balanced attribute distributions, explicit compositional reasoning mechanisms, or hybrid approaches incorporating structured attribute representations.

For practical applications, these findings recommend against relying solely on contrastive models for fine-grained attribute-based search. Queries involving materials, sizes, or artistic styles require fallback strategies: query expansion ("large house" $\rightarrow$ "mansion"), explicit attribute filters (post-hoc size filtering), or hybrid systems combining embedding similarity with metadata-based constraints. Actions and object categories remain the model's strength, suitable for direct semantic retrieval.