A Systematic Analysis of the gLocal Transformation for Human-Model Alignment

Jasper Valk

Layout: typeset by the author using LaTeX.
Cover illustration: Jasper Valk

A Systematic Analysis of the gLocal Transformation for Human-Model Alignment

Jasper Valk
1266402

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 900
1098 XH Amsterdam

Supervisor
Dr. N.J.E. van Noord

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 900
1098 XH Amsterdam

Semester 2, 2024-2025

Abstract

This thesis presents a systematic analysis of the robustness and practical consequences of human-model alignment in vision-language models using the gLocal transformation. The gLocal method is designed as a post-hoc linear mapping, a transformation applied to pretrained model embeddings after the original training process. This gLocal transformation aligns pretrained CLIP ViT-L/14 model embeddings with human similarity judgments, based on large-scale triplet annotations from the THINGS dataset. Experiments investigate the impact of annotation scarcity and label noise by varying the amount and quality of human supervision. Downstream few-shot classification performance is evaluated across several datasets, under a range of alignment and control conditions. Results show that reducing the number of annotated triplets to as little as 0.01% leads to only minimal loss in few-shot accuracy, and that random or trivial transformations often perform comparably to explicit alignment in these settings. However, analysis of the resulting embedding spaces demonstrates that gLocal alignment reliably increases the similarity between model and human judgments, as measured by representational metrics. These changes indicate a shift toward more human-like conceptual organization in the internal structure of the model, even when downstream task gains are minimal. Comprehensive sanity checks and reproduction of prior results confirm the reliability of the experimental pipeline. The findings suggest that they contribute to improved interpretability and cognitive plausibility of neural representations, particularly when measured with respect to human-likeness in representational space.

Contents

Chapter 1

Introduction

Deep neural networks have achieved strong performance across a range of visual recognition tasks. Models such as CLIP (Contrastive Language–Image Pretraining) (Radford et al. 2021), trained on large-scale image collections and associated metadata, demonstrate impressive generalization to new datasets without task-specific fine-tuning. These models serve as a basis for current research in visual representation learning.

Despite these strengths, such models often fall short in a key dimension: alignment with human perception. Representations learned through contrastive objectives are optimized to distinguish between images based on textual context, but do not necessarily capture the ways in which humans naturally compare or group visual concepts. This gap is particularly problematic for tasks such as few-shot learning, where effective generalization relies on meaningful and human-aligned representations. Recent studies have shown that contrastive objectives can lead to representational gaps between model embeddings and human similarity judgments, reflecting more fundamental differences in how models and humans perceive similarity and category structure (Muttenthaler, Greff, et al. 2024; Fahim, Murphy, and Fyshe 2024). For instance, models trained on image-text pairs may excel at broad semantic distinctions but often fail to capture fine-grained or perceptual similarities that are intuitive for humans. These misalignments raise concerns for applications that require interpretable or cognitively grounded representations.

A prominent approach to addressing this gap is to directly align model representations with human behavioral data. Triplet comparisons, where annotators judge whether object A is more similar to B or to C, provide an effective and scalable form of supervision, offering insight into human similarity judgments. Large-scale datasets such as THINGS (Hebart et al. 2019) supply the basis for studying representational alignment at scale.

The gLocal transformation, introduced by Muttenthaler, Linhardt, et al. (2023),

builds on this approach by learning a lightweight post-hoc mapping of pretrained model embeddings. It is designed to optimize agreement with human triplet judgments while preserving the global semantic structure of the original embedding space. This dual-objective loss, regulated by a tunable parameter α, is intended to balance local alignment and global consistency. The initial results reported improvements in both alignment accuracy and downstream task performance.

Despite promising results, the robustness and reliability of alignment techniques under realistic conditions remain largely unexplored. In practice, collecting high-quality human annotations is costly and time-consuming; labels are often limited, incomplete, or noisy. In addition, alignment pipelines are often assumed to function correctly, with little attention paid to systematic validation or sanity checks that could reveal subtle implementation errors or misinterpretations.

This thesis addresses these limitations by critically evaluating the claims and practical limits of human-model alignment via the gLocal transformation of Muttenthaler, Linhardt, et al. 2023. The main contribution lies in systematically evaluating the robustness of alignment methods and validating whether the claimed benefits persist under real-world constraints. The analysis focuses on scenarios with limited or noisy human similarity data and introduces a comprehensive framework of sanity checks to validate both the pipeline and the alignment process. This approach provides methodological precision and aims to clarify whether observed improvements reflect genuine human alignment or result from experimental errors.

A series of experiments evaluates pretrained vision-language models, named CLIP ViT-L/14, for alignment with human judgments across several datasets. Experiments systematically vary the quantity and quality of triplet annotations, introduce controlled label noise, and assess downstream generalization in few-shot classification settings. In addition, a suite of sanity checks including identity, random, null, and destructive transformation, as well as intentionally incorrect triplet supervision is implemented to ensure the integrity and interpretability of the alignment pipeline. The central research question in this thesis is: *How robust is human-model representation alignment when using limited or noisy human similarity data, and how can systematic sanity checks improve the reliability of alignment research?* This main question is addressed through two subquestions: *How robust is the gLocal transformation when trained with limited quantities of triplet data or when triplet labels contain noise?* and *How can systematic sanity checks be used to assess both pipeline reliability and the robustness of vision-language models to different types of transformations?*

By systematically probing both the practical limits and the methodological reliability of gLocal-based alignment, this work aims to clarify the amount and quality of human supervision required for effective alignment but also to identify which imperfections can be tolerated.

Chapter 2

Background and Related Work

This chapter discusses the theoretical and empirical foundation of human-model alignment in vision-language models. This chapter first reviews vision-language architectures, then examines techniques for aligning model representations with human perception, emphasizing the gLocal transformation approach. The chapter ends by exploring the current challenges and unanswered questions in matching neural representations with human perception.

2.1 Introduction to Vision-Language Models

Vision-language models (VLMs) have shown great progress in artificial intelligence by supporting integrated reasoning over visual and textual information. The primary motivation for VLMs is to close the representational gap between how machines process images and how humans describe visual content using language (Radford et al. 2021). Early methods to working with multimodal learning often combined separately trained vision and language models, resulting in limited cross-modal reasoning capacity (Karpathy and Fei-Fei 2015). Recent work in deep learning and the availability of large-scale multimodal datasets have facilitated the development of models that learn joint visual-textual representations that show more flexible and robust transfer between tasks (Xu et al. 2015).

Modern VLMs typically employ two main components: an image encoder and a text encoder. The image encoder is based on convolutional neural networks (CNNs) or increasingly, vision transformers or ViTs, which are used in this work. ViTs offer enhanced scalability and capture global image structure more effectively than CNNs, which contributes to their success in the best VLMs. The text en-

coder, generally implemented as a transformer-based language model, maps text into a high-dimensional embedding space. Both encoders are trained such that semantically similar images and texts are mapped to nearby locations within a shared embedding space. This is illustrated in Figure 2.1. Semantic similarity is typically measured using cosine similarity:

$$s(x, y) = \frac{x \cdot y}{\|x\|\|y\|}. \tag{2.1}$$

This shared embedding space enables zero-shot image classification, cross-modal retrieval, and other multimodal tasks (Radford et al. 2021).

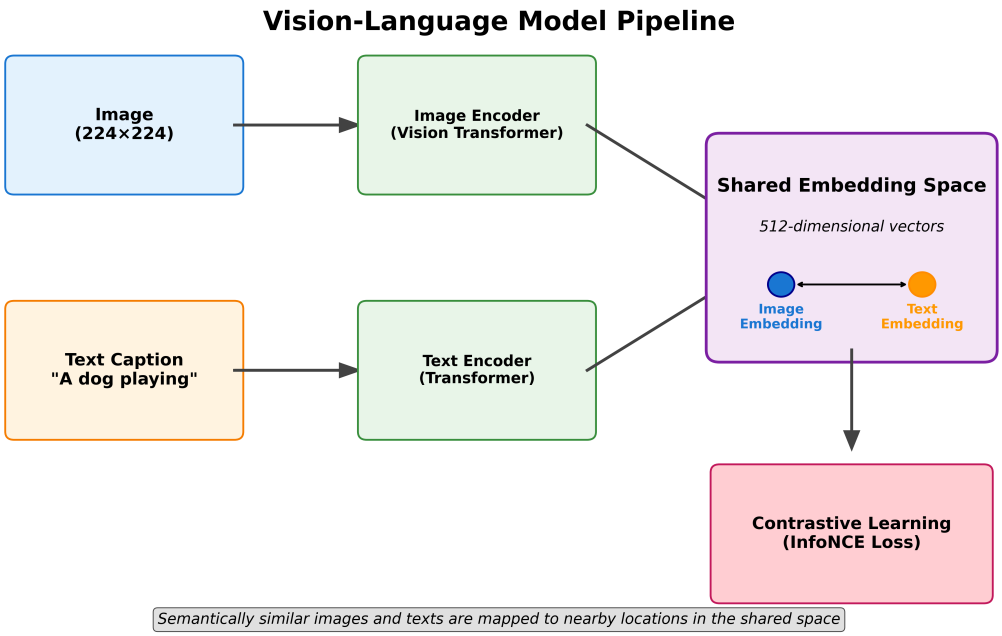


Figure 2.1: A typical vision-language model (VLM) pipeline. The model encodes an image and a text prompt into a shared embedding space, where similarity is computed with cosine similarity.

CLIP (Contrastive Language–Image Pretraining), introduced by Radford et al. (ibid.), employs large-scale contrastive learning in image-caption pairs and achieves robust generalization in diverse tasks. OpenCLIP extends this approach to even larger and more diverse datasets to further increase robustness. These models show that pre-training with natural language supervision eliminates the requirement for manual annotation, resulting in robust zero-shot performance (Kaiser et al. 2017).

Current VLMs remain misaligned with human perceptual and conceptual similarity in several aspects. Although these models capture broad semantic rela-

tionships, they often do not reflect the semantic ways in which humans group or compare visual concepts (Geirhos et al. 2020). This limitation is especially relevant for tasks that require interpretable or cognitively grounded representations, such as few-shot learning or human-in-the-loop systems. Improving alignment with human similarity judgments remains an important challenge for the development of trustworthy and human-centered AI systems.

2.2 Contrastive Learning and Representations

Contrastive learning is the foundational training strategy for most VLMs. The objective is to learn embeddings such that paired and semantically similar inputs, like images and their captions are mapped closer together, while unpaired or dissimilar examples are pushed apart. This is done through the InfoNCE loss (Oord, Li, and Vinyals 2018; Chen et al. 2020):

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(x, x^+))}{\sum_{x^-} \exp(\text{sim}(x, x^-))}, \tag{2.2}$$

where x^+ denotes the positive (paired) example and x^- runs over negatives in the batch. Cosine similarity is used for $\text{sim}(\cdot, \cdot)$. Because of avoiding reliance on explicit class labels contrastive learning supports highly scalable pretraining on diverse data sources (Radford et al. 2021).

Contrastive objectives have enabled the development of models that excel at capturing high-level category structure and achieve excellent performance in zero-shot transfer tasks. However, the supervision provided by contrastive learning is dominantly local: it establishes that only paired examples are close, without directly constraining the global structure of the embedding space (Muttenthaler, Linhardt, et al. 2023). As a result, representations may cluster high-level categories but fail to organize concepts in a way that is aligned with human semantics.

Empirical evaluations indicate that contrastive representations are less predictive of human similarity judgments for abstract or fine-grained comparisons than for broad and concrete categories (Sucholutsky et al. 2023; Hebart et al. 2019). For example, most VLMs can distinguish "dog" from "car" but often misrepresent the similarity between visually or conceptually related objects that humans would consider similar. This shortcoming has led to research on explicit alignment of neural representations with human perception.

2.3 Human Similarity Judgments and Triplet Data

Although VLMs capture many semantic relationships, they do not necessarily reflect human similarity perception. The THINGS dataset by Hebart et al. 2019

address this gap by collecting triplet judgments of the form: "Is image A more similar to image B or C?" This triplet data is straightforward for humans to provide and are robust to individual annotator noise. The accuracy of a model in predicting human triplet judgments, triplet or odd-one-out accuracy (OOO), is a standard metric for alignment:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left[\hat{y}_i = y_i^{\text{human}} \right], \tag{2.3}$$

where $\hat{y}_i$ is the model's predicted OOO and y_i^{human} is the human judgment. Despite their utility, triplet annotations are expensive to collect and typically sparse. This motivates the development of efficient alignment methods that can learn from limited or noisy human feedback.

2.4 Representation Alignment

Aligning model representations with human judgments is an active area of research. Initial approaches focused on fine-tuning model parameters with additional behavioral supervision, but these methods are computationally intensive and poorly scalable. Simpler approaches, such as linear probes or metric learning, attempt to project embeddings into spaces that are better aligned with human similarity, but these are often insufficient to capture the complexity of human perceptual spaces, especially under limited supervision (Sucholutsky et al. 2023).

A common metric for evaluating representational alignment is triplet accuracy: the proportion of triplet comparisons where the model agrees with human judgments. More advanced alignment strategies include post-hoc linear or shallow non-linear transformations that operate on frozen embeddings. This optimizes alignment without expensive retraining (Muttenthaler, Linhardt, et al. 2023).

2.5 The gLocal Transformation

The gLocal transformation, introduced by Muttenthaler, Linhardt, et al. (ibid.), is a lightweight, post-hoc linear mapping designed to align pretrained model embeddings with human similarity judgments (see Section 2.3). The key contribution of gLocal is its dual-objective loss, which balances local alignment to human judgments (via triplet loss) with global preservation of the semantic structure present in the original embedding space (via a contrastive loss over a large set of unlabeled images).

In intuitive terms, gLocal optimizes two goals at once: it encourages the embedding space to agree locally with human similarity judgments, while also preserving

the global organization of semantic relationships already present in the pretrained model.

These human similarity judgments are given in the form of triplets, sourced from the THINGS dataset (Hebart et al. 2019). In each triplet (A, B, C), object A is judged by human annotators to be more similar to B than to C. The method learns a linear transformation $T \in \mathbb{R}^{d \times d}$, which is applied to the original feature vectors $x \in \mathbb{R}^d$ extracted from pretrained models (in this study, $d = 768$ for OpenCLIP ViT-L/14). The goal is to produce transformed embeddings Tx that better reflect human-perceived similarities while preserving the global semantic structure of the original embedding space, shown in Figure 2.2.

Mathematical Formulation

The training objective of gLocal combines three components:

- **Triplet loss** ($\mathcal{L}_{\text{triplet}}$): Encourages the transformed embeddings to align with human triplet judgments. For a triplet (A, B, C), where A is judged more similar to B than to C, the loss is defined as

  $$\mathcal{L}_{\text{triplet}} = \max\left(0, \tau + s(Tx_A, Tx_B) - s(Tx_A, Tx_C)\right)$$

  where $s(\cdot, \cdot)$ denotes dot product similarity, that is, $s(x, y) = x^\top y$. This notation is used for similarity, not distance. τ is a margin hyperparameter (Muttenthaler, Linhardt, et al. 2023). Features are standardized to zero mean and unit variance per dimension, but not normalized to unit length.

  To quantify alignment between model and human similarity judgments, OOO accuracy is computed on these triplets. Given representations x_A, x_B, x_C for images A, B, C, cosine similarity is defined as

  $$c(x, y) = \frac{x \cdot y}{\|x\|_2 \|y\|_2}$$

  The predicted most similar pair in a triplet is

  $$\text{argmax}_{(i,j) \in \{(A,B),(A,C),(B,C)\}}\, c(x_i, x_j)$$

  OOO accuracy is the proportion of triplets for which the predicted least similar image matches the human judgment.

- **Contrastive loss** ($\mathcal{L}_{\text{contrastive}}$): Preserves the global structure of the embedding space by encouraging the transformed features to maintain their original relationships on a large set of unlabeled images. This is implemented as a

mean squared error (MSE) between pairwise dot product similarities in the original and transformed spaces:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{|P|} \sum_{(i,j)\in P} (s(Tx_i, Tx_j) - s(x_i, x_j))^2$$

where P is a set of image pairs sampled from the dataset.

- **Regularization** ($\mathcal{R}(T)$): A regularization term (Frobenius norm) to prevent overfitting and encourage smoothness in the transformation:

$$\mathcal{R}(T) = \|T\|_F^2$$

The total loss function is a weighted sum of these components:

$$\mathcal{L} = \mathcal{L}_{\text{triplet}} + \alpha \cdot \mathcal{L}_{\text{contrastive}} + \lambda \cdot \mathcal{R}(T)$$

where α and λ are hyperparameters controlling the influence of the contrastive loss and regularization, respectively. In this thesis, the gLocal transformation is applied to features extracted from OpenCLIP ViT-L/14 models, with all implementation and experimental details following the original approach of Muttenthaler, Linhardt, et al. (2023).

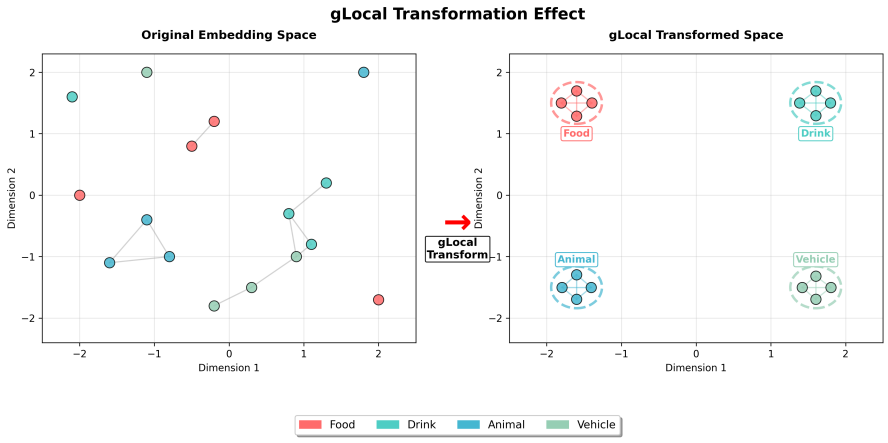


Figure 2.2: Conceptual illustration of the gLocal transformation. The original embedding space captures local structure but lacks global organization consistent with human semantics. The gLocal transform reorganizes clusters (e.g., food and drink) while maintaining local similarity. Adapted from Muttenthaler, Linhardt, et al. (2023).

2.6 Open Challenges

Several prior studies have explored representational alignment between neural models and human judgments. Early efforts focused on fine-tuning pretrained models using human-annotated supervision (Karpathy and Fei-Fei 2015), but these methods do not scale well to large architectures or diverse datasets. Metric learning and linear probing approaches offer computational efficiency but struggle to model the rich structure of human similarity spaces under realistic data constraints (Sucholutsky et al. 2023).

Recently, post-hoc alignment techniques have become more popular due to their flexibility and effectiveness. Lightweight linear or shallow nonlinear transformations can align model representations with human data without retraining the base model (Muttenthaler, Linhardt, et al. 2023). The gLocal approach balances the need for interpretability and generalization by integrating global semantic preservation with local triplet alignment.

However, several open challenges remain. Many existing studies assume access to large quantities of high-quality human annotations. This is an unrealistic view in most practical applications. Additionally, systematic validation of alignment methods through robustness analysis or diagnostic sanity checks has been largely overlooked in previous research. This questions methodological reliability and the risk of undetected implementation errors. Negative or unexpected results can provide useful insights into the practical limitations of alignment techniques.

This work addresses these issues by systematically evaluating the robustness of human-model alignment under constraints of limited and noisy supervision and by introducing a comprehensive framework of sanity checks and diagnostic controls. This approach ensures that observed alignment improvements reflect genuine changes in representational quality, rather than experimental errors.

Chapter 3

Methodology

This chapter describes the experimental framework for evaluating human-model alignment using the gLocal transformation, with a primary focus on methodological reliability. The approach extends standard alignment and classification analyses by systematically incorporating diagnostic sanity checks. Unlike other studies, which often assume correct pipeline behavior, this methodology explicitly tests whether observed effects are due to changes in representation quality or to implementation errors.

A complete set of control transformations was used to validate the pipeline's sensitivity. These sanity checks establish clear reference points. This ensures that the evaluation framework reliably shows meaningful improvements from noninformative or destructive changes.

All datasets, feature extraction procedures, and training protocols are selected for consistency with the prior work of Muttenthaler, Linhardt, et al. 2023. The following sections outline dataset selection, feature extraction, alignment procedures, downstream evaluation, and the central role of sanity checks.

3.1 Datasets

This study uses a set of datasets to evaluate the alignment between human and model representations and to assess downstream performance. All datasets are also used in the work of Muttenthaler, Linhardt, et al. (2023).
The following datasets are employed:

- **THINGS** (Hebart et al. 2019): Contains 1,854 concepts of natural objects, annotated with over one million human similarity triplets. These triplets served as the primary supervision signal for training the gLocal transformation.

- **CIFAR-100** (Krizhevsky, Hinton, et al. 2009): Comprises 100 fine-grained object classes, used for few-shot performance analysis.
- **CIFAR-100-coarse** (ibid.): Utilizes the coarse label split of CIFAR-100, grouping the original 100 object classes into 20 superclasses, used for few-shot performance analysis.
- **Describable Textures Dataset (DTD)** (Cimpoi et al. 2014): Consists of 47 texture categories and is used for few-shot performance analysis.
- **SUN397** (Xiao et al. 2010): Offers a diverse set of scene categories across 397 classes, used for few-shot performance analysis.
- **ImageNet**(Deng et al. 2009): Used as an unlabeled data source for computing the contrastive loss component of the gLocal training objective. ImageNet provides large-scale visual coverage and supports the preservation of global semantic structure in the embedding space.

These datasets were selected to investigate the effect of alignment across different visual domains and semantic granularities, ensuring that the results reflect both generalization and robustness.

3.1.1 Vision-Language Model Selection

All experimental evaluations are based on two large-scale vision-language models: `CLIP ViT-L/14 (LAION-400M)` (Schuhmann, Vencu, et al. 2021) and `CLIP ViT-L/14 (LAION-2B)` (Schuhmann, Beaumont, et al. 2022). The ViT-L/14 architecture has shown very good performance in tasks involving representational alignment and transfer.

Both models align image and text modalities through contrastive pre-training. They enable the extraction of dense semantically rich embeddings suitable for downstream analysis. The choice to restrict evaluation to these two variants is due to the fact that these models perform better than the other models that are used in the work of Muttenthaler, Linhardt, et al. 2023; Muttenthaler, Greff, et al. 2024. These studies have shown that these models outperform alternatives, particularly in settings with limited supervision and for post-hoc transformation tasks. In all following sections references to the selected vision-language models denote these two CLIP ViT-L/14 variants.

3.1.2 Input Resizing for Vision Transformers

All image datasets are preprocessed to match the input resolution expected by CLIP ViT-L/14 which operates on fixed-size 224×224 pixel inputs. While high-

resolution datasets like ImageNet and SUN397 already conform to this format, other datasets such as CIFAR-100, CIFAR-100-coarse, and DTD consist of low-resolution images (e.g., 32×32). These are up-sampled to 224×224 pixels prior to feature extraction. This step is necessary to avoid tensor shape mismatches in ViT-based models and to maintain compatibility with pretrained weights.

The resizing is integrated into a custom feature extraction pipeline, adapted from the original gLocal implementation. This modification allows the pipeline to operate consistently across datasets with varying native resolutions. This aligns with practices in previous work (Muttenthaler, Linhardt, et al. 2023).

3.2 Feature Extraction

Feature extraction forms the basis for all alignment and classification experiments in this study. All experiments use the two vision-language models introduced in Section 3.1.1. The selection of these models is motivated by previous work showing strong performance in representation alignment and downstream transfer tasks.

All images are resized to 224×224 pixels to match the input resolution required by the ViT-L/14 architecture. For datasets with lower native resolution, resizing is performed using bicubic interpolation. This approach produces smoother and more consistent images than simpler methods, ensuring compatibility with pretrained model weights and consistent feature extraction.

The features are extracted from the penultimate layer of the vision transformer, following established practice to evaluate the quality of the embedding (Radford et al. 2021; Tian et al. 2020). For each image i, the feature vector $x_i \in \mathbb{R}^{768}$ is normalized to unit length,

$$\tilde{x}_i = \frac{x_i}{\|x_i\|_2}$$

to ensure a consistent embedding space and to prevent scale-dependent effects in all analyses.

The extracted feature matrices are serialized and cached in `.pkl` format with consistent naming conventions. This facilitates reproducibility and computational efficiency across all experiments.

3.2.1 Implementation Adaptations

The main adaptation is the extension of the feature extraction pipeline to support low-resolution datasets, ensuring compatibility with vision transformer architectures. To enable compatibility with CIFAR-scale datasets, a uniform resizing step was added to 224×224 pixels prior to feature extraction. This adaptation pre-

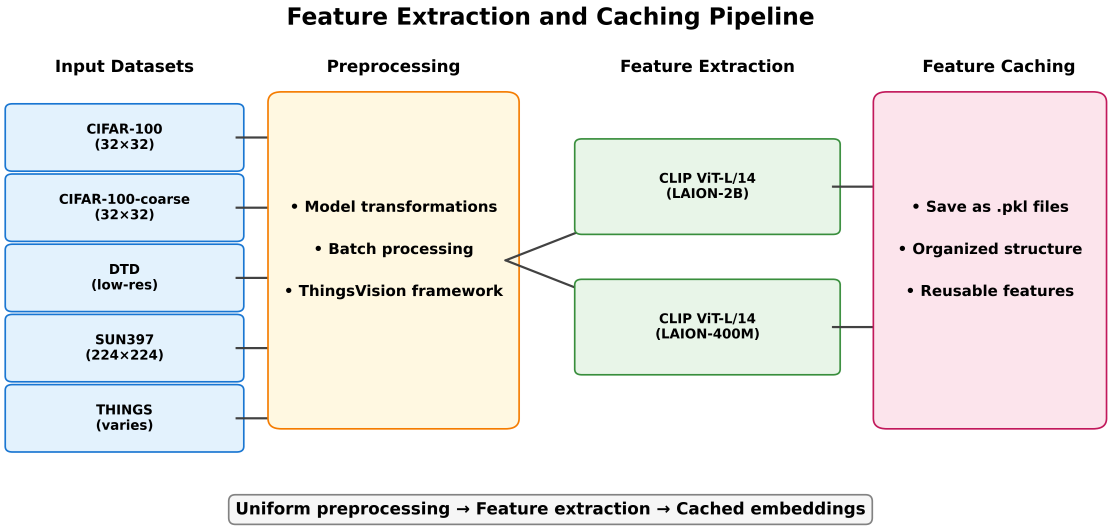


Figure 3.1: Schematic overview of the feature extraction and caching pipeline. Each dataset undergoes uniform preprocessing, model-specific feature extraction, and the resulting embeddings are cached for all downstream experiments.

vents tensor shape mismatches and supports the patch-based structure of vision transformers.

The feature extraction pipeline is adapted from the official implementation of gLocal by Muttenthaler, Linhardt, et al. 2023, with additional support for strict handling of input dimensions. All models are used with pretrained weights without further fine-tuning. The extraction script supports both model-specific configurations and output layers. A caching mechanism was implemented to avoid redundant computations and to enable seamless downstream use for alignment and classification tasks. ImageNet-1K is used only as an unlabeled dataset for the contrastive loss during gLocal training and is not included in downstream classification benchmarks. This maintains strict train-test separation and prevents information leakage.

3.3 Training the gLocal Transformation

The gLocal transformation was trained to align pretrained model embeddings with human similarity judgments, following the approach of Muttenthaler, Linhardt, et al. (2023). The transformation consists of a single linear mapping $T \in \mathbb{R}^{768\times768}$, optimized to increase agreement with supervised data while preserving local structure.

The training objective is a weighted sum of three components: the triplet alignment loss L_{triplet}, a contrastive loss $L_{\text{contrastive}}$ that preserves pairwise similarities among model features, and an ℓ_2 regularization term:

$$L = L_{\text{triplet}} + \alpha L_{\text{contrastive}} + \lambda \|T\|_F^2,$$

where α and λ are hyperparameters.

For most experiments, the transformation was trained using human-annotated triplets from the THINGS dataset. The triplet loss was computed over batches of annotated triplets, with margin parameter $\tau = 1.0$. The contrastive loss was implemented as mean squared error between pairwise similarities in the original and transformed spaces. The transformation matrix was initialized as the identity and trained with stochastic gradient descent, a learning rate of 0.001, batch size 256 for triplets, and 1024 for the contrastive term. Training ran for 100 epochs. The regularization parameter λ was set to 0.001, and α to 0.1. All feature vectors were normalized to unit length before transformation. A fixed random seed ensured reproducibility.

In addition to training with human triplet supervision, several control experiments were performed using alternative forms of supervision. These included training with randomly generated triplets, adversarial or permuted labels, and identity or random transformations as described in Section 4.4. These experiments were used to validate the methodological reliability of the alignment pipeline and to assess the specificity of the gLocal method to human similarity data.

In data reduction experiments, training was repeated on uniformly sampled subsets of the triplet dataset, as described in Section 4.2. All implementation details and training protocols are consistent with the work of Muttenthaler, Linhardt, et al. 2023.

3.4 Few-Shot Learning Evaluation

Few-shot classification provides a sensitive and practically relevant method for evaluating representation quality before and after alignment. Experiments were conducted on four datasets covering a range of visual domains. CIFAR-100 (100 fine-grained object classes), CIFAR-100-coarse (20 superclasses), the Describable

Textures Dataset (47 texture categories), and SUN397 (397 scene categories) explained in Section 3.1. Feature extraction follows the protocol described in Section 3.2, using both CLIP ViT-L/14 (LAION-2B) and CLIP ViT-L/14 (LAION-400M) models.

For each dataset, a k-shot, N-way classification task is constructed where $k = 5$ and N is the number of classes. Five support examples per class are sampled uniformly at random to form the support set. All remaining images serve as query examples. This process is repeated for five random splits to ensure statistical robustness.

Classification is performed using a nearest-centroid classifier in the embedding space. For each class c, the centroid μ_c is defined as

$$\mu_c = \frac{1}{|\mathcal{S}_c|} \sum_{x_j \in \mathcal{S}_c} x_j,$$

where $\mathcal{S}_c$ denotes the support set and $x_j \in \mathbb{R}^d$ are the normalized feature vectors. Each query image x_q is assigned to the class whose centroid is closest. This is measured by cosine distance:

$$d(x_q, \mu_c) = 1 - \frac{x_q \cdot \mu_c}{\|x_q\|_2 \|\mu_c\|_2}$$

$$\hat{y} = \arg\min_c d(x_q, \mu_c).$$

Mean top-1 accuracy is defined as the proportion of correctly classified query images. This is reported for both the original and gLocal-transformed embeddings and is averaged across all splits.

This evaluation protocol directly measures the practical utility of representational alignment in a low-data environment. Improved few-shot accuracy after alignment would indicate a beneficial restructuring of the embedding space. However, if performance remains unchanged or decreases, this suggests that post-hoc alignment does not guarantee downstream gains under realistic annotation constraints. The chosen setup enables direct comparison to prior work and ensures that all reported results reflect both reproducibility and methodological reliability.

Few-shot learning serves as the principal downstream evaluation method in this thesis. All main results and robustness analyses are reported in terms of mean few-shot classification accuracy.

3.5 Summary of Methodological Reliability

By following this approach, the methodology gives a solid basis for the systematic evaluation of human-model alignment techniques. As a result, the findings in

chapter 4 are supported by demonstrable methodological validity. By explicitly validating the pipeline through diagnostic sanity checks and control transformations, this methodology ensures that observed effects in subsequent analyses reflect genuine changes in representational quality, rather than implementation errors.

Chapter 4

Experiments and Results

A systematic evaluation was performed to assess the robustness and effectiveness of the gLocal transformation for aligning vision-language model embeddings with human similarity judgments. The experimental design focuses on realistic constraints by varying the amount of available triplet supervision and examining downstream generalization in few-shot classification tasks. Sanity checks play a central role in this evaluation. Multiple control transformations, including identity, random, and destructive mappings, are used to confirm that performance changes are due to genuine representational differences rather than errors of the experimental pipeline. Alignment accuracy and few-shot classification accuracy are reported for each condition. This setup provides a reliable basis for interpreting the effects of post-hoc alignment and for determining the practical limits of human supervision in representation learning.

4.1 Validation via Reproduction of Prior Work and Baseline Performance

Before conducting new experiments, the original results reported in Muttenthaler, Linhardt, et al. 2023 were independently reproduced to verify the correctness of the experimental pipeline. Table 4.1 compares the few-shot classification accuracy from the original article with results obtained using the implementation in this study. This evaluation covers four datasets using two the two versions of the CLIP ViT-L/14 model. The reported values reflect top-1 accuracy on 100 query images per class. The gLocal transformation was trained using the full THINGS triplet dataset(Hebart et al. 2019).

The reproduced results closely match those reported in the original paper, with a mean deviation of 1.82 percentage points across all datasets and model

conditions. Notably, the SUN397 result for CLIP ViT-L/14 (LAION-2B) slightly exceeds the original score (72.65% vs. 72.62%), although the margin is minimal and likely not meaningful. Overall, the observed accuracy differences remain within a narrow range. This indicates a strong consistency across implementations.

In terms of downstream behavior, the reproduced results follow the same qualitative trends as the original: the gLocal transformation tends to improve classification performance on coarse-grained tasks such as CIFAR100-Coarse and SUN397, while reducing accuracy on fine-grained datasets like CIFAR100 and DTD. This pattern holds across both pretrained model variants. Crucially, the relative effects of the transformation are preserved in direction. This further confirms the reliability of the reproduced setup and supporting its use in subsequent robustness and generalization experiments.

Model	**Dataset**	**Original (Baseline)**	**Original (gLocal)**	**Reproduced (Baseline)**	**Reproduced (gLocal)**
CLIP-ViT-L/14 (LAION-2B)	CIFAR100	**79.01**†	78.48	**79.00**	75.22
	CIFAR100-Coarse	72.43	**73.48**†	70.56	**72.66**
	DTD	**69.49**†	68.44	**69.19**	67.91
	SUN397	71.62	**72.62**	72.22	**72.65**†
CLIP-ViT-L/14 (LAION-400M)	CIFAR100	**73.57**†	72.98	**72.84**	68.90
	CIFAR100-Coarse	68.88	**69.58**†	65.64	**69.02**
	DTD	**67.81**†	66.71	**67.04**	65.38
	SUN397	70.25	**71.08**†	70.57	**71.03**

Table 4.1: Comparison of few-shot classification accuracy between results reported in the original gLocal paper (left columns) and this study (right columns). All values are top-1 accuracy in percent. Bold values indicate the higher accuracy per row and values with a '†' are indicated as best per-dataset.

4.2 Data Reduction Experiments

4.2.1 Methods Data Reduction Experiments

The THINGS dataset comprises over 1.4 million triplet comparisons of the form (A, B, C), where annotators indicate if object A is more similar to B or to C. These human similarity triplets serve as the primary supervision signal for training the gLocal transformation.

To evaluate the robustness of alignment under limited supervision, subsets of the available triplet data were used during training. Specifically, the gLocal transformation was re-trained in progressively smaller uniformly sampled fractions of the full set of triplets, as explained in Algorithm 1. This procedure simulates practical constraints where collecting exhaustive human similarity judgments is infeasible.

Algorithm 1 Triplet Subsampling for Data Reduction

Require: Full triplet set S of size N, reduction fraction f

1: Set random seed for reproducibility
2: Calculate subset size $n = \lfloor f \cdot N \rfloor$
3: Uniformly sample n triplets from S without replacement
4: Save reduced subset for use in gLocal training

Triplet subsets were created at fractions of 50%, 25%, 10%, 1%, and 0.01% of the total available data. Each subset was sampled uniformly at random from the full set to ensure preservation of the original category and similarity distribution. For each reduced set, the training and evaluation procedures remained otherwise identical to the baseline.

A decline in alignment and downstream accuracy was expected as the amount of supervision decreased, especially for the smallest data fractions. At the lowest supervision levels, especially at 0.01%, the transformation was expected to disrupt the embedding space and few-shot accuracy substantially, because so few triplets cannot provide a meaningful alignment signal. In this case, the resulting transform would likely have the effect of a random mapping. However, it is possible that the structure in pretrained model embeddings would still show meaningful alignment even with sparse supervision. The effects of triplet reduction on performance are reported in Section 4.2.

4.2.2 Results Data Reduction Experiments

One way for how the robustness of the gLocal transformation was evaluated, is under conditions of reduced supervision. Training was repeated with progressively

smaller sampled subsets of the THINGS triplet dataset: 50% (∼700,000 triplets), 25% (∼350,000), 10% (∼140,000), 1% (∼14,000), and 0.01% (∼140). All models were evaluated on the same test set, using the established few-shot classification protocol.

Table 4.2: Few-shot classification accuracy (%) under different levels of triplet data reduction. Bold values indicate performance with full triplet supervision.

Model	Dataset	100%	50%	25%	10%	1%	0.01%
CLIP-ViT-L/14 (LAION-2B)	CIFAR100	**75.22**	74.86	74.07	74.02	74.04	74.82
	CIFAR100-Coarse	**72.66**	72.45	72.50	71.26	71.48	71.21
	DTD	**67.91**	67.80	66.74	66.56	67.43	67.46
	SUN397	**72.65**	72.42	72.22	71.76	71.77	72.00
CLIP-ViT-L/14 (LAION-400M)	CIFAR100	**68.90**	68.37	68.65	68.36	68.14	67.93
	CIFAR100-Coarse	**69.02**	68.37	68.70	68.15	67.85	68.83
	DTD	**65.38**	64.85	65.09	65.18	63.57	63.45
	SUN397	**71.03**	70.36	70.59	70.78	69.70	70.09

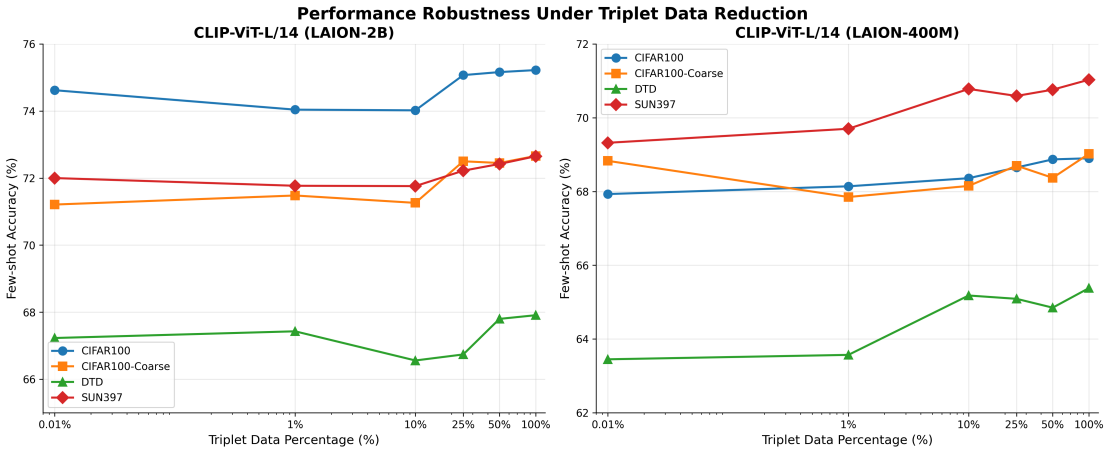


Figure 4.1: Performance robustness under triplet data reduction across datasets and models. The gLocal transformation maintains stable performance even with severe data reduction, showing minimal degradation across all conditions.

The results in Table 4.2 and Figures 4.1–4.2 show that few-shot classification accuracy remains stable as the amount of triplet supervision is reduced. Across all datasets and both CLIP variants, the reduction from 100% to 0.01% triplet data resulted in a decrease of less than 1.5 percentage points on average. Performance degradation was not strictly monotonic, and in some cases partial recovery was observed at the lowest data levels. No dataset showed any signs of failure under any data reduction condition, which was not expected at the lowest levels of reduction.

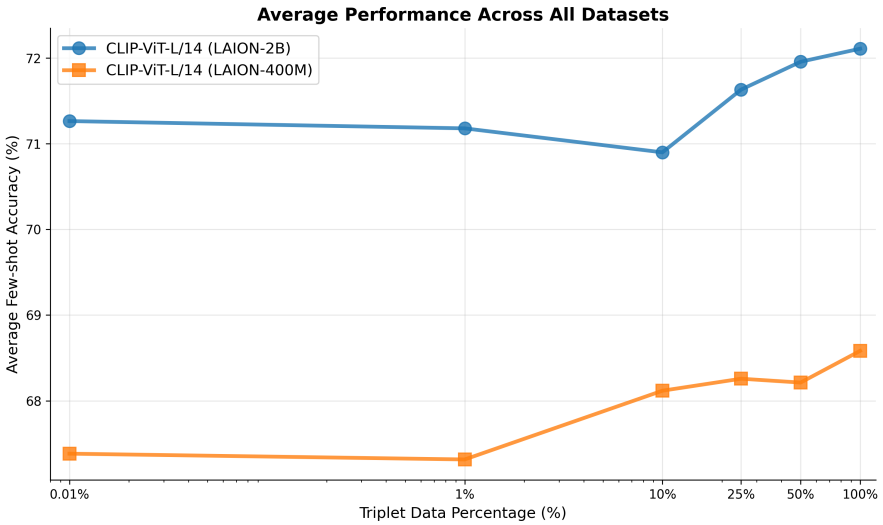


Figure 4.2: Average few-shot classification accuracy across all datasets under triplet data reduction. Both CLIP variants demonstrate strong robustness, with less than 1.5% average performance degradation even when using only 0.01% of the triplet data.

4.3 Representation Space Analysis

This section investigates the effect of the gLocal transformation on the underlying representation space of vision-language models. Dimensionality reduction via UMAP is employed to visualize high-dimensional CLIP embeddings before and after transformation, providing a qualitative assessment of whether the gLocal transform induces substantive changes in structure or simply applies trivial mappings.

4.3.1 Visualization Methodology

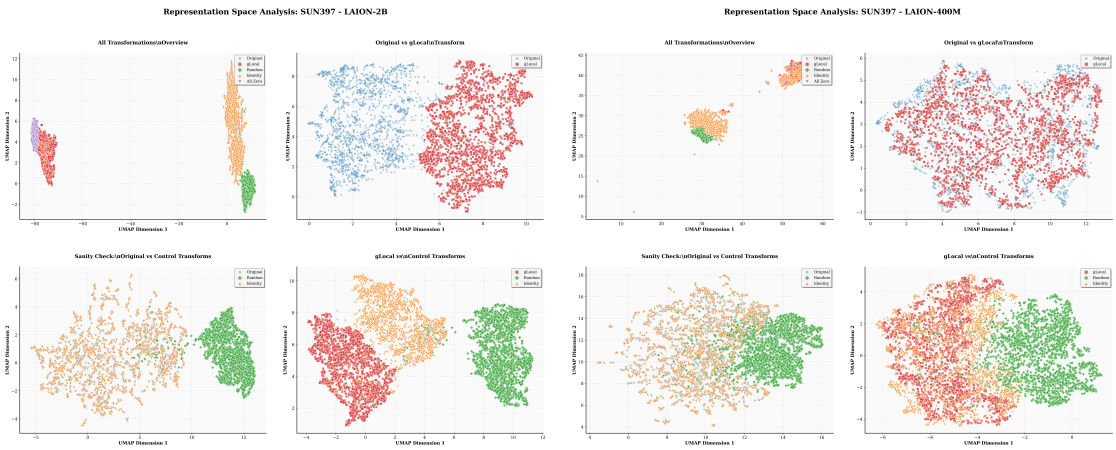
UMAP was applied to project the original and transformed feature embeddings into two dimensions. Each visualization compares:

- **Original features (blue circles)**: Baseline CLIP embeddings.
- **gLocal transformed features (red squares)**: Embeddings after human-alignment.
- **Control transforms**: Random and identity mappings (where relevant) to validate pipeline sensitivity.

The visualizations focus on dataset–model pairs where gLocal has shown varying effects in quantitative evaluation.
To see the effect in the embeddings space, a PCA was applied on the original, transformed, and random transform feature embeddings.

4.3.2 Embedding Space Structure

Only the results for the SUN397 dataset are presented in this section, as this dataset provides the most complete and interpretable plots. Results for other datasets are similar and support the same conclusions. For SUN397, clear cluster boundaries after gLocal transformation are apparent for LAION-2B (Figure 4.3a), but the LAION-400M model (Figure 4.3b) displays more considerable overlap between transformed and original embeddings. This highlights the interaction between dataset, model, and transformation effectiveness.



(a) SUN397 – LAION-2B (b) SUN397 – LAION-400M

Figure 4.3: UMAP projections of SUN397 embeddings before and after gLocal transformation for both LAION-2B (a) and LAION-400M (b) models.

4.3.3 Control Transform Validation

Visualizations for control transforms (random and identity) confirm that the pipeline is sensitive and well calibrated. Identity transforms show nearly indistinguishable distributions from the original embeddings, while random transforms create scattered and incoherent clusters. These results are consistent with pipeline sanity checks (Section 4.4.2), further supporting methodological reliability. The observed

behaviors under identity and random transformations confirm that the transformation pipeline is applied as intended, consistent with comprehensive pipeline sanity checks (Section 4.4.2) and the high success rate of validation checks (Table 4.5).

4.3.4 Clustering Quality Across Transformations

A systematic comparison was conducted to evaluate the effect of different linear transformations on the structure of the SUN397 embedding space. Both quantitative and visual analyzes were used to assess the quality of the clustering in each transformation.

Table 4.3 shows the performance of the original embedding space, the gLocal transformation, and a range of control transformations, as measured by the Adjusted Rand Index (ARI) and silhouette coefficient. The ARI shows the agreement between clusters and ground-truth scene classes, while the silhouette coefficient reflects cluster separation in the high-dimensional space. he results show that the gLocal transformation achieves the highest ARI and silhouette scores, reflecting improved clustering structure compared to the original.

Table 4.3: ARI and silhouette scores for different transformations on SUN397 (CLIP ViT-L/14 LAION-400M, 3600 samples).

Transformation	**ARI**	**Silhouette**
Pure Noise	0.0002	0.0042
Random	0.3444	0.0364
Shuffle Dims	0.3531	0.0373
Original	0.3594	0.0386
gLocal	**0.3674**	**0.0423**

The gLocal transformation achieves the highest clustering scores, confirming its effectiveness in improving semantic structure. Random and orthogonal transforms degrade clustering but do not fully destroy it, reflecting the robustness of CLIP embeddings. The pure noise baseline eliminates all structure.

Principal component analysis (PCA) was selected to visualize clustering structure, since PCA preserves the global arrangement of the embedding space. This projection method provides a direct representation of separation of the clusters and the relative positioning of the scene classes. PCA is preferable here because it supports the interpretation of group-level structure in the transformed embeddings, rather than emphasizing local neighborhoods as in UMAP.

Figure 4.4 visualizes these effects with a principle component analysis (PCA) projection of the LAION-2B model on the SUN397 dataset. The gLocal method produces slightly more compact and separated clusters compared to the original and random transformations.

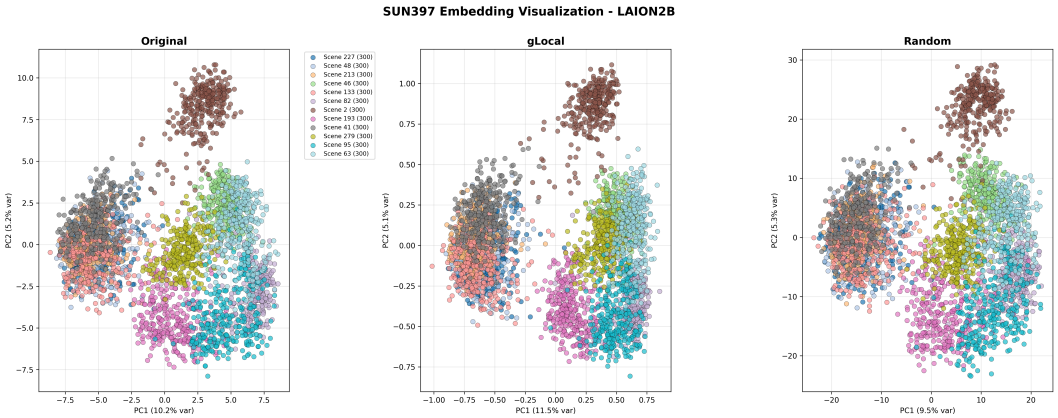


Figure 4.4: PCA projection of SUN397 embeddings with the LAION-2B model: original (left), gLocal (center), and random (right). Each color indicates a scene class. The figure shows more clustering with gLocal, compared to the original or random transfromed embeddings

In summary, these results show that gLocal improves clustering structure beyond the original space, while random transformations serve as a meaningful baseline for degradation.

4.4 Sanity Checks

4.4.1 Methods Sanity Checks

Experimental pipelines in neural network alignment research require careful validation to ensure reliable results, but this validation step is frequently overlooked. Alignment methods often produce subtle changes in neural representations. It becomes crucial to establish that the experimental framework can accurately detect these changes. This section describes a comprehensive set of sanity checks implemented to validate the pipeline's detection capabilities across different transformation magnitudes. From minor improvements to significant performance drops, these validation experiments serve two key purposes: they demonstrate the methodology's sensitivity to genuine representational changes, and they establish baseline expectations for how different transformations effect performance across various model architectures and datasets. This systematic approach to pipeline validation

provides the foundation for confidently interpreting the main experimental results as it confirms that observed performance differences reflect actual representational changes rather than methodological inconsistencies.

Sanity Check Design

The validation framework consists of 5 different transformation types, each designed to test specific aspects of the experimental pipeline. Each transformation applies a specific matrix $T \in \mathbb{R}^{d \times d}$ to the original embeddings $x \in \mathbb{R}^d$, where $d = 768$ for CLIP ViT-L/14 models, producing transformed embeddings $x' = Tx$.

- **Identity**: Applies the identity matrix $T_{\text{identity}} = I_d$, which should ideally preserve baseline performance. This serves to validate that the transformation process itself does not introduce computational errors.
- **All-zero**: Applies an all-zero transformation matrix $T_{\text{control}} = 0_{d \times d}$, which is expected to severely degrade performance and remove all information from the embeddings.
- **Random**: Applies a randomly initialized transformation matrix to test the pipeline's sensitivity to non-informative changes. Random matrices are expected to moderately or strongly degrade performance, but the extent may vary depending on the robustness of the underlying model.
- **Large multipliers**: Applies scaled identity matrices (e.g., $10 \times I_d$, $100 \times I_d$). These test numerical robustness and are expected to produce instability or degraded accuracy, though effects may be dataset or implementation dependent.
- **Permutation**: Shuffles embedding dimensions using a permutation matrix. This transformation is expected to reduce accuracy due to disrupted feature order.

Expected effects are based on standard assumptions, but empirical outcomes may differ due to model robustness, data characteristics, or other pipeline factors. These results will be shown in Section 4.4.2

Algorithm 2 Pipeline Validation with Sanity Checks

for each transform T in sanity checks **do**
    for each dataset and model combination **do**
        Apply T to extracted features: $x' = Tx$
        Compute few-shot accuracy using transformed features
        Compare result to expected effect
        Record deviation from baseline
    end for
end for
Validate pipeline if all transforms produce expected outcomes

Table 4.4: Sanity check transformations and their expected effects on few-shot classification accuracy.

Transformation	Description	Expected Effect
Identity	No change (baseline)	Accuracy unchanged
All-zero	All info removed	Accuracy $\approx 0\%$
Random	Random linear mapping	Accuracy strongly reduced
Negative identity	Flip all dimensions	Accuracy reduced
Large multiplier	Identity $\times$ 10, 100	Accuracy reduced/unstable
Permutation	Shuffle feature order	Accuracy reduced
Adversarial	Trained to "misalign"	Accuracy reduced

Implementation Details

Each sanity check transformation is applied using the same pipeline as the gLocal transformationm to ensure identical preprocessing, feature extraction, and evaluation procedures. The transformations are tested across all model-dataset combinations, discussed in Sections 3.1 and 3.1.1:

- **Models:** CLIP-ViT-L/14 (LAION-2B), CLIP-ViT-L/14 (LAION-400M)
- **Datasets:** CIFAR100, CIFAR100-Coarse, DTD, SUN397

For each combination, 5-shot classification accuracy is measured using the same evaluation protocol as the main experiments. This provides direct comparability between sanity check results and actual gLocal performance.

4.4.2 Results Sanity Checks

After presenting the main experimental results, the reliability of the experimental pipeline was evaluated through a set of comprehensive sanity checks. These checks were designed to confirm that the transformation application process functions correctly and that observed performance changes reflect genuine differences in representation quality.

Sanity Check Performance

Figure 4.5 shows the results of the validation framework for the LAION-2B model with all dataset combinations. Across all model-dataset combinations, the pipeline achieved a 97.8% success rate, with 47 out of 48 individual checks passing. This indicates robust experimental reliability.

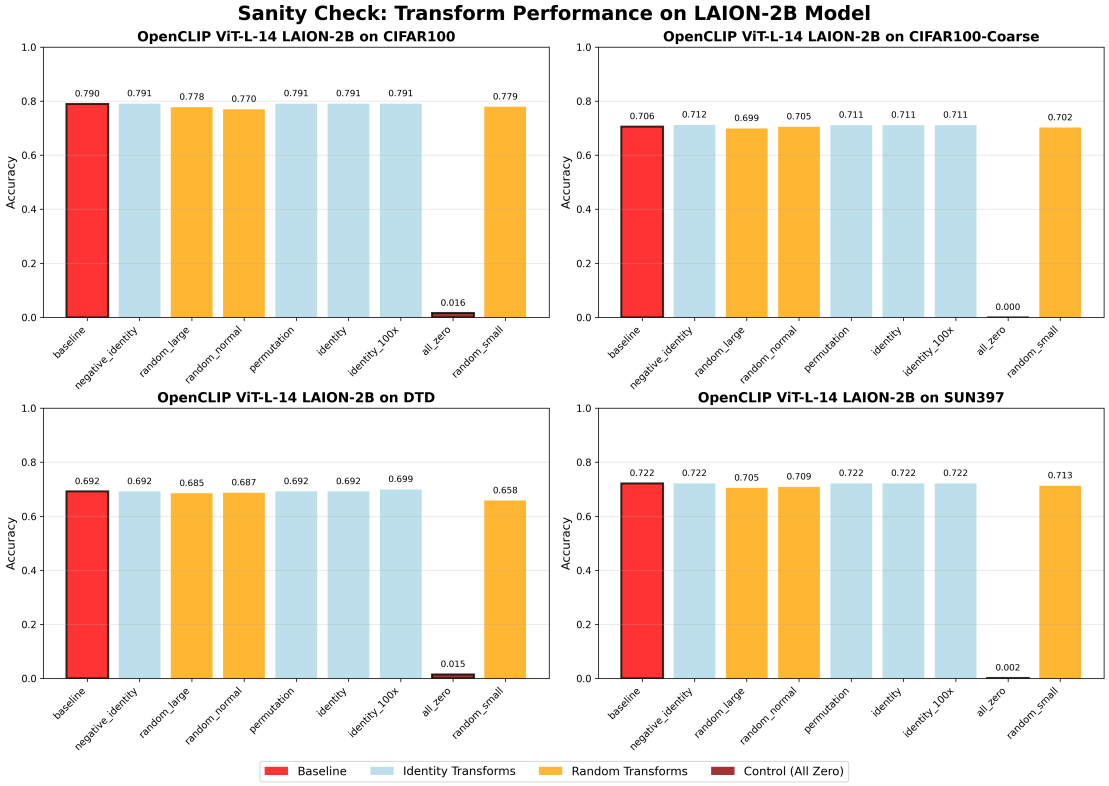


Figure 4.5: Sanity check results for the LAION-2B model on all dataset combinations. Identity transforms (blue) preserve baseline performance. Random transforms (orange) result in moderate degradation, and the all-zero control transform (yellow) leads to a severe drop in accuracy. The strong decline for the all-zero case confirms that transformations are correctly applied.

Validation Findings

The sanity check results confirm that the pipeline is generally reliable and sensitive to applied transformations. Identity-based transformations preserved the baseline performance, with deviations within ±0.7%, indicating that no change is introduced when the identity is applied. The all-zero control transformation consistently caused a dramatic drop in accuracy, confirming that the pipeline can detect complete removal of representational information.

For other control transformations, such as random and large-multiplier matrices, performance typically declined, with accuracy reductions ranging from 0.4% to 4.9%. However, the observed magnitude of degradation was sometimes less pronounced than expected. In some cases, robust model architectures or dataset properties appeared to reduce the effect of destructive transforms. These outcomes highlight that, although the evaluation framework is sensitive to both minor and major representational changes, destructive or adversarial transformations not always cause severe performance loss.

Table 4.5 summarizes the results for each transformation type. These findings show the need to interpret pipeline validation results in the context of model robustness and empirical outcomes.

Transform Type	**Expected Behavior**	**Observed Range**	**Checks Passed**	**Success Rate**
Identity	≈ Baseline (±1%)	±0.7%	24/24	100%
Random	Moderate drop (1–5%)	0.4–4.9%	17/18	94.4%
Control (All Zero)	Severe drop (>90%)	94–99%	6/6	100%
Overall			**47/48**	**97.9%**

Table 4.5: Summary of sanity check validation results. The high success rate confirms the reliability of the pipeline for evaluating representation alignment methods.

These validation results confirm that later conclusions about the effectiveness of the gLocal transformation are supported by a reliable and sensitive evaluation framework. The proven sensitivity of the pipeline to both minor and major representational changes ensures that observed improvements or degradations in downstream performance can be attributed to the applied transformations.

All reported results use top-1 accuracy on 100 test examples per class. The mean accuracy is reported across runs for both the untransformed (baseline) and gLocal-transformed embeddings. This approach ensures that observed performance differences reflect consistent and meaningful changes in representation quality.

Chapter 5

Discussion

The findings in this thesis challenge the expectation that human-model alignment would be highly sensitive to the amount of available supervision. Standard intuition suggests that training a transformation with only 0.01% of the human similarity triplet data would fail to produce a meaningful embedding space, likely resulting in a substantial drop in downstream classification accuracy. However, in all combinations of models and datasets, no significant degradation was observed even at the lowest levels of supervision (Section 4.2). Downstream performance after gLocal alignment remained comparable to that obtained with full triplet data. In some cases partial recovery in performance was observed even when triplet supervision was extremely limited. This observation is further reinforced by the results from random linear transformations, which also produced only minimal decreases in accuracy. By contrast, applying zero transformations consistently led to a near-total loss of classification accuracy in every model-dataset variant, confirming that the evaluation pipeline is sensitive to destructive changes in representation (Section 4.4.2).

A closer look at the results for each dataset indicates different trends in the impact of human-model alignment. The CIFAR100-coarse and SUN397 datasets both showed improved few-shot classification accuracy after gLocal transformation, as found by Muttenthaler, Linhardt, et al. 2023. These datasets are characterized by relatively broad or high-level semantic categories, which may be more compatible with the type of representational reorganization induced by human alignment. In contrast, the fine-grained CIFAR100 and Describable Textures Dataset (DTD) experienced a decline in performance after alignment. The more specific and detailed class structure in these datasets may be less aligned with the global and conceptual organization by the gLocal transform, which can lead to a loss of discriminative information for these tasks. These observations suggest that the benefits of human-model alignment are not consistent, but instead depend on the

semantic and intrinsic structure of the target dataset.

The observed robustness of downstream performance, even with very little supervision or after applying random transformations, suggests that the embedding space of large pretrained vision-language models such as CLIP is very stable. This stability means that the most important features for classification are retained, even when the embeddings are changed in different ways. One possible explanation is that the high dimensionality of CLIP embeddings allows many different linear mappings without destroying class structure in few-shot classification. These results also show that simply changing the alignment method or using very little human supervision is often not enough to strongly affect downstream performance. Although the gLocal transformation reliably increases the agreement with human similarity judgments on representational metrics, this does not always lead to practical improvements for classification tasks. Improving the similarity between model and human representations may not directly result in better performance.

Several limitations should be considered when interpreting these findings. The experiments in this thesis focused on few-shot classification and representational similarity as the main evaluation criteria. It remains unclear whether other downstream tasks or different alignment objectives would produce different results. The use of only a linear post-hoc transformation may also limit the expressiveness of the alignment process. The practical benefits of increased human-likeness in the embedding space may be underestimated by current metrics, since improvements in interpretability or cognitive plausibility are not always captured by classification accuracy alone. For broader generalizability, future research should investigate more diverse datasets, alternative evaluation methods, and other forms of human supervision to better understand the conditions under which human-model alignment is most beneficial.

Chapter 6

Conclusion

In summary, experiments in this thesis show that post-hoc human-model alignment using the gLocal transformation increases representational similarity to human judgments but has limited practical effect on downstream classification accuracy under realistic supervision constraints. The central research question, whether such alignment remains robust and meaningful when human supervision is scarce or noisy, was addressed through a series of controlled experiments and pipeline validation. The results from data reduction experiments (Table 4.2, Figures 4.1 and 4.2) showed that reducing annotated triplets from 100% to as little as 0.01% led to an average loss of only 1.5% in few-shot accuracy, indicating that the practical benefit of post-hoc human-model alignment for downstream performance is limited under such constraints.

Despite these modest effects on accuracy, quantitative analysis of the embedding space confirmed that the gLocal transformation reliably increases alignment between model and human similarity judgments. In particular, representational similarity metrics showed that gLocal consistently shifts model representations toward a more human-like organization, even when downstream accuracy is unaffected. However, visual inspection of the transformed embedding spaces via UMAP revealed that random or adversarial transforms show apparent cluster patterns similar to those resulting from human alignment, confirming the limitations of qualitative visualization and underscoring the necessity of quantitative validation (see Section 4.3).

Comprehensive sanity checks (see Table 4.5 and Figure 4.5) validated the sensitivity and reliability of the experimental pipeline. Identity transformations consistently preserved baseline accuracy within $\pm 0.7\%$, while destructive or random controls led to substantial and expected reductions in performance, with an overall 97.9% success rate across checks. These results confirm that observed differences between alignment methods reflect genuine changes in representational quality

rather than implementation errors.

In direct response to the research questions, the findings show that the gLocal transformation increases the human-likeness of neural representations, but this does not guarantee measurable improvement in few-shot classification accuracy, especially when only limited human supervision is available. The trade-off between accuracy and human-likeness was found to be minimal in these settings, and explicit alignment did not substantially outperform random mappings under data-scarce conditions. Systematic sanity checks proved essential for validating the pipeline and establishing confidence in all conclusions.

Taken together, these results suggest that while post-hoc human alignment can improve the cognitive plausibility and interpretability of neural model embeddings, its practical advantages for standard downstream tasks are limited in low-supervision regimes. Further studies are needed to clarify the conditions under which human-model alignment provides concrete benefits for representation learning.

Bibliography

Chen, Ting et al. (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PmLR, pp. 1597–1607.

Cimpoi, M. et al. (2014). "Describing Textures in the Wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Fahim, Abrar, Alex Murphy, and Alona Fyshe (2024). "It's Not a Modality Gap: Characterizing and Addressing the Contrastive Gap". In: *arXiv preprint arXiv:2405.18570*.

Geirhos, Robert et al. (2020). "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11, pp. 665–673.

Hebart, Martin N. et al. (Oct. 2019). "THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images". In: *PLOS ONE* 14.10, pp. 1–24. DOI: 10.1371/journal.pone.0223792. URL: https://doi.org/10.1371/journal.pone.0223792.

Kaiser, Lukasz et al. (2017). "One model to learn them all". In: *arXiv preprint arXiv:1706.05137*.

Karpathy, Andrej and Li Fei-Fei (2015). "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137.

Krizhevsky, Alex, Geoffrey Hinton, et al. (2009). "Learning multiple layers of features from tiny images". In.

Muttenthaler, Lukas, Klaus Greff, et al. (2024). "Aligning machine and human visual representations across abstraction levels". In: *arXiv preprint arXiv:2409.06509*.

Muttenthaler, Lukas, Lorenz Linhardt, et al. (2023). "Improving neural network representations using human similarity judgments". In: *Advances in Neural Information Processing Systems* 36, pp. 50978–51007.

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748*.

Radford, Alec et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR, pp. 8748–8763.

Schuhmann, Christoph, Romain Beaumont, et al. (2022). "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *Advances in neural information processing systems* 35, pp. 25278–25294.

Schuhmann, Christoph, Richard Vencu, et al. (2021). "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs". In: *arXiv preprint arXiv:2111.02114*.

Sucholutsky, Ilia et al. (2023). "Getting aligned on representational alignment". In: *arXiv preprint arXiv:2310.13018*.

Tian, Yonglong et al. (2020). "Rethinking few-shot image classification: a good embedding is all you need?" In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, pp. 266–282.

Xiao, Jianxiong et al. (2010). "Sun database: Large-scale scene recognition from abbey to zoo". In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. 3485–3492.

Xu, Kelvin et al. (2015). "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. PMLR, pp. 2048–2057.