Lost in Embeddings: Information Loss in Vision–Language Models

Wenyan Li[1] **Raphael Tang**[2] **Chengzu Li**[3] **Caiqi Zhang**[3]
Ivan Vulić[3] **Anders Søgaard**[1]
[1]University of Copenhagen [2]Microsoft [3]University of Cambridge
{weli, soegaard}@di.ku.dk v-raptang@microsoft.com
{cl917, cz391, iv250}@cam.ac.uk

Abstract

Vision–language models (VLMs) often process visual inputs through a pretrained vision encoder, followed by a projection into the language model's embedding space via a connector component. While crucial for modality fusion, the potential information loss induced by this projection step and its direct impact on model capabilities remain understudied. We introduce two complementary approaches to examine and quantify this loss by analyzing the latent representation space. First, we evaluate semantic information preservation by analyzing changes in k-nearest neighbor relationships between image representations, before and after projection. Second, we directly measure information loss by reconstructing visual embeddings from the projected representation, localizing loss at an image patch level. Experiments reveal that connectors substantially distort the local geometry of visual representations, with k-nearest neighbors diverging by 40–60% post-projection, correlating with degradation in retrieval performance. The patch-level embedding reconstruction provides interpretable insights for model behavior on visually grounded question-answering tasks, finding that areas of high information loss reliably predict instances where models struggle.[1]

1 Introduction

Vision–language models (VLMs) have advanced on many tasks, e.g., visual question answering and image captioning by combining pretrained vision encoders with pre-trained language models. Many of these models employ a small neural network module, known as a connector (or a projector), to bridge the gap between the visual and textual representation spaces. The connectors project visual representations into sequences of embeddings that a language model can process (Chen et al., 2024a; Liu et al., 2023; Deitke et al., 2024; Laurençon et al., 2024; Chen et al., 2024b; Zhang et al., 2025; Sun et al., 2024). Common connector architectures include multi-layer perceptrons (MLPs) or attention-based approaches (Jaegle et al., 2021; Laurençon et al., 2024).

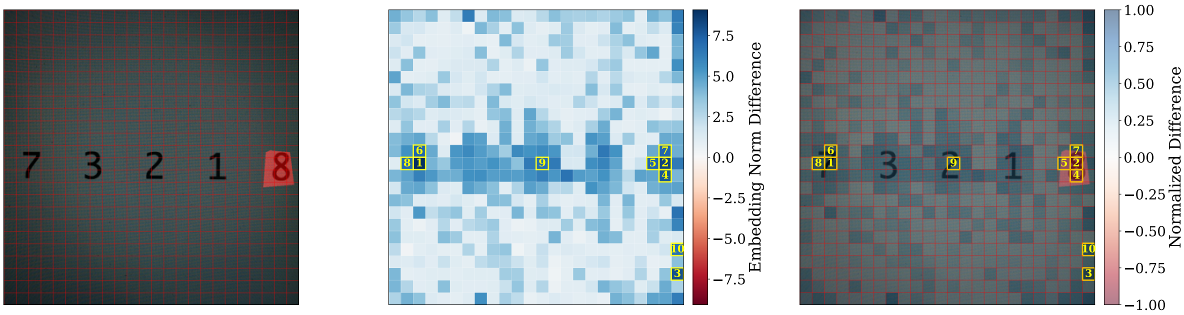
While these connector modules enable efficient cross-modal integration (Li and Tang, 2024), projecting rich visual features into embeddings compatible with language models typically involves dimensional conversion and representation restructuring. Naturally, this raises questions about potential information loss[2] during projection, and how such loss impacts downstream task performance. As shown in Figure 1, the loss of critical visual details most relevant to answering the question imposes inherent limitations on the reasoning capabilities, since the language model's performance is constrained by the quality and completeness of the visual information it receives.

Despite the growing research on VLM connector architectures and their impact on downstream performance (Lin et al., 2024; Zhu et al., 2025), there has been limited investigation into how well they preserve visual information in the latent space. Quantifying this information loss presents substantial challenges; traditional methods like canonical correlation analysis (Hotelling, 1936) struggle with variable-length high-dimensional visual features processed through diverse connector architectures in vision-language models. Performance degradation can also take more than one form, adding to the complexity of its study. For instance, it can take the form of a *direct information loss* due to an inherently lossy connector, or a *geometric collapse* where distinct features become entangled in the projected embedding space.

[1]Code: https://github.com/lyan62/vlm-info-loss.

[2]In this paper, we use "information loss" to broadly describe possible degradation of visual information, including aspects of the representation that cannot be recovered or directly observed after the projection. In a stricter sense, this could also be viewed as a representational gap or discrepancy rather than true loss, e.g. changes in the local geometry as reflected by k-nearest-neighbor relationships.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22676–22693
November 4-9, 2025



(a) Input image with red answer mask (b) Embedding norm signed difference (c) Image overlay with norm difference

Figure 1: Visualization of patch-wise information loss in the embeddings explains the incorrect predicted answer in VizWiz Grounding VQA. For the question "What is the fifth number?", LLaVA incorrectly predicted "18". Figure 1b display the difference between the L^2 norm of the original and the reconstructed patch embeddings. Blue regions indicate where original embeddings have larger norms than predicted embeddings, while red regions show where predicted embeddings have larger norms. The top 10 high-loss patches are marked by yellow squares. Figure 1c shows high loss occurring in several answer-relevant patches contribute to the incorrect prediction.

To bridge this gap in the literature, we present an evaluation framework to quantify information loss in VLM connectors from both the geometric perspective and that of localized information loss. We first examine if the connector projection changes the geometric structure of latent visual representations. By introducing **k-nearest neighbors overlap ratio**, we measure how much the neighborhoods of image embeddings change before and after the projection in the latent representation space, thereby estimating how well geometric and semantic relationships are preserved.[3] Second, we quantify localized information loss by training a model to reconstruct the original visual embeddings from the projected embeddings. This **patch-level visual embedding reconstruction** allows us to pinpoint the high-loss regions in the image—areas where visual features are hard to recover after projection (Figure 1). This two-step approach provides both quantitative analysis and interpretable visualizations, offering insights into the nature of information transformation during vision-text integration.

2 VLMs and Connectors

Integrating visual and textual inputs is fundamental for VLMs to process multimodal information effectively. Existing VLMs typically employ two main approaches (Li and Tang, 2024): models like LLama3.2 (gra, 2024) and BLIP (Li et al., 2023) leverage cross-modal attention mechanisms, while others such as LLaVA (Liu et al., 2023) and Qwen-2-VL (Bai et al., 2025) adopt connectors to project visual representations into latent vectors compatible with large language models (LLMs).[4]

Lin et al. (2024) categorize connectors into two types: feature-preserving and feature-compressing connectors. Feature-preserving connectors include MLPs that preserve the number of patch embeddings, such as the two-layer MLP connector in LLaVA. In contrast, feature compressing connectors project image patch embeddings to a shorter sequence, often involving transformer-based or convolution architectures with pooling operations over the original vision embedding (Jaegle et al., 2021). Feature compressing connectors include the perceiver sampler in Idefics2 (Laurençon et al., 2024) and the patch merger in Qwen-2-VL (Bai et al., 2025). In this paper, we estimate information loss in both types of connectors.

2.1 Formalizing Encoders and Connectors

We now give a formal definition of connectors. First, we consider the textual input. Let Σ be an alphabet of symbols. A **string encoder**, ϕ, is a function that maps a string σ to a sequence of real-valued representations. Formally,

$$\phi\colon \Sigma^N \to (\mathbb{R}^D)^N, \tag{1}$$

where $N \in \mathbb{N}$ is a parameter in the dependent type that denotes the length of the input string, and D is the dimensionality of the representation. Next, we turn to the visual input. Let

[3] Here, the semantic relationship denotes the semantic similarity between pairwise embeddings measured by L^2 distance or their inner product similarity.

[4] In this paper, we do not consider VQ-VAE (van den Oord et al., 2017) based VLMs, which are more often used for text-to-image generation.

$\Delta = \{1, \ldots, 256\}^{H \cdot W \times C}$ be an array of image patches, where H and W represent the height and width dimensions, and C is the number of color channels per pixel. A two-dimensional image of patch dimensions $M_1 \times M_2$ can thus be represented as an element of $\Delta^{M_1 \times M_2}$. Where $\Delta^{M_1 \times M_2}$ denotes the set of all possible $M_1 \times M_2$ grids of patches. The **vision encoder** is formalized as a dependent type:

$$\psi\colon \Delta^{M_1 \times M_2} \to (\mathbb{R}^{D'})^{M_1 \times M_2}, \tag{2}$$

where M_1 and M_2 are parameters in the dependent type, representing the grid dimensions of the image patches, and D' is the visual embedding dimension. This maps a grid of image patches to a grid of embedding vectors.

A **connector** module transforms the vision encoder's output to match the dimensionality of the text encoder—projecting visual embeddings of dimension D' to text-compatible dimension D. We define the connector as a function of type:

$$\textsc{conn}\colon (\mathbb{R}^{D'})^{M_1 \times M_2} \to (\mathbb{R}^{D})^{M_C}, \tag{3}$$

where we typically have $M_C \leq M_1 M_2$. We also use $\textsc{c}$ as shorthand for $\textsc{conn}$.

For combining the output of the string encoder and the vision encoder, we define a **flattener** that combines visual and textual embeddings into a unified sequence:

$$\textsc{flat}\colon (\mathbb{R}^{D})^{M_C} \times (\mathbb{R}^{D})^{N} \to (\mathbb{R}^{D})^{M_C+N} \tag{4}$$

This creates a sequence of length $M_C + N$ by concatenating the flattened grid of visual embeddings with the sequence of text embeddings.

The complete vision–language models we consider can then be expressed as the a composition of these functions:

$$\text{VLM}(x, \sigma) = \text{LM}(\textsc{flat}(\textsc{conn}(\psi(x)), \phi(\sigma))) \tag{5}$$

where $x \in \Delta^{M_1 \times M_2}$ is an input image, $\sigma \in \Sigma^N$ is an input text sequence, and LM is an autoregressive language model that predicts probability of next tokens.

We focus on quantifying the information loss at the connector module defined in Equation 3. Formally, the information loss over the connector is a function $\mu : (\psi(x), \textsc{conn}(\psi(x))) \to \mathbb{R}_{\geq 0}$. We explore how such loss correlates with and explains model performance.

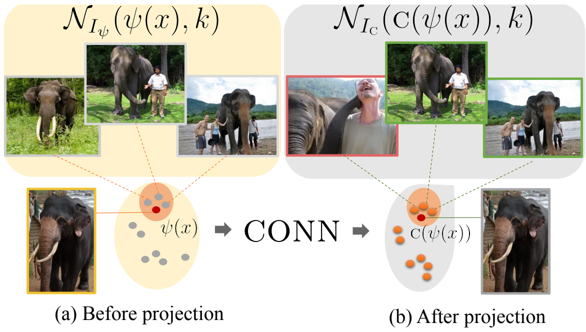


Figure 2: The k-nearest neighbors overlap ratio measures the overlap of an image's neighbors before and after projection. In this example, with $k = 3$, the overlap ratio is 0.67 because two out of the three nearest neighbors are identical in both representation spaces.

3 Quantifying Information Loss

We propose two methods for quantifying information loss over the projection step described above. The first method quantifies the preservation of structural information in semantic embeddings by comparing each image representation's k-Nearest Neighbors (k-NN, Fix and Hodges (1951)) before and after projection. The overlap ratio of the k-NN neighbors captures how well local geometry of the semantic embeddings are preserved in the latent space. Figure 2 gives an example where two of the three nearest neighbors overlap before and after projection. The second method evaluates patch-level representation (Figure 1) distortion by training an *ad hoc* neural network to reconstruct the original image embedding from its projected representation, detailed in Section 3.2.

3.1 k-Nearest Neighbors Overlap Ratio

To quantify geometric information loss during projection in visual representation spaces, we propose the **k-nearest neighbors overlap ratio (KNOR)**, which measures how well k-NN relationships among image embeddings are preserved before and after projection through the connector.

Let I be a finite set of images, ψ a vision encoder, and $\textsc{conn}$ ($\textsc{c}$ for short) a connector as described in §2.1. We use $I_\psi = \{\psi(x)\}_{x \in I}$ to indicate the family of embedded images, and $I_{\textsc{c}} = \{\textsc{conn}(\psi(x))\}_{x \in I}$ for the projection of the embedded images. The k-NN overlap ratio for an image x is defined as

$$\mathcal{R}(x, k) \overset{\text{def}}{=} \frac{\left|\mathcal{N}_{I_\psi}(\psi(x), k) \cap \mathcal{N}_{I_{\textsc{c}}}(\textsc{c}(\psi(x)), k)\right|}{k} \tag{6}$$

Where $\mathcal{N}_{I_\psi}(\psi(x), k)$ is the set of k-nearest neighbors of $\psi(x)$ among the pre-projected embeddings, and $\mathcal{N}_{I_C}(\mathrm{C}(\psi(x)), k)$ is the set of k-nearest neighbors of $\mathrm{C}(\psi(x))$ among the projected embeddings. The **average overlap ratio** is given by

$$\overline{\mathcal{R}}(k) \stackrel{\text{def}}{=} \frac{1}{|I|} \sum_{x \in I} \mathcal{R}(x, k) \tag{7}$$

The average overlap ratio measures how well the local geometric structure is preserved after projection. An optimal connector would maintain the same k-NN sets for $\psi(x)$ and $\mathrm{C}(\psi(x))$. Lower overlap ratio corresponds to more geometric information loss due to projection, while higher overlap suggests faithful geometric retention.

3.2 Embedding Reconstruction

KNOR reflects structural information loss during projection, indicating how well local geometric relationships among image embeddings are preserved. However, it does not identify the loss of fine-grained visual features at the patch level.

To address this, we further quantify and localize patch-level information loss by attempting to reconstruct the original vision embeddings from their projected representations.

Specifically, given a connector CONN defined in Equation 3 and set of images $I \subset \Delta^{M_1 \times M_2}$, we train a **reconstruction model** $f_\theta : (\mathbb{R}^D)^{M_C} \to (\mathbb{R}^{D'})^{M_1 \times M_2}$ to minimize reconstruction loss. For each patch index $(i, j) \in M_1 \times M_2$, we define the per-patch loss as

$$\mathcal{L}_{\text{patch}}(x, i, j) \stackrel{\text{def}}{=} \|\psi(x)_{(i,j)} - f_\theta(\mathrm{C}(\psi(x)))_{(i,j)}\|_2^2 \tag{8}$$

which measures the squared Euclidean distance between the original vision embedding and its reconstruction for each patch. The total reconstruction loss is therefore the sum of the patch-wise losses across all patches and images:

$$\mathcal{L}_{\text{recon}}(I) \stackrel{\text{def}}{=} \sum_{x \in I} \sum_{\substack{(i,j)\in \\ M_1 \times M_2}} \mathcal{L}_{\text{patch}}(x, i, j) \tag{9}$$

This patch-wise reconstruction enables us to identify and visualize the spatial distribution of information loss across the image.

4 Experimental Setup

We quantify information loss using both methods across three open-weights connector-based vision-language models on six datasets spanning question answering, captioning, and retrieval tasks. We assume that greater structural and semantic information loss during projection through the connector leads to reduced neighborhood overlap, while greater patch-wise information loss results in higher reconstruction error.

4.1 Pretrained VLMs

We consider three open-weights connector-based vision-language models including LLaVA (Liu et al., 2023), Idefics2 (Laurençon et al., 2024), and Qwen2.5-VL (Bai et al., 2025). LLaVA uses a two-layer MLP as the connector, preserving total number of patches for each image. In contrast, Idefics2 uses an attention-based perceiver resampler (Jaegle et al., 2021) that projects image embeddings to a fixed-length sequence of embeddings. Qwen2.5-VL uses a MLP-based patch merger which merges every four neighboring patch representations into one. We use the 7B-instruct model variants for LLaVA and Qwen2.5-VL, and the Idefics2-8B-instruct model.

4.2 Evaluation Datasets

We evaluate on six diverse datasets, each of which probes different aspects of visual understanding.

- **SEED-Bench** (Li et al., 2024) provides categorized multiple-choice questions spanning cognitive tasks from basic scene understanding to complex visual reasoning.
- **VizWiz Grounding VQA** (Chen et al., 2022) includes images taken by blind or low-vision individuals regarding scenarios that require visually-grounded question answering.
- **VQAv2** (Antol et al., 2015) covers open-ended questions that test general visual comprehension.
- **CUB-200-2011** (Wah et al., 2011) is a commonly used dataset for fine-grained image retrieval that covers 200 species of birds.
- **Flickr30k** (Young et al., 2014) and **COCO** (Lin et al., 2014) Karpathy test set (Karpathy and Fei-Fei, 2017) are used for image captioning evaluation.

Together, these datasets offer complementary perspectives on how different types of visual information are preserved during projection and how information loss impacts various downstream tasks.

4.3 Embedding Reconstruction Models

We build models to reconstruct image patch embeddings from connector outputs. These reconstruction models are intentionally designed with

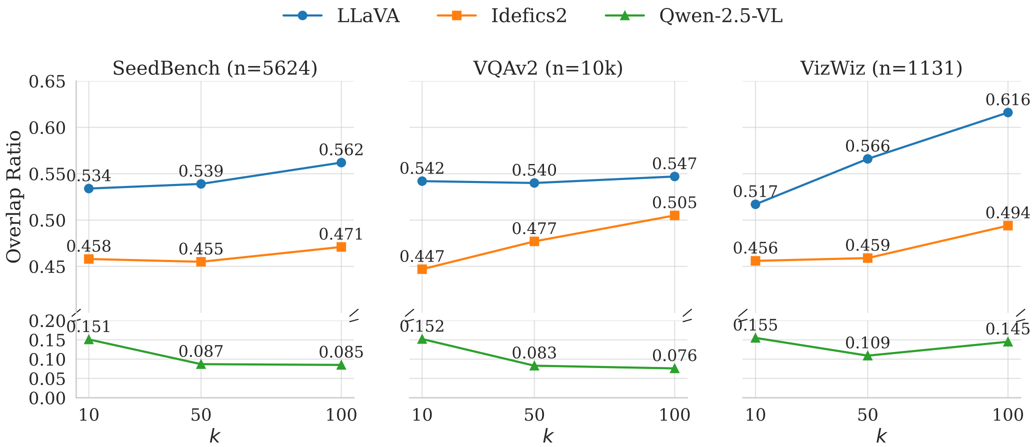


Figure 3: Neighborhood overlap ratios across three datasets: SeedBench validation, a 10,000-sample subset of VQAv2 validation, and Vizwiz grounding VQA validation. Analysis using 10, 50, and 100 nearest neighbors shows overlap ratios below 0.62 for all models, suggesting connectors poorly preserve geometric relationships and neighbor rankings for the visual representations.

Model	$M_1 M_2 \times D'$	$M_C \times D$	$\lvert$CONN$\rvert$	$\lvert f_\theta \rvert$
LLaVA	576×1024	576×4096	21M	27M
Idefics2	576×1152	64×4096	743M	844M
Qwen2.5-VL	576×1280	144×3584	45M	843M

Table 1: Model parameters and embedding dimensions. |CONN| denotes number of parameters in the connector and $|f_\theta|$ represents number of parameters of the reconstruction model. Pre- and post-projection embedding dimensions are listed as $M_1 M_2 \times D'$ and $M_C \times D$.

larger capacity than the original connectors, including expanded hidden dimensions and additional hidden layers. This controlled setup ensures our models are trained to recover the original visual representations without creating new bottlenecks in the reconstruction process.

Architecture We tailor our reconstruction models to each VLM's connector architecture. For LLaVA, which preserves the number of image patches during projection, we use a simple three-layer MLP with a 2048-dimension hidden layer. For Idefics2 and Qwen2.5-VL, which compress sequence length from $M_1 \times M_2$ to M_C, we implement transformer-based models to handle the differences in sequence length. The reconstruction model projects connector outputs to hidden embeddings with positional encodings before processing them through a 16-layer, 16-head transformer encoder with 2048-dimensional vectors. Table 1 summarizes the parameters of the reconstruction models and their input and output dimensions.

Training We train each of the embedding reconstruction models on the COCO 2017 train set (Lin et al., 2014) for 30 epochs with early stopping. We apply a learning rate of $1\text{e}{-4}$, dropout of 0.1, and a total batch size of 128. For training stability, we apply normalization to both pre- and post-projection embeddings using mean and standard deviation of the dataset.

5 Neighbor Rankings and Structural Information are Not Preserved

We calculate KNOR (Section 3.1) for images in the SeedBench validation set, a subset of the VQAv2 validation set with $10,000$ images, and the validation set of Vizwiz grounding VQA dataset. It is intuitive that higher neighborhood overlap ratios suggest that the projection better preserves the structural relationships between image embeddings. As the neighborhood rankings directly impact image retrieval tasks, we also evaluate retrieval performance on the CUB dataset using both pre- and post-connector visual embeddings.

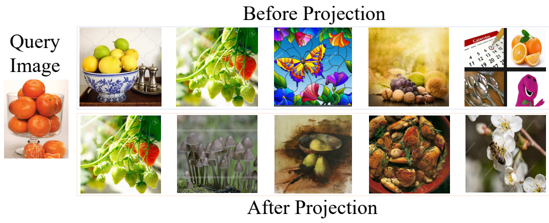
5.1 Low Overlap Ratio for All Models

In Figure 3, we show the neighborhood overlap ratio across $k = 10$, 50, and 100 nearest neighbors, averaging through all unique images in the evaluation datasets.[5] The neighborhood overlap ratios for LLaVA and Idefics-2 are around 50%. LLaVA achieves its highest overlap of 61.6% at $k = 100$, while Qwen2.5-VL loses nearly 90% of

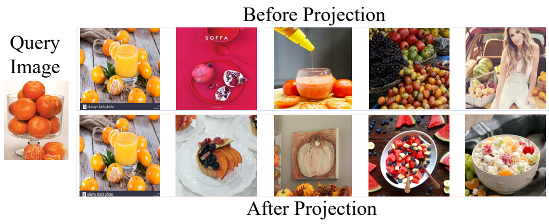
[5]Visual embeddings pre- and post-connector projection have a 1-1 mapping to the input image, and these visual embeddings are not impacted by the language model prompts.

Model	Emb	$\bar{R}$	ρ	Recall	
				R@1	R@5
LLaVA	Pre	0.40	0.08	8.34	21.82
	Post		0.11	6.16	17.22
Idefics2	Pre	0.39	0.23	**13.10**	**30.81**
	Post		**0.28**	10.87	25.28
Qwen-2.5-VL	Pre	0.08	0.10	4.23	11.74
	Post		0.11	10.65	26.44

Table 2: Zero-shot retrieval performance on the CUB test set using L^2 distance as the similarity measure. R@k denotes Recall at rank k. We also report average overlap ratio $\bar{R}$. The Spearman correlation coefficient ρ is calculated between R@5 and k-nearest neighbor overlap ratio for each sample, with $k = 100$. All correlation scores are statistically significant with $p < 1\mathrm{e}{-3}$.



(a) Five nearest neighbors of LLaVA image embeddings



(b) Five nearest neighbors of Idefics2 image embeddings

(c) Five nearest neighbors of Qwen2.5-VL image embeddings

Figure 4: Comparison of five nearest neighbors searched with pre-projection (top) and post-projection (bottom) embeddings using different models. The first image in each row is the query image, followed by its nearest neighbors. For Qwen2.5-VL, despite a low neighborhood overlap ratio, post-projection embeddings retrieve more semantically similar images.

the neighborhood ranking information. This suggests a significant reordering of nearest neighbors post-projection across all models. While LLaVA maintains higher structural preservation compared to Qwen2.5-VL and Idefics-2, it shows notable neighbor reshuffling, especially at smaller neighborhood sizes (k=10).

In Figure 4, we visualize the nearest neighbors of a given query image, revealing significant neighbor reordering across all models. However, for Qwen2.5-VL, the neighbors obtained with post-projection embeddings are more semantically similar to the query image. We suspect that this phenomenon could stem from its continuous training of the image encoder in the pretraining stage and the patch merging, which yields more semantically meaningful post-projection embeddings. Other VLMs such as LLaVA use a frozen vision encoder, where the connector is updated to inherit features from the pretrained encoder. However, in Qwen2.5-VL, continued pretraining with an unfrozen vision encoder produces fundamentally different learned visual embeddings. This indicates that the pre- and post-projection visual representations are not equivalent, but may not necessarily lead to worse semantic representations of the image.

5.2 Image Retrieval Evaluation

To verify if neighborhood reordering correlates with a degradation in the semantic representation of images, we evaluate on the CUB-200-2011 image retrieval test set (Wah et al., 2011). We perform zero-shot image retrieval with pre- and post-connector embeddings for each query image, excluding the query image itself from the gallery. The pre-and post-projection embeddings are indexed with FAISS (Douze et al., 2024), and we experiment with retrieving similar images based on both the L^2 distance and the inner product similarity (Table 8 in Appendix) of the image representations.

We report Recall@1 (R@1) and Recall@5 (R@5) in Table 2. Consistent with the neighborhood overlap visualization in Figure 4, we observe degradation in R@5 of 41.4% for LLaVA and 18.8% for Idefics2 when using post-projection image embeddings for retrieval. In contrast, Qwen2.5-VL shows improved retrieval performance with post-projection embeddings, suggesting that its low overlap ratio reflects substantial differences between pre- and post-projection representations.

To examine how structural preservation relates to retrieval, we compute Recall@5 and the 100-nearest-neighbor overlap ratio (KNOR) for each sample, then calculate their Spearman correlation.

Model	COCO	Flickr30k
Reconstruction loss (avg / std)		
LLaVA	0.087 / 0.016	0.097 / 0.019
Idefics2	0.796 / 0.082	0.854 / 0.074
Qwen-2.5-VL	1.069 / 0.117	1.069 / 0.115
Overall CIDEr Scores		
LLaVA	81.28	56.79
Idefics2	53.64	39.22
Qwen-2.5-VL	13.04	12.85

Table 3: Reconstruction loss on COCO and Flickr30k test sets. Top: reconstruction loss averaged over all samples, where LLaVA achieves lowest reconstruction error. Bottom: CIDEr scores of zero-shot captioning.[6] For both datasets, we observe better overall captioning performance with lower average reconstruction loss.

Model	COCO	Flickr30k
CIDEr Scores for High Loss / Low Loss samples		
LLaVA	73.98 / 86.96	51.79 / 61.74
Idefics2	40.84 / 66.13	29.24 / 53.22
Qwen-2.5-VL	12.45 / 13.56	13.15 / 12.35
Spearman Correlation (ρ / p)		
LLaVA	-0.077 / 0.000	-0.096 / 0.000
Idefics2	-0.214 / 0.000	-0.226 / 0.000
Qwen-2.5-VL	0.001 / 0.975	0.027 / 0.403

Table 4: Top: The comparison of CIDEr scores for top 25% highest and 25% lowest reconstruction loss samples, reported as "High Loss / Low Loss" Bottom: Spearman correlations (ρ) of per-sample reconstruction loss and captioning CIDEr scores.

As shown in Table 2, all models show a positive correlation, with coefficients of about 0.1 for Qwen2.5-VL and LLaVA, and a stronger correlation of 0.3 for Idefics2. All p-values are below $1e-3$, indicating statistical significance. This positive per-sample correlation means that, within a given model, images whose local neighborhoods are better preserved tend to achieve higher retrieval performance. For Qwen2.5-VL, we observe a small but positive per-sample correlation between local overlap ratio and recall score. This suggests that while most of the pre-projection structure was discarded to create a more semantically meaningful space, retaining certain stable neighborhoods remains advantageous for specific images. Retrieval examples are shown in Figure 11 in the Appendix.

6 Reconstruction and Model Behavior

While the neighborhood overlap ratio reflects structural information loss in the semantic representation space, we further examine the information loss at the image patch level. Specifically, as in Equation 8, we reconstruct patch-level visual representation $\psi(x)$ of an image from its projected counterpart $\text{CONN}(\psi(x))$. Higher reconstruction loss indicates greater difficulty in recovering the features that are captured in the original visual embeddings. This patch-level comparison between original and reconstructed embeddings enables us to precisely quantify and locate the visual information loss at a more fine-grained level.

6.1 Reconstruction Loss Impacts Captioning

Our embedding reconstruction evaluation follows two steps: 1) we train a reconstruction model for each VLM using paired pre- and post-projection embeddings from images in the COCO 2017 train set (as described in Section 4.3); 2) we apply these reconstruction models to predict the original image representations from their projected counterparts.

For image captioning, we measure the reconstruction loss for images in the Flickr30k validation set and COCO Karpathy test split. We use CIDEr score (Vedantam et al., 2015) to evaluate the quality of the generated captions. Table 3 summarizes the overall average reconstruction loss of the three models on the captioning test datasets. For both datasets, we observe lower average reconstruction loss yields better captioning performance. We also investigate how reconstruction loss impacts captioning for each individual image by calculating the correlation between per-sample CIDEr score and reconstruction loss per-image. In Table 4, the spearman correlation indicates higher reconstruction loss for a given image corresponds to worse captioning for Idefics2 and LLaVA, indicating by the negative correlation with p values smaller than $1e{-}5$. Please see more visualization in Figure 12. For Qwen-VL, we did not observe obvious correlation for individual images. The large gap of CIDEr scores between the highest and lowest reconstruction loss samples for LLaVA and Idefics2 suggests substantial impact on downstream tasks.

[6] We notice Qwen-2.5-VL is particularly sensitive to the task prompt; we use the prompt in the official repo: https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct.

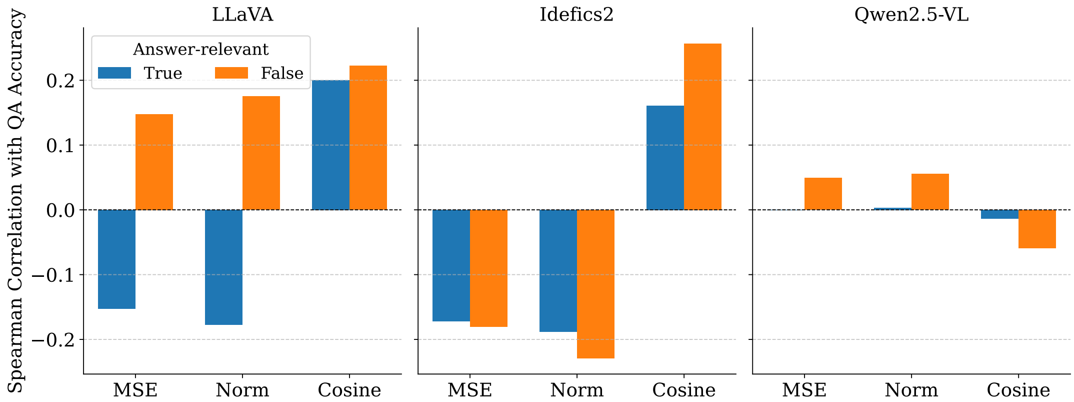


Figure 5: Correlation between reconstruction loss and question-answering accuracy on the VizWiz grounding VQA task. For LLaVA and Idefics2, all correlations have a p-value < $5e{-}5$, indicating statistically significant relationships, whereas no clear correlation is observed for Qwen2.5-VL. The reconstruction loss occurs in both answer-relevant and irrelevant patches. Loss in relevant patches negatively affects performance of LLaVA and Idefics2. "Norm" represents differences between the L^2 norm of the embeddings.

6.2 Loss at Patch-level Visual Features Explains Question Answering Behaviors

To distinguish whether the reconstruction loss stems from selective feature preservation or actual information loss, we visualize the patch-level loss for images in the VizWiz grounding VQA validation dataset. This dataset is particularly suitable for our analysis as it provides answer grounding—binary masks indicating image regions relevant to each question. By examining the relationship between the reconstruction loss for the answer-relevant image patches and question-answering accuracy, we can assess whether the projection preserves task-relevant visual information.

We report the Spearman correlation between the reconstruction loss and the question answering accuracy in Figure 5. For LLaVA, we observe a negative correlation between prediction accuracy and reconstruction loss in answer-relevant patches, while a positive correlation is found in irrelevant patches. This indicates that information loss in answer-relevant patches negatively impacts model performance, whereas loss in irrelevant patches has a less significant effect. For Idefics2, we can see that information loss in any patches would hurt question answering accuracy. We do not observe significant correlation for Qwen-2.5-VL, which is consistent with our findings in the captioning tasks.

As shown in Figure 1, identifying distorted features allows us to pinpoint visual information that becomes inaccessible or less reliable for the language model. For instance, reconstruction loss in the patches of the fifth number "8" rank among the top ten of all image patches, suggesting that the model may have struggled to answer the question due to lost details necessary for identifying the number. This analysis introduces a new visualization approach to examine VLM limitations, particularly in scenarios requiring reasoning or recognizing fine-grained visual features. Please see more visualization examples in Appendix D.

7 Analysis

Procrustes analysis We also attempt to find the optimal geometrical transformation from the post-projection embedding space to the pre-projection one through Procrustes analysis (Gower, 1975) – a method often used for supervised alignment of embeddings (Artetxe et al., 2018). The alignment error reflects the degree of structural similarity of the two embedding spaces. We use mean-pooled image embeddings from LLaVA, Idefics2, and Qwen2.5-VL. As the pre- and post-projection embeddings have different embedding dimensions and sequence lengths, our analysis follows three steps to complete the embedding alignment. We first take the mean-pooled image representation by averaging over the sequence length, producing fixed-size vectors of size D' and D. We then use PCA (Hotelling, 1933) on the mean-pooled post-projection embeddings to project them to the same dimension of the mean-pooled pre-projection embeddings.

Orthogonal transformation matrix $\mathbf{R}$ was derived through singular value decomposition of the cross-covariance matrix $\bar{X}^\top \bar{T}$, where $\bar{X} \in \mathbb{R}^{D'}$ represents mean-pooled pre-projection embeddings

Model	Mean	Std	Min	Max
LLaVA	16.62	3.16	8.76	23.65
Idefics2	4.93	0.08	4.78	5.70
Qwen2.5-VL	4.41	0.09	4.24	5.05

Table 5: Procrustes analysis results. We report the alignment error on SeedBench image representations before and after connector projection.

Model	Size	VizWiz	SeedBench	FoodieQA
MLP	27M	Avg 0.050	0.056	0.051
		Std 0.013	0.011	0.007
MLP	39M	Avg 0.064	0.070	0.065
		Std 0.015	0.013	0.008
Transformer	40M	Avg 0.237	0.231	0.228
		Std 0.019	0.025	0.014

Table 6: Reconstruction loss with different architectures across VizWiz, SeedBench, and FoodieQA datasets. Reported values include average loss (Avg) and standard deviation (Std).

and $\bar{T} \in \mathbb{R}^{D'}$ the PCA-transformed post-projection embeddings. Then the orthogonal transformation matrix is learned to best align these two sets of embeddings by minimizing the Euclidean distance. The reconstruction error are reported in Table 5. Figure 6 visualizes the alignment of LLaVA embeddings through procrustes analysis.

Our analysis reveals fundamental limitations in linear alignment of the image embeddings. The high alignment errors of 16.62 for LLaVA and 4.41 for Qwen2.5-VL indicate the inherent difficulty of preserving geometric relationships through rigid transformations. While serving as a critical baseline for structural fidelity assessment, this constrained linear approach explains why our proposed non-linear embedding reconstruction approach achieves significantly lower errors.

Ablation on Reconstruction Model Size and Structure We train three reconstruction models of different sizes for LLaVA: a 27M three-layer MLP, a 39M five-layer MLP, and a 40M Transformer. In Table 6, we observe that the 27M model is sufficient for reconstructing LLaVA visual embeddings, and a larger model does not yield better validation loss.

8 Related Work

A series of analyses has been conducted to investigate the modality gap and representation limitations of contrastive-based VLMs (Schrodi et al., 2024; Liang et al., 2022; Tong et al., 2024). These studies reveal that the representational shortcomings in CLIP embeddings subsequently impact the visual perception capabilities of VLMs relying on such vision encoders. For connector-based VLMs, Zhang et al. (2024) demonstrates that the latent space sufficiently retains the information necessary for classification through probing across different layers, and Lin et al. (2024) demonstrates the impact of different connectors on VLMs' downstream performance. However, there remains a significant gap in understanding whether fine-grained visual information, crucial for tasks such as visual grounding (Krishna et al.) and question answering (Chen et al., 2022), is lost in the process. In this paper, we focus on the connector-based models to understand the information transformation. To the best of our knowledge, our paper is the first to directly quantify information loss of the connectors from the representation perspective, offering deeper insights into where and what specific information is lost from the visual features.

9 Conclusion and Future Work

Our study provides a systematic evaluation of how connectors in existing VLMs transform information and reshape the representation space when projecting visual embeddings into the language embedding space. Through neighborhood overlap ratios and embedding reconstruction, we establish a quantitative framework that captures two critical aspects of the information loss: 1) structural shift of global semantic relationships shown by the 40-60% divergence in nearest-neighbor rankings and 2) patch-level reconstruction loss negatively impacts captioning performance and explains model failures in fine-grained visually grounded question answering. The patch-level reconstruction also enables visualization of local information loss, offering interpretable explanations for model behaviors.

Our findings suggest two key properties of an effective connector: preserving or improving semantic representation of images and preserving visual information most relevant to the text context. These findings could guide further improvements in VLM connectors. For example, the reconstruction loss at the embedding level could potentially be incorporated during model pretraining as regularization. Future work could also explore designing dynamic projection layers or better visual feature selection mechanisms for modality fusion.

Ethics Statement

We foresee no ethical concerns with our research project. In particular, ours is merely a scientific study of VLMs and provides no artifacts that can be used in a real-world scenario.

Limitations

In this study, we evaluate the information loss introduced by connectors in VLMs. However, several limitations should be noted. First, due to variations in model architectures and pretraining strategies, our findings may not generalize beyond the connector-based VLMs analyzed, particularly to architectures that employ cross-attention for modality fusion. Second, our experiments focus on models in the 7B–8B parameter range; extending the analysis to both smaller and larger models could yield deeper insights into how model scale influences information loss. Third, our pixel-level reconstruction experiments (Appendix E) produced inconclusive results, likely due to limitations of the image generation model and training dataset size. Additionally, while we quantify the information loss empirically with our k-NN overlap ratio and embedding reconstruction methods, a more formal theoretical characterization would strengthen their reliability. Finally, our reconstruction experiments cannot fully disentangle whether the observed information loss arises from the connector layer itself or from the learning limitations of the reconstruction network, suggesting that more advanced methods may be needed to refine the analysis.

Acknowledgments

Wenyan Li is supported by the Lundbeck Foundation (BrainDrugs grant: R279-2018-1145). We sincerely appreciate the help and feedback provided by Ryan Cotterell, Vésteinn Snæbjarnarson, and Clemente Pasti from the Rycolab at ETH Zürich, especially regarding the formal mathematical formulations.

References

2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024a. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

E. Fix and J.L. Hodges. 1951. *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine.

John C. Gower. 1975. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021.

Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 123(1).

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR.

Songtao Li and Hao Tang. 2024. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*.

Junyan Lin, Haoran Chen, Dawei Zhu, and Xiaoyu Shen. 2024. To preserve or to compress: An in-depth study of connector selection in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'ar, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2024. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann Lecun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings - 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. Caltech-ucsd birds 200. Technical report, California Institute of Technology.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, Zirui Wang, Afshin Dehghan, Peter Grasch, and Yinfei Yang. 2025. MM1.5: Methods, analysis & insights from multimodal LLM fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are visually-grounded language models bad at image classification? In *Advances in Neural Information Processing Systems*.

Xun Zhu, Zheng Zhang, Xi Chen, Yiming Shi, Miao Li, and Ji Wu. 2025. Connector-s: A survey of connectors in multi-modal large language models. *arXiv preprint arXiv:2502.11453*.

A Connectors in Autoregressive Vision-Language Models

Idefics2 Idefics2 leverages a perceiver resampler (Jaegle et al., 2021) as the connector. The perceiver resampler forms an attention bottleneck that encourages the latent representations to attend to the most relevant inputs in a high-dimensional input array through iterative cross-attention layers. In other words, the cross-attention module projects the high-dimensional inputs into a fixed-dimensional learned representation. Please refer to Laurençon et al. (2024) for more details.

LLaVA LLaVA (Liu et al., 2023) uses a two layer MLP to project the image embeddings to the language model's embeddings space. The MLP projector preserves the image feature length – number of patches extracted by the image encoder.

Qwen2.5-VL Qwen2.5-VL (Bai et al., 2025) uses a patch merger (two-layer MLP) to reduces the length of the input image features. The image representations of the neighboring four patches in the image are first merged, and then passed through a two-layer MLP to project the image representation to the LM embedding dimension.

B Ablation on Index Method for k-NN Overlap Ratio

We evaluated k-NN overlap ratio using three different embedding types as search indices: original embeddings, mean-pooled image embeddings, and normalized embeddings (Table 7). Since the performance differences were minimal, we selected mean-pooled embeddings for both pre- and post-projection image representations in calculating k-NN overlap ratios.

Overlap Ratio	**Index Type**					
	IndexFlatL2		IndexFlatL2 (mean pooling)		IndexFlatIP (normalized vectors)	
	mean	**std**	**mean**	**std**	**mean**	**std**
top100	0.466	0.122	0.563	0.107	0.504	0.129
top50	0.488	0.128	0.556	0.120	0.425	0.142
top10	0.490	0.149	0.551	0.160	0.377	0.161
Vector Size						
Before projection	576×1024		1×1024		576×1024	
After projection	576×4096		1×4096		576×4096	

Table 7: Ablation on KNN results when using original embeddings, mean pooled image embeddings, and normalized embeddings. We chose to use the mean-pooled embeddings for efficiency due to large embeddings size.

C Additional Evaluation Results

C.1 CUB image retrieval performance

In Table 8, we show the image retrieval performance on CUB test set using L^2 and inner product for similarity measure. The performance are consistent regardless of the index method used.

C.2 Reconstruction loss on VQA datasets

For visual question answering tasks, we measure the reconstruction loss for images in the validation set of VizWiz grounding VQA, Seed-Bench, and FoodieQA. Table 9 presents overall reconstruction loss. Among all tested models, LLaVA's projected embeddings maintain the highest reconstruction fidelity. The overall reconstruction loss reflects the overall difficulty of recovering information encoded in the visual representations.

D Visualization

D.1 Visualization for Procrustes Analysis

In Figure 6, we visualize the alignment for LLaVA pre- and post-projection embeddings, as well as the embeddings learned through the linear transformation learned. From the visualization we can observe that

Model	L2		IP	
	R@1	R@5	R@1	R@5
Pre-projection				
LLaVA	8.34	21.82	9.46	24.78
Idefics2	**13.10**	**30.81**	**13.38**	**30.98**
Qwen-2.5-VL	4.23	11.74	6.83	24.23
Post-projection				
LLaVA	6.16 ↓	17.22 ↓	5.54 ↓	20.49 ↓
Idefics2	**10.87** ↓	25.28 ↓	**10.99** ↓	25.15 ↓
Qwen-2.5-VL	10.65 ↑	**26.44** ↑	8.26 ↑	**26.70** ↑

Table 8: Zero-shot retrieval performance on CUB test set using L^2 distance and inner product for similarity measure. R@k denotes Recall at rank k. Arrows indicate performance change direction after projection.

Dataset	MSE	LLaVA	Idefics2	Qwen2.5-VL
VizWiz	Avg	**0.115**	0.907	1.069
	Std	0.086	0.298	0.684
SeedBench	Avg	**0.106**	0.872	1.069
	Std	0.071	0.307	0.610
FoodieQA	Avg	**0.113**	0.918	1.069
	Std	0.057	0.283	0.673

Table 9: Embedding reconstruction loss of images in the VizWiz, SeedBench, and FoodieQA datasets. We report both average loss (avg) and standard deviation (std). LLaVA's visual embeddings exhibit lowest reconstruction error among all models. The reconstruction performance is consistent to what we have observed for the images in COCO and Flickr30k.

the linear transformation is not able to align the pre- and post-projection embeddings well.

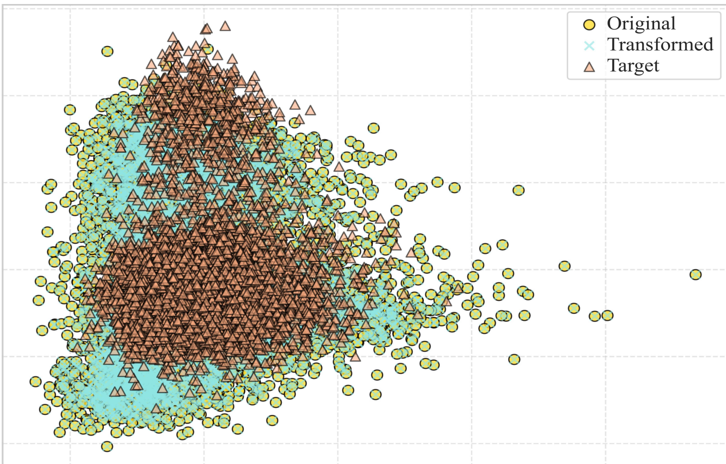


Figure 6: Alignment visualization for LLaVA pre- and post-projection embeddings through PCA.

D.2 Patch-level Loss Visualization for Vizwiz Grounding VQA

In Figure 7, we visualize additional examples of high reconstruction loss patches that contributes to model's failure on answering questions that requires recognizing text in the objects.

D.3 Visualization of Neighborhood Reordering

In Figure 10, we present more k-NN examples on comparison of searching with pre-projection (top) v.s. post-projection (bottom) embeddings. In Figure 11, we present CUB image retrieval visualization with pre- and post-projection embeddings.

D.4 Visualization of reconstruction loss and captioning performance

E Image Reconstruction with Different Embeddings

Beyond neighbor-overlapping and embedding reconstruction, we aim to investigate how information loss manifests in the reconstructed images themselves. To explore this, we project different representations of visual features onto the input embedding space of a powerful image decoder to assess their reconstruction

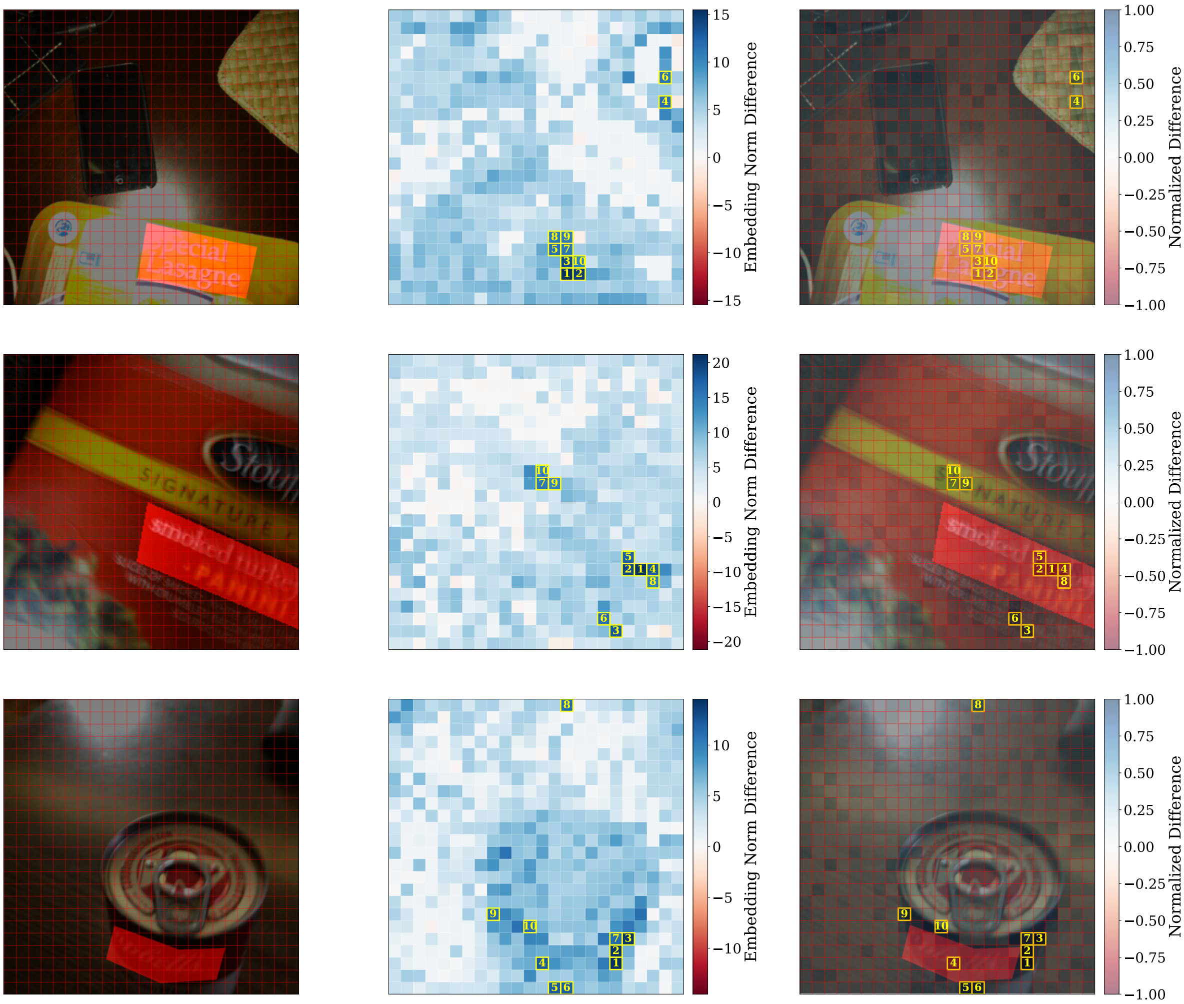


Figure 7: Additional visualization of high reconstruction loss patches that contributes to model's failure on answering questions that requires recognizing text in the objects. Left: input images with answer-relevant regions in red masks. Middle: signed difference between post-projection embeddings norms and pre-projection embedding norms. Right: normalized norm differences overlay with the input image, with highest loss patches marked in yellow.

quality. However, image reconstruction performance depends on various factors, including the expressiveness of the image decoder. As such, this section serves as a preliminary exploration, and we encourage future work in this direction.

For our experiments, we use a fine-tuned VAE decoder[7], trained on the original VAE checkpoint from Stable Diffusion, trying to alleviate the influence of the decoder as a limiting factor in reconstruction quality. To align the sequence length between the vision encoder in the VLM and the expected input length of the VAE decoder, we employ a 6-layer Transformer encoder-decoder module with 4 attention heads. We train the aligner module on the COCO 2017 training set for 100 epochs with three objectives: 1) Embedding loss minimizing the difference between the VAE encoder embeddings and the aligned embeddings from the VLM's visual encoder; 2) Reconstruction loss measuring the mean squared error (MSE) between the original and reconstructed images; 3) Latent loss quantifying the divergence between the mean and variance of the Gaussian distribution for diffusion.

For the VLM, we use the LLaVA model in our experiments. We evaluate reconstruction performance on both an in-distribution image from the COCO 2017 dev split and an out-of-distribution image, as shown in Figure 13. When using embeddings before projection, the overall pixel-wise MSE reconstruction loss is 0.2128, compared to 0.2443 after projection. Figure 13 illustrates the reconstructed images for both cases, where pre-projection embeddings yield similar contour preservation with post-projection embeddings.

[7]https://huggingface.co/stabilityai/sd-vae-ft-mse

Figure 8: Idefics high kNN overlap ratio example, where we can observe the reordering among semantically similar vision embeddings.

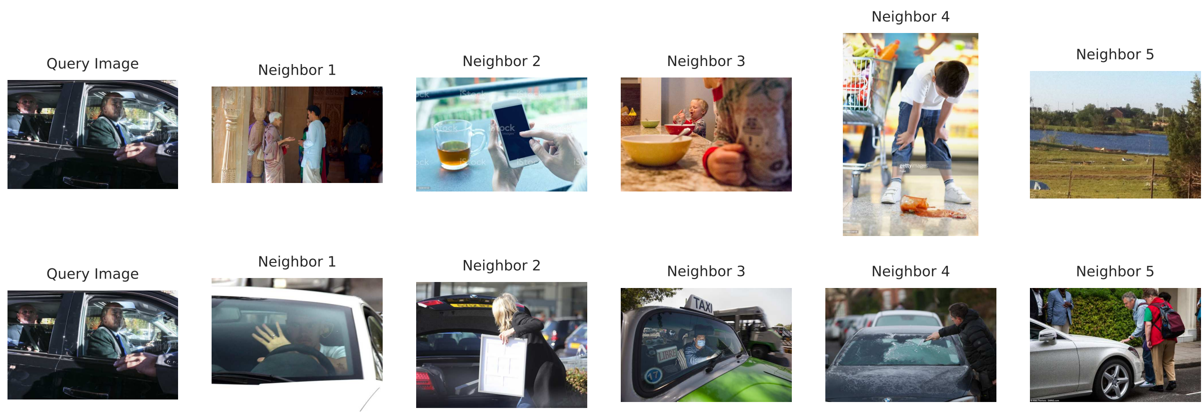


Figure 9: Qwen kNN example where the post-projection embeddings are better at retrieving semantically similar images (bottom).

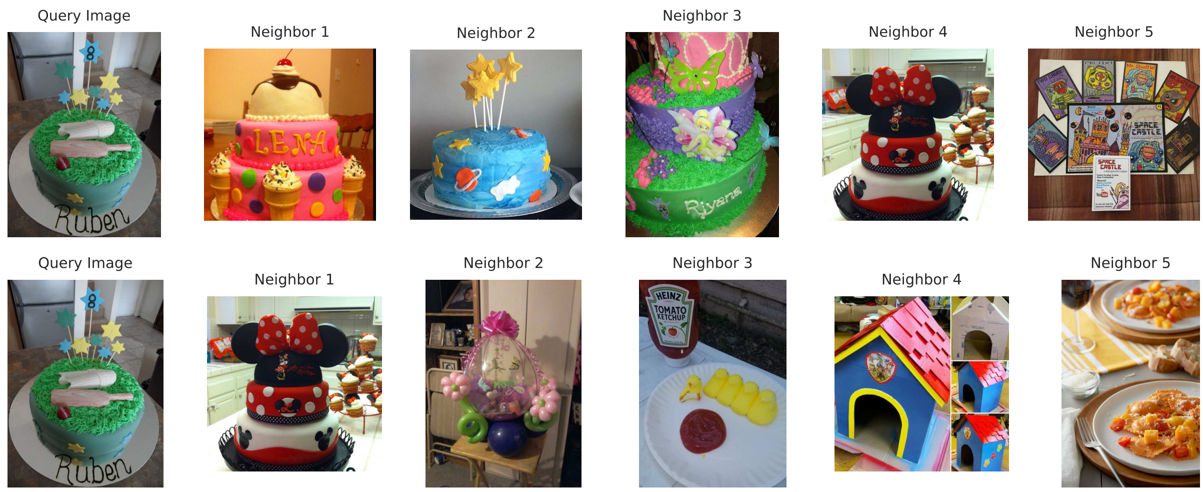
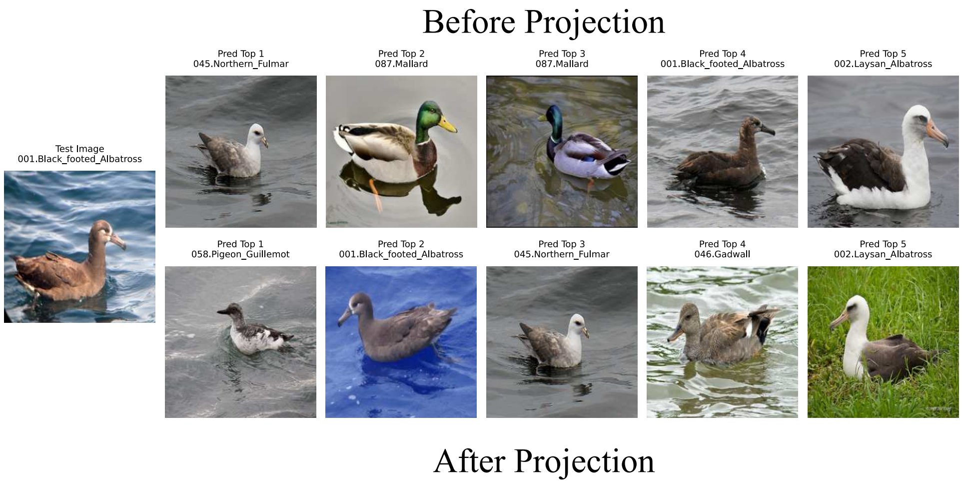


Figure 10: LLaVA low kNN overlap ratio example. We can observe the degradation in post-projection embedding.

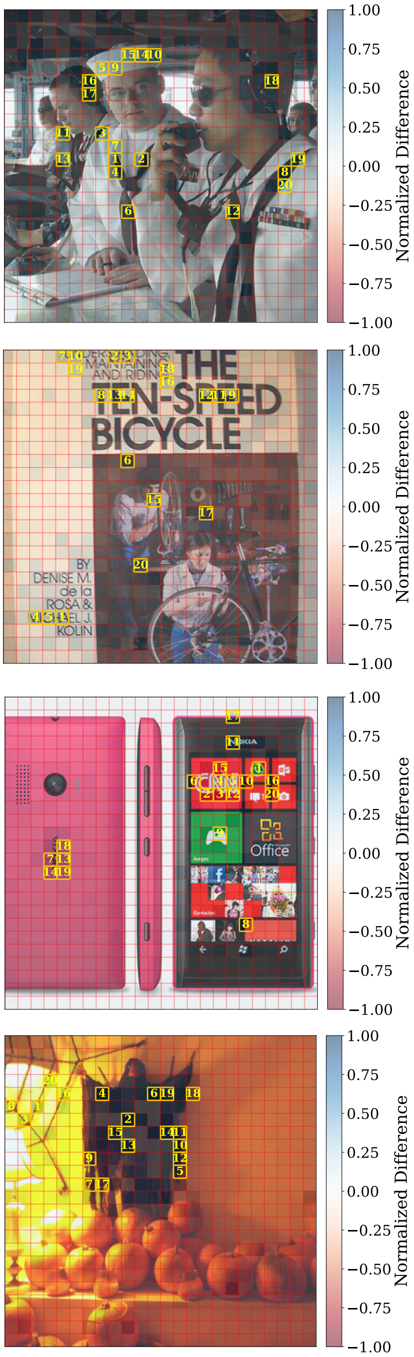
(a) Top five retrieved images of LLaVA image embeddings



(b) Top five retrieved images of Idefics2 image embeddings

(c) Top five retrieved images of Qwen2.5-VL image embeddings

Figure 11: Comparison of top five retrieved images of pre-projection (top) and post-projection (bottom) embeddings using different models on CUB test sample. Zero-shot retrieval based on fine-grained visual details is hard for all tested models.



Reference Captions

- People in **navy** uniforms and one person talking on a walkie- talkie.
- Group of **sailors** in command center with one talking on walkie talkie.

Generated Caption

- A group of men in uniforms are sitting at a table.

Reference Captions

- A book about understanding and maintaining a **ten-speed bicycle**.
- A sign explaining the components of **a 10 speed bike.**

Generated Caption

- A man and a woman are working on a bicycle.

Reference Captions

- Various angle shots of the **Nokia** Windows cell phone.
- A pink smartphone with Windows 8 on the screen.

Generated Caption

- A pink cell phone is displayed next to a red screen.

Reference Captions

- The Halloween display includes a **spiderweb** and lots of pumpkins.
- Multiple pumpkins and a skeleton on the wall.

Generated Caption

- A spooky Halloween display features a witch figure and a bunch of pumpkins.

Figure 12: Visualization of low CIDEr score captioning samples and the reconstruction loss overlay with the input image. We can observe that details regarding the high loss patches are missing from the generated captions. High loss patches are marked in yellow squares.

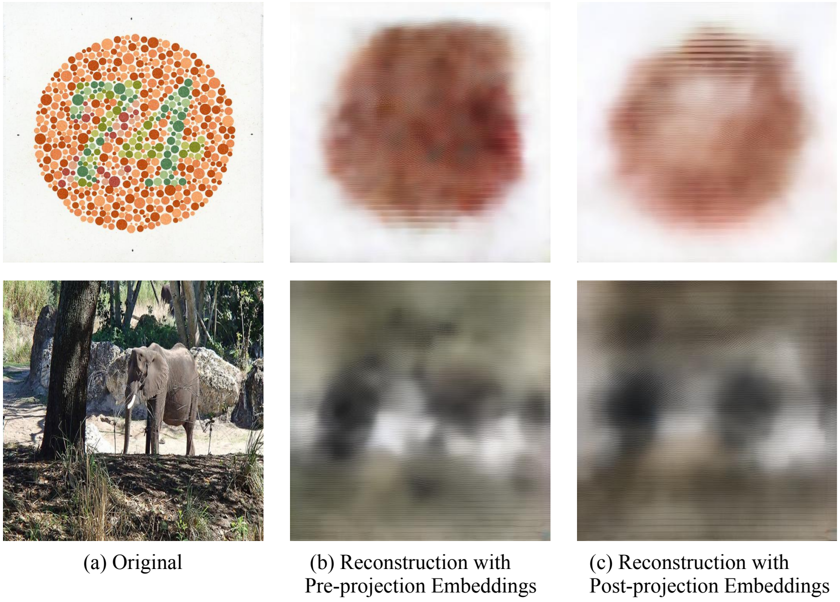


Figure 13: Image reconstruction with LLaVA pre-and post-projection embeddings on out-of-distribution (top) and in-distribution (bottom) examples.