Lost in Embeddings: Information Loss in Vision-Language Models

Anonymous ACL submission

Abstract

Vision–language models typically process visual inputs through a pretrained vision encoder followed by projection into the language model's embedding space. While crucial for modality fusion, this projection step induces under-characterized information loss that directly impacts model capabilities. We propose two novel approaches to quantify visual information loss introduced at this projection step. First, we evaluate the preservation of semantic information and structural relationships by analyzing changes in nearest-neighbor rankings between representations. Second, to locate information loss for the image representation at a patch level, we directly measure information loss through visual embedding reconstruction. Focusing on connector-based VLMs, our experiments reveal projection layers fundamentally alter visual semantic relationships – nearest neighbor similarity rankings diverge by 40-60% post-projection, directly explaining observed retrieval performance drops. Our embedding reconstruction approach provides interpretable insights for model behavior on visual question-answering tasks, finding that areas of high information loss reliably predict instances where models struggle.

1 Introduction

Vision–language models (VLMs) have demonstrated remarkable capabilities in visual question answering tasks by leveraging pretrained vision encoders. A series of models employ connector modules to bridge the semantic gap between visual and textual modalities, projecting visual representations into embedding sequences that language models 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), as implemented in LLaVA (Liu et al., 2023), or more sophisticated transformer-based perceiver sampler used in Idefics (Laurençon et al., 2024) that convert image patches to a fixed-length sequence of visual tokens.

While these connector modules enable efficient cross-modal integration (Li and Tang, 2024), their impact on information fidelity remains poorly understood. The transformation of rich visual features into a format compatible with language models inevitably involves dimensional conversion and representation restructuring. This raises fundamental questions about the nature and extent of potential *information loss* during this critical projection step. As highlighted in Figure 1, such information loss could impose inherent limitations on the model's reasoning capabilities, as the language model's performance is bounded by the quality and completeness of the visual information it receives. Despite the growing body of research on VLM connector architectures and their downstream performance (Lin et al., 2024), there has been limited systematic investigation into how different connector designs correlate visual information loss in the latent space.

To bridge this gap in the literature, we present a comprehensive evaluation framework to quantify information loss in VLM connector modules. We first measure information loss through careful examination of the geometric structure of latent visual representations. Then, through patch-level visual feature reconstruction, we are able to pinpoint the high-loss regions in the image — areas where visual features are hard to recover after projection. This two-step approach provides both quantitative metrics and interpretable visualizations, offering insights into the nature of information transformation during vision-text integration.

The main findings of this paper are:

- We propose a novel evaluation framework comprising two approaches to quantify the information loss at the connector component for vision-language models.
- Our neighborhood overlap analysis shows significant degradation of 40%–60% of geomet-

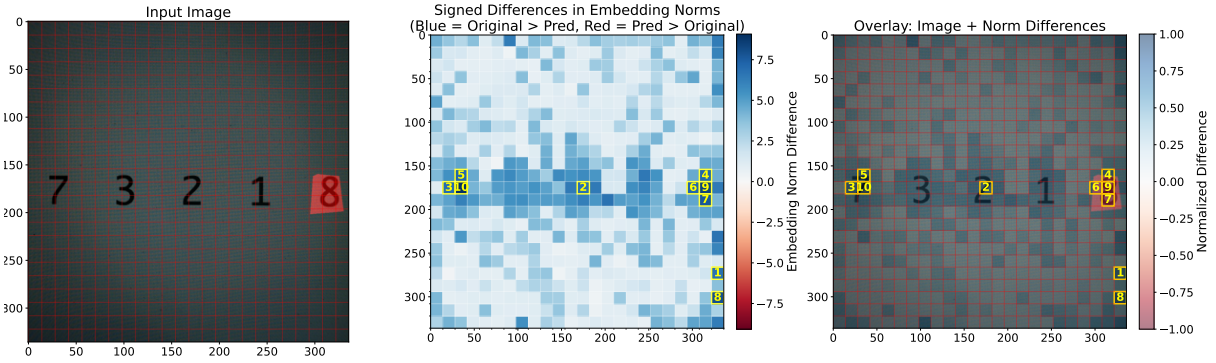


Figure 1: Example of visualization of patch-wise information loss in the embeddings explains the incorrect predicted answer in VizWiz Grounding VQA. For the question "What is the fifth number?", LLaVA incorrectly predicted "18". The signed differences is difference between the L^2 norm of the original and reconstructed patch embeddings. The yellow squares highlight the top ten high loss patches that contribute to the wrong prediction.

ric relationships during the projection process across all tested models, and the preservation of local structure varies across different model and datasets.

- Our embedding reconstruction identifies patch-level information loss and provides interpretable visualizations for error analysis in VLMs, directly linking local information loss to model performance.

2 Related Work

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

Lin et al. (2024) categorize connectors into two types: feature preserving and feature compressing connectors. Feature preserving connector including MLPs that preserves the patch numbers and embedding dimensions, such as the two-layer MLP connector in LLaVA. While feature compressing connectors project vision embeddings to a reduced number of patch embeddings, which often involves transformer-based or convolution architecture, and pooling over the original vision embedding. The feature compressing category includes connectors such as the perceiver sampler in Idefics2 (Laurençon et al., 2024) and the patch merger in Qwen-2-VL (Wang et al., 2024). In this paper, we estimate information loss considering both type of connectors.

Limitations & Analysis of VLMs 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 impacts 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) demonstrate 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 and visual question answering, 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.

3 Preliminaries

In this paper, we consider vision-language models that consist of a vision encoder, a text encoder, and a *connector* module for modality fusion. Specifically, for an input image $\boldsymbol{x} \in \mathbb{R}^{w \times h \times c}$ (width w, height h, channels c), the **visual encoder**

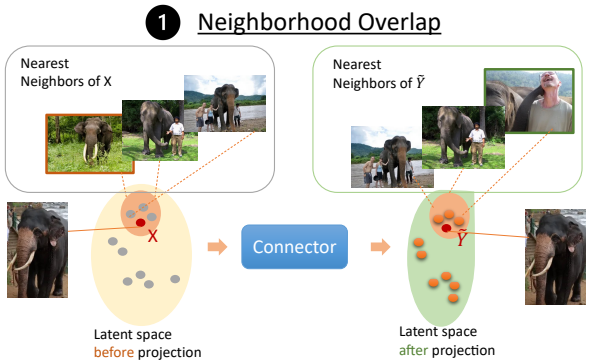


Figure 2: For an image, we can calculate the overlap of its neighbors before and after projection. For example, the overlap ratio for the given image in this figure is 0.67 as two of its three nearest neighbors are the same in both representation space.

$\phi_v : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{n_v\times d_v}$ produces a sequence of patch embeddings $\mathbf{X} = \phi_v(\boldsymbol{x})$, where n_v is the number of patches and d_v is the embedding dimension. We write Σ^* to denote the set of text sequences over an alphabet of characters Σ. The **text encoder** $\phi_\ell : \Sigma^* \to \mathbb{R}^{n_\ell\times d_\ell}$ converts the text sequence $\boldsymbol{y} \in \Sigma^*$ to text embeddings $\mathbf{Y} = \phi_\ell(\boldsymbol{y})$, with n_ℓ the sequence length and d_ℓ the text embedding dimension.

After the image embeddings are obtained, the **connector** $\textsc{conn}: \mathbb{R}^{n_v\times d_v} \to \mathbb{R}^{n_\ell\times d_\ell}$ projects the visual embeddings $\mathbf{X}$ into the language model's space, producing the projected representations $\widetilde{\mathbf{Y}} = \textsc{conn}(\mathbf{X})$. The projected image representations are then concatenated with the text representations as a combined input sequence $[\widetilde{\mathbf{Y}}; \mathbf{Y}]$. The language model further processes this sequence and predicts probability distribution over the next tokens. Please see formal definition in Appendix A.

4 Quantifying Information Loss

We propose two methods for quantifying information loss during the critical projection step in VLMs — where the connector projects visual features of an image into the shared semantic space used for language understanding. The first method, as illustrated in Figure 4, quantifies structural perseveration of semantic embeddings by measuring the overlap between each image representation's k-Nearest Neighbors (k-NN, Cover and Hart (1967)) before and after projection. The second method evaluates patch-level representation distortion by training an *ad hoc* neural network to reconstruct the original image embedding from its projected representation (Figure 3).

4.1 k-Nearest Neighbors Overlap Ratio

To quantify geometric information loss during projection in visual representation spaces, we propose the **k-nearest neighbors overlap ratio**, a metric grounded in the preservation of the k-NN relationship (Cover and Hart, 1967) between data points. Specifically, consider a set of unique images $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. For each image $\boldsymbol{x}_i, i = 1, \ldots, N$, let $\mathbf{X}_i = \phi_v(\boldsymbol{x}_i) \in \mathbb{R}^{n_v\times d_v}$ be its original vision embedding, and $\widetilde{\mathbf{Y}}_i = \textsc{conn}(\mathbf{X}_i) \in \mathbb{R}^{n_v\times d_v}$ be its projected embedding through the connector. We define the k-NN overlap ratio for image $\boldsymbol{x}_i$ as

$$\mathrm{R}(i,k) = \frac{\left|\mathcal{N}_{[\mathbf{X}]}(\mathbf{X}_i, k) \cap \mathcal{N}_{[\widetilde{\mathbf{Y}}]}(\widetilde{\mathbf{Y}}_i, k)\right|}{k} \quad (1)$$

Where $\mathcal{N}_{[\mathbf{X}]}(\mathbf{X}_i, k)$ is the set of k-nearest neighbors of $\mathbf{X}_i$ among the pre-projection embeddings of all other images $[\mathbf{X}] = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$. Likewise, $\mathcal{N}_{[\widetilde{\mathbf{Y}}]}(\widetilde{\mathbf{Y}}_i, k)$ is the set of k-nearest neighbors of $\widetilde{\mathbf{Y}}_i$ among the projected embeddings of all the other images $[\widetilde{\mathbf{Y}}] = \{\widetilde{\mathbf{Y}}_1, \ldots, \widetilde{\mathbf{Y}}_N\}$. As a global measure, the *average* overlap ratio is calculated as

$$\overline{\mathrm{R}}(k) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{R}(i,k) \quad (2)$$

An ideal projection would preserve the local geometric structure, ensuring that the k-NN sets for $\mathbf{X}_i$ and $\widetilde{\mathbf{Y}}_i$ remain identical for each image $\boldsymbol{x}_i$. Deviations in these neighborhoods—measured by the overlap ratio (Equation 1)—reflect information loss introduced during projection. Lower overlap indicates greater distortion, while higher overlap suggests faithful geometric retention.

4.2 Embedding Reconstruction

While neighborhood overlap ratios reveal structural information loss during projection—indicating how well geometric relationships between embeddings are preserved—they do not identify patch-level visual feature loss during the connector projection.

To address this, we further quantify and localize patch-level information loss in the embedding space by attempting to reconstruct the original vision encoder embeddings from their projected representations. Specifically, given a dataset of images $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, we train a **reconstruction**

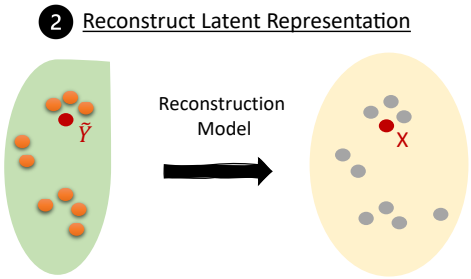


Figure 3: We can also quantify the information loss by reconstructing the visual representation from the projected latent vectors of a given image.

model $f_\theta : \mathbb{R}^{n_\ell \times d_\ell} \rightarrow \mathbb{R}^{n_v \times d_v}$ to minimize the following reconstruction loss

$$\mathcal{L}_{\text{recon}}(\mathcal{D}) = \sum_{i=1}^{N} \|\mathbf{X}_i - f_\theta(\widetilde{\mathbf{Y}}_i)\| \qquad (3)$$

where $\|\cdot\|$ denotes the Euclidean distance, $\mathbf{X}_i \in \mathbb{R}^{n_v \times d_v}$ is the original embedding sequence for the image $\boldsymbol{x}_i$, and $\widetilde{\mathbf{Y}}_i \in \mathbb{R}^{n_\ell \times d_\ell}$ is the connector projected embedding. For each image $\boldsymbol{x}_i$, the reconstruction loss $\|\mathbf{X}_i - f_\theta(\widetilde{\mathbf{Y}}_i)\|$ yields a loss matrix of the same size as the original embedding, which enables patch-level visualization.

5 Experimental Setup

We test our information loss hypothesis by experimenting with three open-weights connector-based vision-language models across five evaluation datasets, including visual question answering and image retrieval tasks.

5.1 Evaluation Datasets and VLMs

We evaluate on five diverse evaluation datasets, each probes different aspects of visual understanding.

- **SEED-Bench** (Li et al., 2023a) provides categoriezed multiple-choice questions spanning cognitive tasks from basic scene understanding to complex visual reasoning.
- **FoodieQA** (Li et al., 2024) focuses on more fine-grained feature understanding in the food domain through multiple-choice questions.
- **VizWiz Grounding VQA** (Chen et al., 2022) includes real-world visual assistance scenarios with grounding-based 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.

Together, these datasets offer complementary perspectives on how different types of visual information are preserved during projection.

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 (Wang et al., 2024). LLaVA uses a two-layer MLP as the connector, preserving total number of patches for each image. In contrast, Idefics2 uses a attention-based perceiver resampler (Jaegle et al., 2021) that projects image embeddings to a fixed-length 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.

5.2 Embedding Reconstruction Models

We build models to reconstruct image patch embeddings from connector outputs. These reconstruction models are intentionally designed with 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 As the connector in the LLaVA model preserves the number of image patches before and after the projection of the visual embeddings, we use a simple 3-layer MLP with a hidden dimension of 2048. For Idefics2 and Qwen2.5-VL, whose connector reduces the sequence length of the embeddings from n_v to n_l, we first project the connector outputs to hidden embeddings, combined with learnable positional encodings, and then process it through a 16-layer transformer encoder with 16 attention heads. The hidden vector dimension is 2048. The parameters of the reconstruction models and their input and output dimensions are reported in Table 1.

Training We train each of the embedding reconstruction models on the COCO 2017 train set images (Lin et al., 2014) for 30 epochs with early stopping. We apply a learning rate of $1e-4$ and dropout of 0.1, and a total batch size of 128.

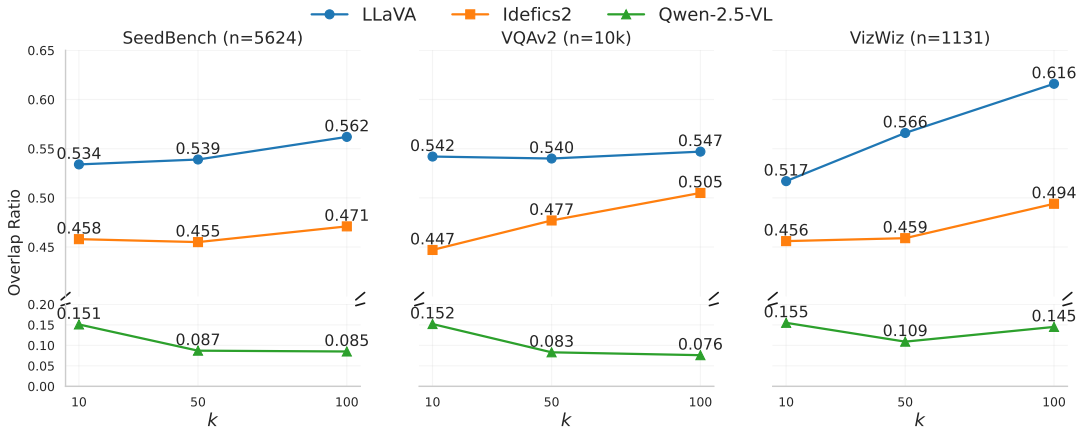


Figure 4: Neighborhood overlap ratios across three datasets: SeedBench validation set, a 10,000-sample subset of VQAv2 validation, and Vizwiz grounding VQA validation. Analysis using 10, 50, and 100 nearest neighbors shows overlap ratios below 0.62 for all models, suggesting connectors poorly preserve distance relationships and neighbor rankings for the visual representations.

Model	$\|M_C\|$	E_{Pre}	E_{Post}	$\|M_R\|$
LLaVA	21M	576×1024	576×4096	27M
Idefics2	743M	576×1152	64×4096	844M
Qwen2.5-VL	45M	576×1280	144×3584	843M

Table 1: Model parameters and embedding dimensions. $|M_C|$ denotes number of parameters in the connector and $|M_R|$ represents number of parameters of the reconstruction model. E_{Pre} and E_{Post} refer to pre- and post-projection embedding dimensions, respectively.

6 Neighbor Rankings and Semantic Information are Not Preserved

We evaluate the neighborhood overlap ratio (Section 4.1) using images in the SeedBench validation set, a subset of the VQAv2 validation set with $10,000$ elements, and the validation set of Vizwiz grounding VQA dataset. It is intuitive that higher neighborhood overlap ratios suggest that the projection better preserves the relationships between visual embeddings. As the neighborhood rankings directly relates to the image retrieval task, we also evaluate retrieval performance using both pre- and post-connector visual embeddings.

6.1 Low Overlap Ratio for All Models

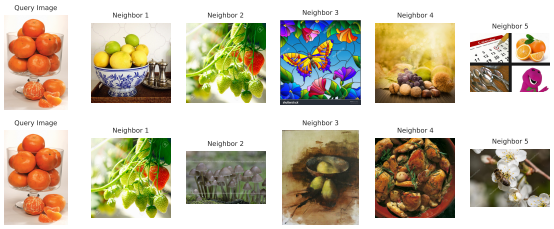
In Figure 4, we show the neighborhood overlap ratio across $k = 10$, 50, and 100 nearest neighbors, averaging through all unique images in the evaluation datasets.[1] We can observe that the neighborhood overlap ratios are around 50% for all three models, with LLaVA achieving 61.6% overlap as the maximum when considering 100 nearest neighbors. This suggests a significant reordering of nearest neighbors post-projection across all models. Specifically, LLaVA maintains higher structural preservation compared to Qwen2.5-VL and Idefics-2, whereas Qwen2.5-VL lost almost 90% of the neighborhood ranking information. However, even LLaVA shows notable neighbor reshuffling, especially at smaller neighborhood sizes (k=10).

In Figure 5, 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 this does not necessarily lead to worse semantic representations of the image.

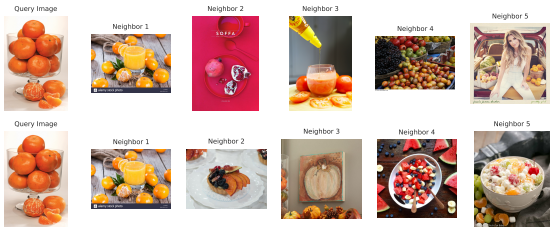
6.2 Image Retrieval Evaluation

To verify if neighborhood reordering correlates with a degradation in the semantic representation of images, we evaluate on the CUB-200-2011 im-

[1] 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.



(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 5: 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.

age 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 of the image representations.

We report the recall scores at rank 1 (R@1) and rank 5 (R@5) in Table 2. Consistent with our observations from the neighborhood overlap visualization (Figure 5), we observe semantic degradation of 41.4% and 18.8% of R@5 for for LLaVA and Idefics model, respectively. In contrast, for the Qwen2.5-VL model, the improved image retrieval performance with post-projection embeddings suggests that the low overlap ratio stems from the substantial differences between the two sets of visual

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 2: 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 3: MSE loss for embedding reconstruction 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 test representations.

embeddings, with the post-projection embeddings capturing more semantic features. This suggests that post-projection image representations can be used for image retrieval tasks for Qwen2.5-VL, while the original image representations work better for the other two models.

7 Reconstruction Loss and Model Behavior

While the neighborhood overlap ratio reflects information loss in semantic representations and latent space geometry, we further examine the information loss at the image patch level. Specifically, we reconstruct patch-level visual representation $\mathbf{X}$ of an image from its projected counterpart $\widetilde{\mathbf{Y}}$ (Figure 3). Higher reconstruction loss indicates greater difficulty in recovering the features that was 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 that is lost during connector projection.

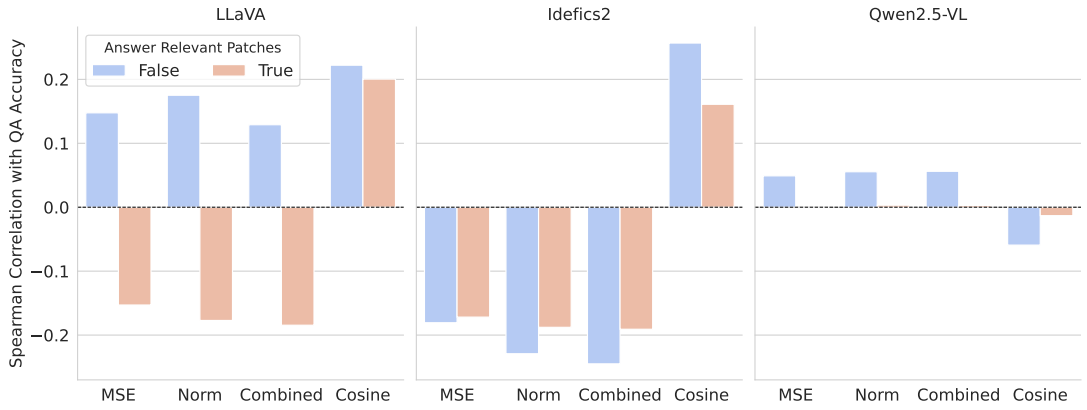


Figure 6: Correlation between reconstruction loss and question-answering accuracy. For LLaVA and Idefics2, all correlations have a p-value < 5e-5, indicating statistically significant relationships, whereas no clear correlation is observed for Qwen2.5-VL. The reconstruction loss occurs in both answer-relevant and irrelevant patches. Loss in relevant patches has negatively affects performance of LLaVA and Idefics2. “Norm" represents differences between the L^2 norm of the embeddings. “Combine" represents the norm difference weighted by cosine similarities.

7.1 Reconstruction Loss

Our embedding reconstruction evaluation follows two steps: 1) We first train a reconstruction model for each VLM, using pre- and post-projection embedding pairs of the images in the COCO 2017 train set; 2) Then we use the reconstruction models to predict image representations from their projected counterparts.

For training stability, we apply normalization to both pre- and post-projection embeddings using mean and standard deviation of the dataset. We measure the reconstruction loss for images in the validation set of VizWiz grounding VQA, SeedBench, and FoodieQA.

Table 3 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. We further visualize the patch-level loss to test whether the transformation through the connector projection acts as a selective filter of visual features or results in genuine information loss. This distinction helps clarify if the projection merely prioritizes certain visual aspects while preserving them, or if it actually degrades the encoded visual information.

7.2 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 6. 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.

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 viusal features. Please see more visualization examples in Appendix C.

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

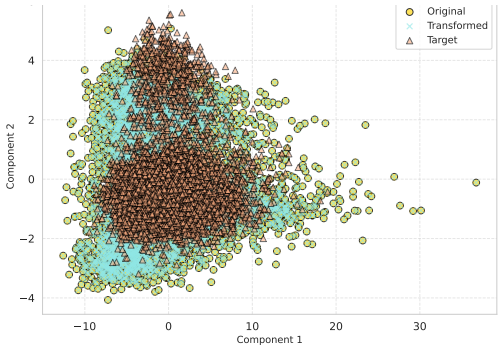


Figure 7: Alignment visualization for LLaVA pre- and post-projection embeddings through PCA.

8 Analysis

Procrustes analysis We performed Procrustes analysis (Gower, 1975) on 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 n_v and n_l. We then use PCA (Maćkiewicz and Ratajczak, 1993) 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 $\bar{X} \in \mathbb{R}^{n_v}$ represents mean-pooled pre-projection embeddings and $\bar{T} \in \mathbb{R}^{n_v}$ 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 4. We visualize the LLaVA embedding alignment with PCA in Figure 7.

Our Procrustes analysis reveals fundamental limitations in linear alignment of the image embeddings, with high alignment errors of 16.62 for LLaVA and 4.41 for Qwen2.5-VL demonstrating the inherent difficulty of preserving geometric relationships through rigid transformations. While this method establishes a critical baseline for structural fidelity assessment, its constrained linear formulation explains why our non-linear reconstruction approaches achieve significantly lower errors.

Ablation on Reconstruction Model Size and Structure We build three reconstruction models of different size for LLaVA: a 27M three layers MLP, a 39M five-layer MLP, and a 40M Transformer. In Table 5, we can observed that the 27M model is sufficient for reconstructing LLaVA visual embeddings and a larger model does not yield better validation loss.

Model	Size	VizWiz	SeedBench	FoodieQA
MLP	27M	Avg 0.050 Std 0.013	0.056 0.011	0.051 0.007
MLP	39M	Avg 0.064 Std 0.015	0.070 0.013	0.065 0.0075
Transformer	40M	Avg 0.237 Std 0.019	0.231 0.025	0.228 0.014

Table 5: Evaluation of MSE across VizWiz, SeedBench, and FoodieQA datasets. Reported values include average loss (Avg) and standard deviation (Std).

9 Conclusion and Future Work

Our study provides a systematic evaluation of how connectors in vision-language models (VLMs) induce information loss 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 correlating with 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.

In the future, we will investigate information loss quantification through image reconstruction to quantify information loss at a pixel-level. We hope to further explore approaches to mitigate the impact of information loss in VLMs, such as incorporating better feature selection mechanisms.

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 be specific to the connector-based VLMs analyzed and may not generalize to architectures that employ cross-attention for modality fusion. Second, our experiments focus on connectors in VLMs within the 7B–8B parameter range. Expanding the analysis to models of different sizes could provide deeper insights into the relationship between model scale and information loss. Third, our pixel-level reconstruction experiments (Appendix D) yielded inconclusive results in quantifying information loss, possibly due to limitations in our chosen image generation model and training dataset size. Additionally, while we empirically validate our k-NN overlap ratio and embedding reconstruction metrics, a formal theoretical characterization would further strengthen their reliability. Finally, our reconstruction experiments cannot conclusively determine whether the observed information loss stems from the connector layer itself or from potential learning limitations of the trained reconstruction network.

References

2024. The llama 3 herd of models.

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)*.

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.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

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.

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

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.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. 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*.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

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*. Association for Computational Linguistics.

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.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

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*.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. Caltech-ucsd birds 200. Technical Report CNS-TR-2011-001, California Institute of Technology.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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? *Conference on Neural Information Processing Systems (NeurIPS)*.

A Autoregressive Vision-Language Models

A **string** $\boldsymbol{y}$ is a sequence of characters from an alphabet Σ, we denote with Σ^* the set of possible strings over Σ. We define an image $\boldsymbol{x}$ as an array in $\mathbb{R}^{w\times h\times c}$, where w denotes the image's width, h its height, and c the number of color channels.

A **vision-language model** over strings in Σ^*, and images in $\mathbb{R}^{w\times h\times c}$ is a probability distribution $p_\theta : \Sigma^* \times \mathbb{R}^{h\times w\times c} \to [0,1]$ where θ is a set of learnable parameters. In this paper, we are mostly interested in **autoregressive** vision–language models, i.e. language models that are defined through a conditional next token distribution ℓ_θ:

$$p_\theta(\boldsymbol{y}, \boldsymbol{x}) = \ell_\theta(\text{EOS} \mid \boldsymbol{y}, \boldsymbol{x}) \prod_{t=1}^{|\boldsymbol{y}|} \ell_\theta(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) \tag{4}$$

Where EOS is a distinguished end-of-sequence symbol, and we write $\overline{\Sigma} = \Sigma \cup \{\text{EOS}\}$ to denote the alphabet augmented with EOS.

We assume the following general structure on the vision-language model. The input string $\boldsymbol{y}_{<t}$ is mapped by the **language encoder** $\phi_\ell : \Sigma^* \to \mathbb{R}^{d_\ell \times n_\ell}$ to a vector representation $\mathbf{Y} = \phi_\ell(\boldsymbol{y}_{<t})$ of the string, where d_ℓ and n_ℓ are respectively the context window size and the embedding dimension of the language encoder. [2]

B Ablation on Index Method for k-NN

We evaluated k-NN overlap ratio using three different embedding types as search indices: original embeddings, mean-pooled image embeddings, and normalized embeddings (Table 6). 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.

[2] In practice, we assume that the context window is fixed, this means that sequences longer than n_ℓ will be truncated, and shorter sequences will be padded.

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 6: 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.

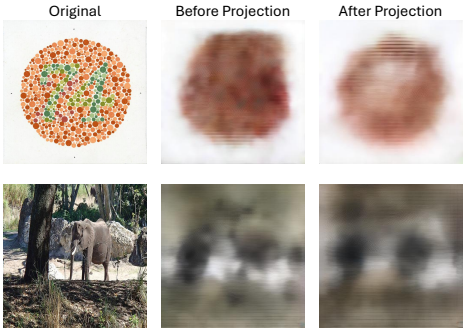


Figure 8: Image reconstruction with LLaVA model on in-distribution and out-of-distribution examples.

C Visualization

C.1 Patch-level Loss Visualization for Vizwiz Grounding VQA

In Figure 9, we visualize examples of high reconstruction loss patches that contributes to model's failure on answering questions that requires recognizing text in the objects (red regions are the answer grounding masks).

C.2 Visualization of Neighborhood Reordering

In Figure 12, we present more kNN examples on comparison of searching with pre-projection (top) v.s. post-projection (bottom) embeddings.

D 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 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[3], 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 8. When using embeddings before projection, the overall pixel-wise MSE reconstruction loss is 0.2128, compared to 0.2443 after projection. Figure 8 illustrates the reconstructed images for both cases, where pre-projection embeddings yield similar contour preservation with post-projection embeddings. We leave this for fur-

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

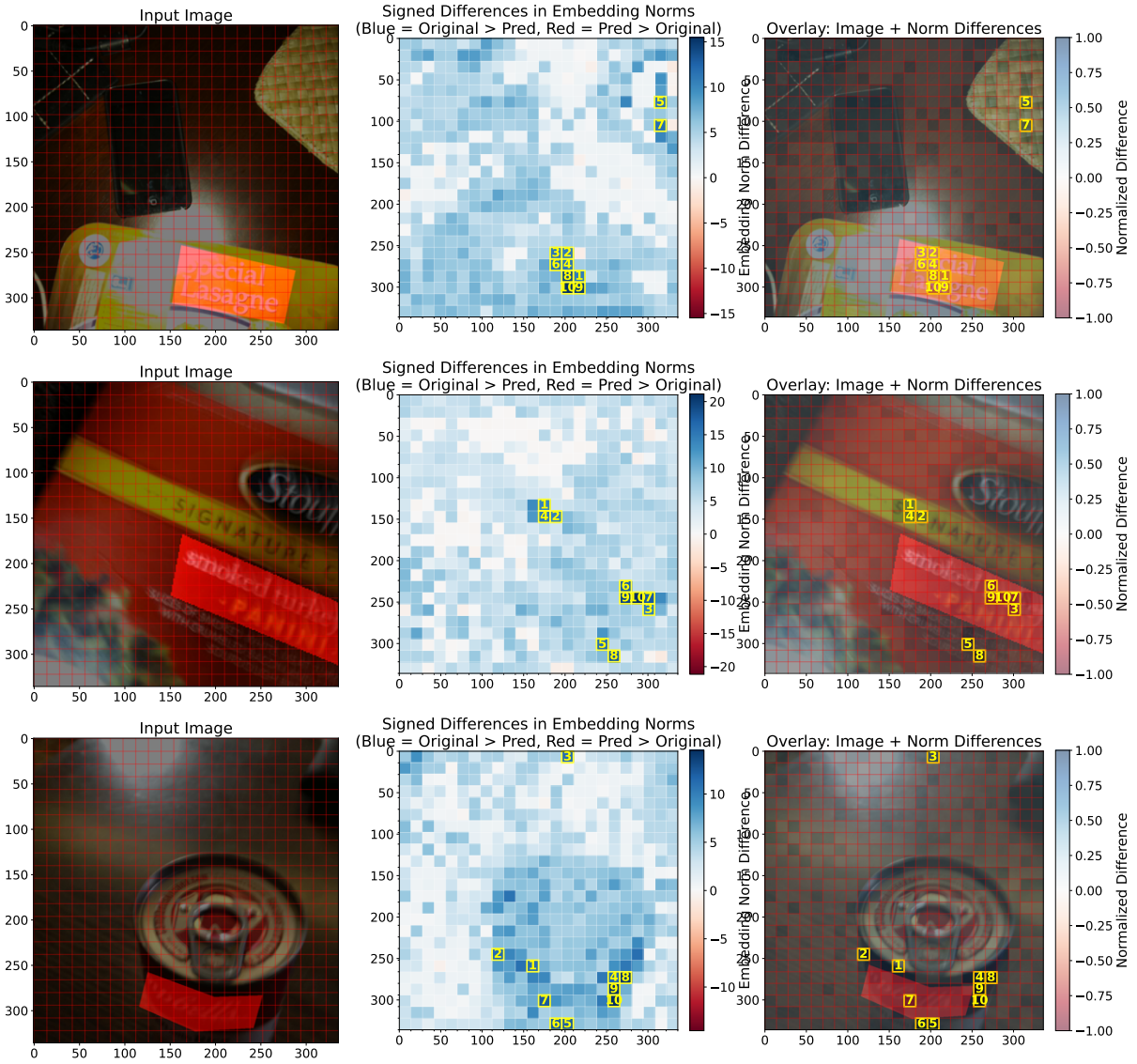


Figure 9: Visualization of high reconstruction loss patches that contributes to model's failure on answering questions that requires recognizing text in the objects (red regions are the answer grounding masks).

ther research in visualizing the nuanced difference between embeddings.

Figure 10: Idefics high kNN overlap ratio example, where we can observe the reordering among semantically similar vision embeddings.

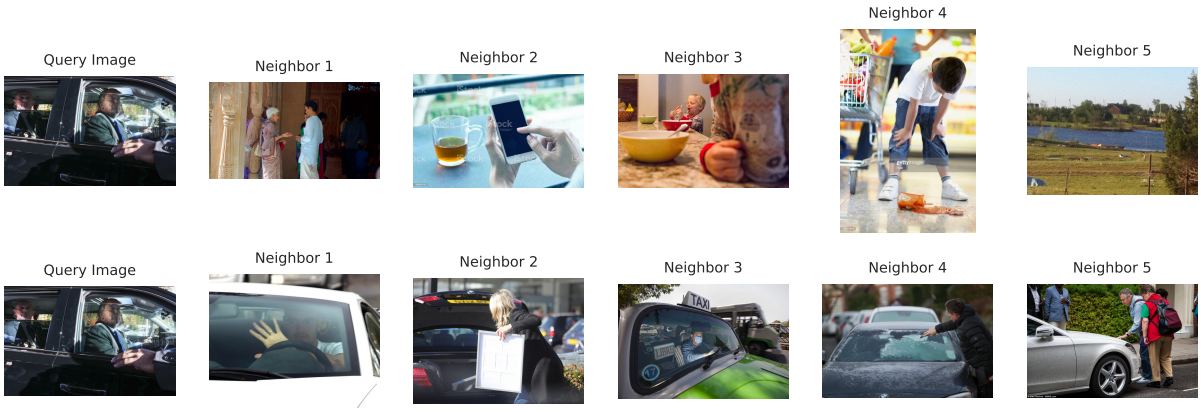


Figure 11: Qwen kNN example where the post-projection embeddings are better at retrieving semantically similar images (bottom row).

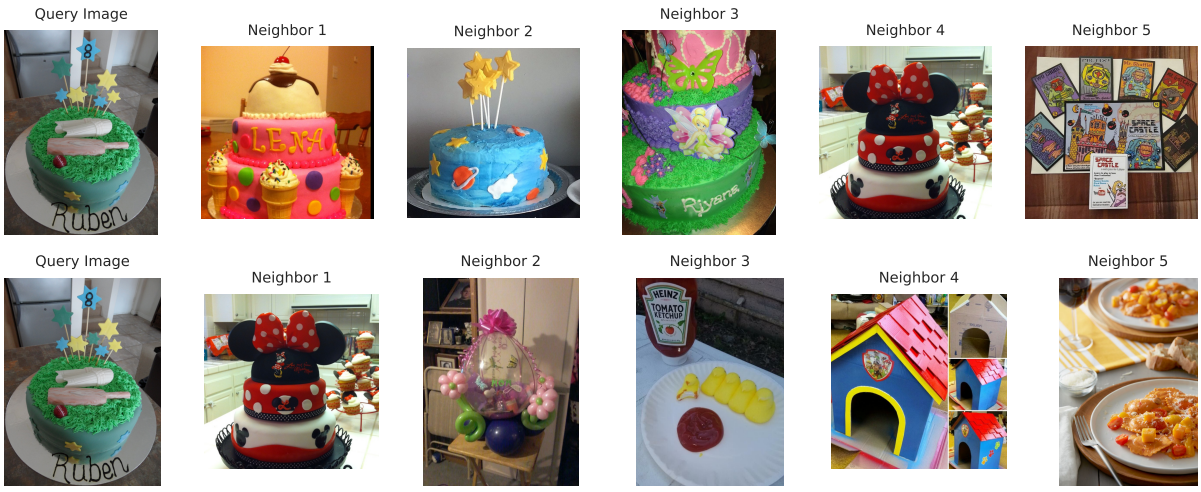


Figure 12: LLaVA low kNN overlap ratio example. We can observe the degradation in post-projection embedding.