How Do Inpainting Artifacts Propagate to Language?

Pratham Yashwante, Davit Abrahamyan*, Shresth Grover*, Sukruth Rao*

UC San Diego

Correspondence: pyashwante@ucsd.edu

Abstract

We study how visual artifacts introduced by diffusion-based inpainting affect language generation in vision-language models. We use a two-stage diagnostic setup in which masked image regions are reconstructed and then provided to captioning models, enabling controlled comparisons between captions generated from original and reconstructed inputs. Across multiple datasets, we analyze the relationship between reconstruction fidelity and downstream caption quality. We observe consistent associations between pixel-level and perceptual reconstruction metrics and both lexical and semantic captioning performance. Additional analysis of intermediate visual representations and attention patterns shows that inpainting artifacts lead to systematic, layer-dependent changes in model behavior. Together, these results provide a practical diagnostic framework for examining how visual reconstruction quality influences language generation in multimodal systems.

1 Introduction

Vision-language models (VLMs) are increasingly deployed within multi-stage pipelines, where visual inputs are processed or reconstructed before being consumed by language models. A common instance of this paradigm is image inpainting, in which missing or corrupted regions are filled prior to downstream generative tasks. Although modern diffusion-based inpainting models produce visually plausible reconstructions, they are optimized primarily for pixel-level realism, allowing subtle but semantically meaningful artifacts to be introduced without being perceptually salient. Figure 1 shows representative examples of this effect. Despite visually coherent reconstructions, localized inpainting artifacts lead to object substitutions, attribute changes, or category-level errors in downstream captions.

Because captioning models lack explicit awareness of which regions have been reconstructed, synthesized content may be treated as genuine visual evidence. This raises a central question: *to what extent does reconstruction fidelity influence downstream caption correctness and semantic grounding?* Despite the widespread use of inpainting and captioning, this relationship remains underexplored. We discuss related work in Appendix A and we also discuss the motivation for studying caption quality as a downstream proxy in Appendix B.

To study this interaction, we introduce a two-stage diagnostic framework. Images are synthetically degraded and reconstructed using diffusion-based inpainting, and both original and reconstructed images are passed to a frozen captioning model. By directly comparing the resulting captions, this enables controlled analysis of how reconstruction artifacts affect language generation without retraining either model.

Using this framework, we conduct a systematic analysis across diverse domains, examining relationships between reconstruction fidelity metrics and caption quality. We further analyze representation-level and attention-level changes within frozen vision encoders to understand how reconstruction artifacts manifest internally. Across datasets, we observe consistent associations between reconstruction fidelity and caption grounding, even when reconstruction and captioning models are trained independently.

Contributions. We (i) introduce a model-agnostic diagnostic framework for analyzing reconstruction-induced effects in vision-language pipelines, (ii) provide multi-dataset empirical evidence linking reconstruction fidelity to caption quality, and (iii) show that inpainting artifacts are associated with layer-wise and spatially localized changes in vision encoders.

*Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 727–745

Orig: a man in a blue shirt
Inp: a woman in a blue shirt

Orig: a man playing tennis in blue shorts
Inp: a man in a white shirt and black shorts

Orig: a brown cow in the middle
Inp: the horse in the middle

Orig: a blue and white plane
Inp: a blue truck with a tail

Figure 1: Qualitative examples illustrating captioning errors induced by center-region inpainting. Incorrect semantic attributes introduced by inpainting are highlighted in red, while the correct interpretation is shown in green.

2 Methodology

We adopt a degradation–reconstruction–captioning framework as shown in Figure 2. Given an input image, we apply a synthetic degradation to a predefined region using perturbations. The degraded image is reconstructed with a diffusion-based inpainting model, yielding a visually plausible but potentially semantically altered input. Both original and reconstructed images are then passed to a captioning model, and the resulting captions are compared using linguistic and semantic metrics.

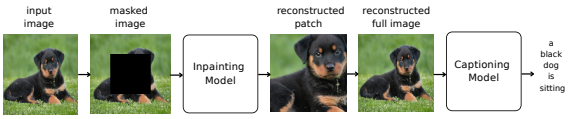

Figure 2: Degradation–reconstruction–captioning framework used to evaluate how inpainting artifacts propagate into downstream language outputs.

Our analysis proceeds along two complementary axes. First, we examine correlations between reconstruction fidelity metrics and caption quality metrics to assess whether improvements in visual fidelity correspond to stronger linguistic grounding. Second, we analyze representational stability within a frozen vision encoder by measuring embedding similarity and layer-wise attention drift between original and reconstructed inputs. By varying only the reconstruction process while keeping all models fixed, this setup isolates the effect of visual artifacts on downstream language behavior.

3 Experiments

Models. For the inpainting stage, we employ diffusion-based models Stable Diffusion (SD) 1.5, 2.0, and 3.0 (Rombach et al., 2022) for masked image completion. These models generate high-fidelity reconstructions while differing in scale and training data diversity which helps us to analyze

Figure 3: Masking examples illustrating center-region degradations on (A) Flickr and (B) RefCOCOg.

how architectural and data variations affect downstream robustness. We evaluate three VLMs used for caption generation: LLaVA (Li et al., 2024), BLIP (Li et al., 2022), and Qwen2.5-VL (Qwen Team, 2024). To study visual representations independently of decoding, we extract embeddings and attention maps from a ViT-Base (ViT-B) encoder (Dosovitskiy et al., 2021).

Evaluation Metrics. Reconstruction fidelity is measured using mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) (Wang et al., 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018). Caption quality is evaluated using BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and semantic similarity metrics including SimCSE (Gao et al., 2021) and SBERT (Reimers and Gurevych, 2019). We further assess representational stability using cosine similarity between visual embeddings and attention drift. See Appendix D for metric definitions and Appendix E for inference settings.

Datasets and Setup. Images are degraded using hard center masking, Gaussian-blurred masking, or low-dimensional compression. Dataset-specific masking details are provided in Appendix C.2. Figure 3 shows how the original image is masked across the three degradation variants. We eval-

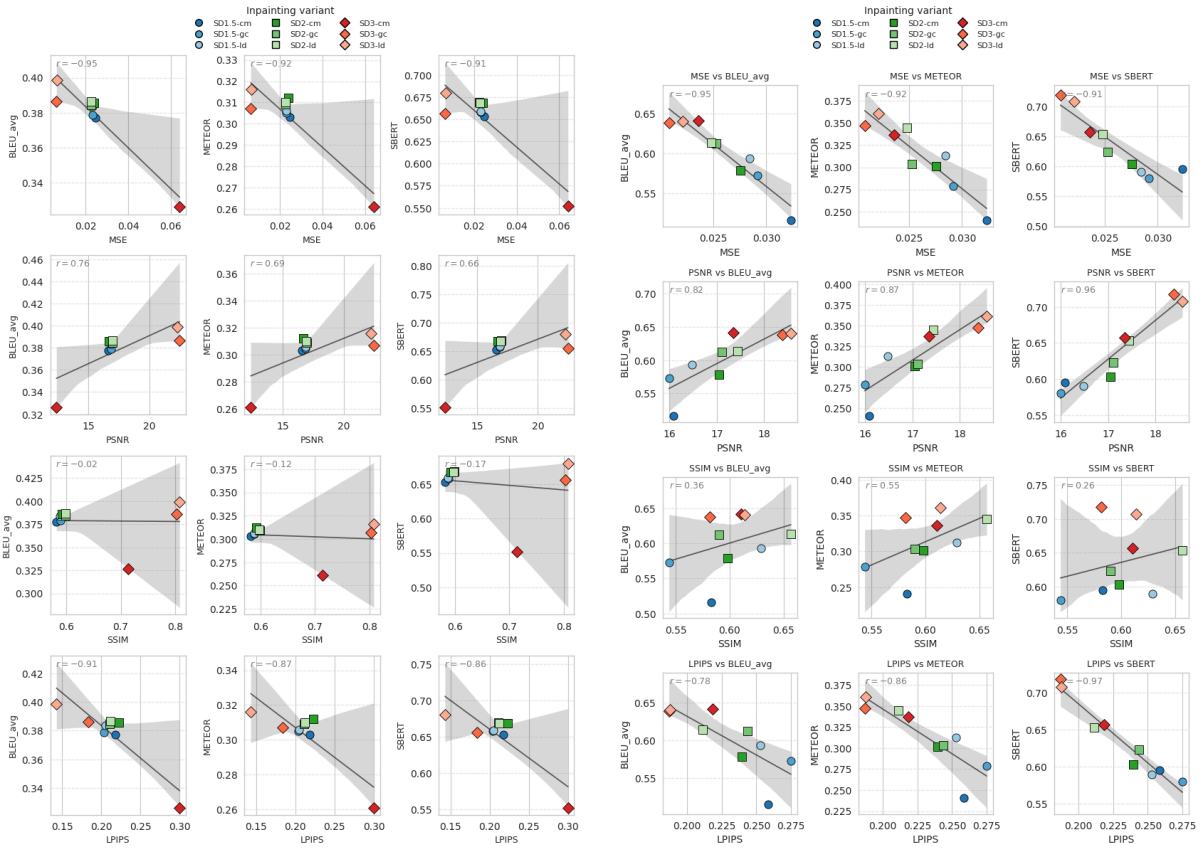

Figure 4: Correlations between reconstruction fidelity metrics and caption quality metrics on Flickr and RefCOCOg. Points show Stable Diffusion inpainting variants under three masking strategies: *cm* (hard center), *gc* (Gaussian), and *ld* (low-dimensional). Caption quality is evaluated using BLIP for Flickr and Qwen2.5-VL for RefCOCOg.

uate our framework across multiple datasets, including natural images (Flickr (Plummer et al., 2015), RefCOCOg (Yu et al., 2016)), medical imagery (ROCOv2 (Rückert et al., 2024), Indiana X-Ray (Demner-Fushman et al., 2016)), audio-spectrograms (GTZAN (Tzanetakis and Cook, 2002)), and structured time-series plots (TRUCE (Jhamtani and Berg-Kirkpatrick, 2021)). Representative examples are shown in Appendix C.

4 Results

Reconstruction fidelity correlates with caption quality. Figure 4 reports correlations between reconstruction fidelity metrics and caption quality metrics across SD variants and masking strategies. Across visually grounded datasets, lower reconstruction error, reflected by lower MSE and LPIPS and higher PSNR, is consistently associated with stronger linguistic alignment.

Among reconstruction metrics, perceptual distance (LPIPS) and pixel-level error (MSE) exhibit the strongest and most consistent correlations with caption quality, indicating that perceptual realism is critical for downstream grounding. In contrast, SSIM shows weak or inconsistent correlations for natural images such as Flickr, suggesting that global structural similarity alone is insufficient to predict caption correctness.

RefCOCOg and TRUCE follow the same directional trends, with reconstruction fidelity remaining predictive of caption quality. In RefCOCOg, correlations are particularly strong, reflecting the importance of preserving localized visual content for region-grounded captions. TRUCE exhibits slightly weaker but still consistent correlations, likely due to the numeric and trend-focused nature of captions, which reduces sensitivity to fine-grained visual detail (Appendix J.3).

Appendix J.4 shows analysis on ROCOv2, showing how reconstruction fidelity relates to caption quality in medical imagery, while Appendix K documents failure cases (GTZAN and X-ray) where limited linguistic variability prevents meaningful reconstruction–caption correlations. Across these

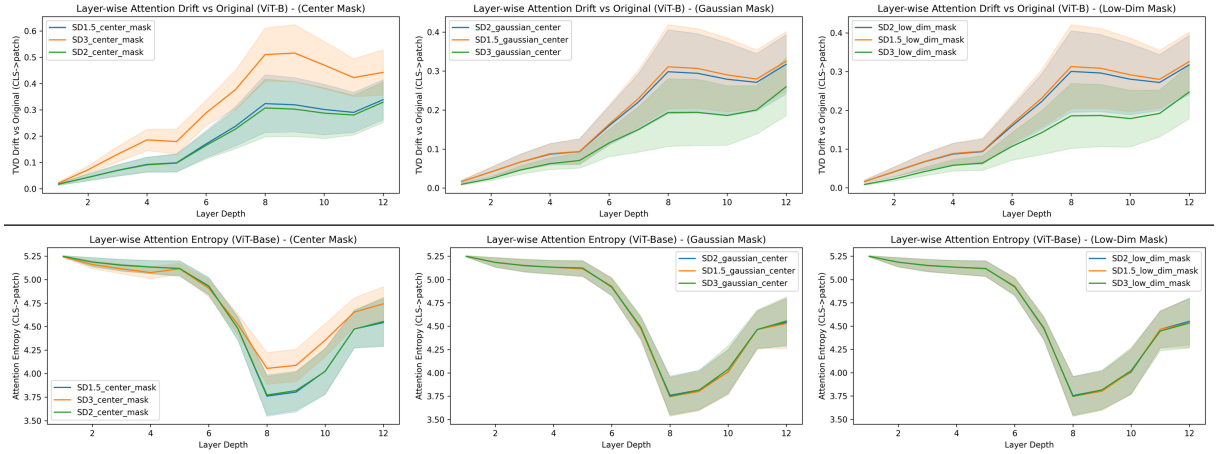

Figure 5: Layer-wise attention drift and entropy under inpainting on Flickr. Drift increases with depth and is higher for center-masked reconstructions.

analyses, ROCOv2 preserves strong, monotonic reconstruction–caption relationships despite higher absolute reconstruction error, whereas GTZAN and X-ray show near-flat caption metrics that remain insensitive to large reconstruction differences. Analyses on leave-one-out stability and guidance scale sensitivity are reported in Appendix J.1.

Inpainting artifacts induce structured semantic failures. Entity-level analysis on Flickr further reveals that inpainting artifacts induce failures spanning object identity, actions, and attributes. Using a spaCy-based (Honnibal et al., 2020) entity overlap pipeline on SD1.5 inpainted images, we obtain mean noun entity precision = 0.58, recall = 0.32, and F1 = 0.40, with lower overlap for verbs (F1 = 0.33) and adjectives (F1 = 0.16), confirming multi-level semantic impact beyond object identity. Annotation of the 100 lowest-F1 cases on Flickr reveals that attribute-level detail loss (98%) and participant reassignment (85%) are the dominant failure modes, while object substitution (29%) and activity substitution (22%) represent the most semantically disruptive errors. See Appendix I for full taxonomy and examples.

Masking strategy affects semantic stability. Beyond aggregate correlations, masking strategy plays a critical role in semantic stability. We see that smoother degradation schemes such as Gaussian-center masking and low-dimensional compression preserve caption quality more effectively than hard center masking. While center-masked reconstructions often appear visually plausible, they induce larger drops in both lexical and semantic metrics, indicating that abrupt spatial discontinuities disrupt object identity and relational cues relied upon by captioning models. A robustness check using Segment Anything Model (SAM) object masks (Kirillov et al., 2023) on Flickr confirms this trend generalizes beyond fixed geometric regions, with object-aligned masks achieving higher reconstruction fidelity and correspondingly stronger caption quality (see Appendix F). Both geometric and semantic masks place reconstruction pressure on captioning-relevant regions, explaining the consistent fidelity–caption relationship (Appendix G).

Inpainting induces layer-dependent attention drift. We analyze attention patterns in a frozen ViT-B encoder to assess how reconstruction artifacts affect internal representations. Figure 5 shows layer-wise Total Variation Distance (TVD) between CLS-to-patch attention maps for original and reconstructed inputs. Across all inpainting variants, attention drift increases with depth and peaks in later layers, while earlier layers remain comparatively stable. Center-masked reconstructions consistently induce the largest drift, indicating that sharper visual disruptions have a stronger impact on higher-level representations. Attention entropy exhibits a complementary pattern, with a dip in mid-layers followed by increased dispersion in deeper layers as semantic features are formed. Late-layer attention drift is negatively correlated with entity F1 ($r = -0.51, p < 0.001$) which links representational drift in deeper encoder layers to object-level caption failures.

Global visual representations shift under inpainting. We further examine how inpainting af-

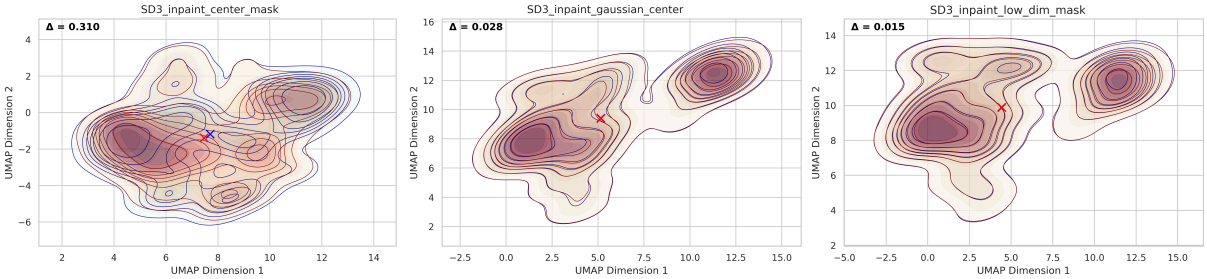

Figure 6: UMAP density visualization of ViT CLS embeddings for SD3 inpainting variants on Flickr. Smoother masking strategies preserve closer alignment with original representations.

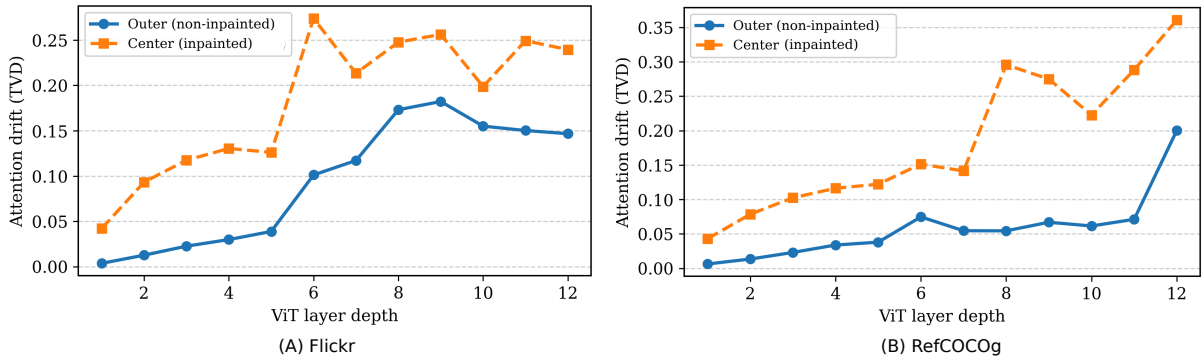

Figure 7: Spatial localization of attention drift under center inpainting. CLS → patch Drift is consistently higher in reconstructed regions and increases with depth.

fects global visual representations. Figure 6 visualizes UMAP projections of ViT CLS embeddings for original and reconstructed images under different SD3 masking strategies on Flickr. Center-masked reconstructions exhibit clear separation from the original embedding distribution, while Gaussian-center and low-dimensional masking yield progressively closer alignment. These trends are consistent with cosine similarity measurements across domains (Figure 9), where smoother strategies preserve higher representational similarity.

Attention drift is spatially localized. Finally, we assess whether attention drift is localized to reconstructed regions. Figure 7 compares CLS-to-patch drift for the inpainted center region versus the unmasked outer region. Across all layers, drift is higher within the reconstructed center, with divergence increasing in deeper layers, while outer regions remain comparatively stable. This confirms that semantic instability is more aligned with reconstructed content rather than global degradation.

5 Discussion

We investigated how artifacts introduced by diffusion-based inpainting influence downstream captioning in VLMs. Across multiple domains with visually grounded captions, improved reconstruction fidelity is consistently associated with more stable captions and visual representations. Pixel-level and perceptual reconstruction metrics are strong predictors of caption quality, whereas global structural similarity alone is often insufficient. We further show that inpainting artifacts primarily affect deeper layers of vision encoders and induce attention drift that is spatially aligned with reconstructed regions.

Overall, our findings indicate that inpainting can meaningfully alter semantic artifacts in multimodal pipelines and motivate reconstruction-aware diagnostics when evaluating vision-language robustness. In practice, systems that pass reconstructed or edited images into VLMs should account for reconstruction fidelity as a factor in downstream language reliability. More broadly, our diagnostic framework provides a reproducible foundation for future work on reconstruction-aware evaluation criteria, robustness benchmarks, and mitigation strategies for artifact-induced semantic failures in multimodal systems. Code can be accessed at `github.com/raosukruth/inpaint-caption`.

Limitations

Our study focuses on diffusion-based inpainting as a representative reconstruction mechanism which we chose for its state-of-the-art performance and broad applicability and whether the findings generalize to other inpainting paradigms or to other visual preprocessing operations such as super-resolution or denoising, remains unclear. While we evaluate multiple Stable Diffusion variants and masking strategies, we do not exhaustively explore reconstruction hyperparameters or alternative architectures, which may influence the degree of reconstruction-induced drift. Our analysis is centered on captioning and visually grounded language generation. The results may not fully extend to other vision-language tasks such as visual question answering or multi-step reasoning. Additionally, some datasets with limited linguistic variability, such as GTZAN and X-ray imagery, restrict the strength of measurable reconstruction-language correlations. Finally, all models are evaluated in a frozen setting to isolate reconstruction effects. This does not capture potential adaptation that may occur in end-to-end trained multimodal systems, which remains an open direction for future work.

Ethical Considerations

All datasets used in this work are publicly available benchmarks and do not contain personally identifiable information. The study does not involve human subjects, user interaction, or the collection of new personal data. All models are used in an inference-only setting and are not deployed in real-world decision-making contexts. Diffusion-based inpainting can introduce visually plausible but semantically incorrect content. By explicitly analyzing and documenting these failure modes, our work aims to support more transparent and reconstruction-aware evaluation of vision-language systems, particularly in settings where robustness is critical. We do not identify any direct societal or ethical risks arising from the experiments or datasets used in this paper.

LLM Usage Statement

VLMs were used as evaluation models, serving as frozen captioning baselines. LLMs were used as a writing assistance tool to improve clarity and presentation. They did not contribute to research ideation, experimental design, or analysis. All conclusions and responsibility for the content rest solely with the authors.

Acknowledgments

We thank the anonymous reviewers for their thoughtful feedback and constructive suggestions.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Seth* Forsgren and Hayk* Martiros. 2022. Riffusion - Stable diffusion for real-time music generation.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2021. Truth-conditional captioning of time series data. *arXiv preprint arXiv:2110.01839*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *Preprint*, arXiv:2304.02643.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Qwen Team. 2024. Qwen2.5-vl: A family of multimodal large language models. https://qwenlm.github.io/blog/qwen2.5-vl/.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. 2024. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1).

Zachary Shah, Neelesh Ramachandran, and Mason Wang. 2024. Riff-controlnet: Controlled audio inpainting using controlnet architecture. https://github.com/zachary-shah/riff-cnet.

George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Appendix

A Related Work

Early image inpainting methods framed reconstruction as a self-supervised learning problem, with Generative Adversarial Network (GAN) based approaches such as Context Encoders (Pathak et al., 2016) and U-Net–style architectures (Ronneberger et al., 2015) emphasizing spatial coherence and perceptual realism. Subsequent work introduced perceptual similarity metrics (Zhang et al., 2018) and large-scale generative models, including recent diffusion-based inpainting systems that produce visually plausible completions (Rombach et al., 2022). These models are typically evaluated using pixel-level or perceptual fidelity metrics, without explicit assessment of their semantic reliability when integrated into downstream text generation pipelines.

Image captioning has evolved around encoder–decoder architectures with attention mechanisms that condition language generation on visual representations (Xu et al., 2015; Anderson et al., 2018). Large-scale pretraining approaches such as BLIP (Li et al., 2022) further strengthened vision–language alignment, but captioning models remain sensitive to visual input quality and grounding cues.

A related line of work studies object hallucination in image captioning, where models generate entities not present in the image. Rohrbach et al. (Rohrbach et al., 2018) showed that strong performance on standard captioning metrics does not necessarily guarantee faithful visual grounding, as captioning models may rely on language priors even when visual evidence is ambiguous or weak. More recent studies extend this analysis to large VLMs, demonstrating that hallucination remains prevalent even in LLM-based systems (Liu et al., 2024). These works primarily focus on model architectures, decoding behavior, and instruction design, rather than on the effects of upstream visual transformations.

Separately, robustness benchmarks such as ImageNet-C and ImageNet-P (Hendrycks and Dietterich, 2019) evaluate classifier stability under common corruptions and perturbations. While influential for vision robustness, these benchmarks do not consider reconstruction-based transformations or their interaction with downstream language generation.

Our work complements these lines of research by analyzing how reconstruction artifacts introduced by diffusion-based inpainting propagate into caption grounding and internal visual representations in frozen vision–language pipelines.

B Motivation for Studying Downstream Caption Quality

Studying downstream caption quality matters for two key reasons.

First, captioning directly tests whether reconstructed images preserve the semantic content that VLMs rely on, beyond what pixel-level metrics alone can capture.

Second, caption degradation reveals real-world failure modes in multimodal pipelines: in many practical settings such as assistive technology, medical reporting, and robotics, reconstructed or edited images are passed into VLMs to generate language outputs, and incorrect artifacts can directly lead to problematic downstream results.

C Datasets

C.1 Dataset Descriptions

In this section, we summarize the datasets used and the subsets selected for our experiments.

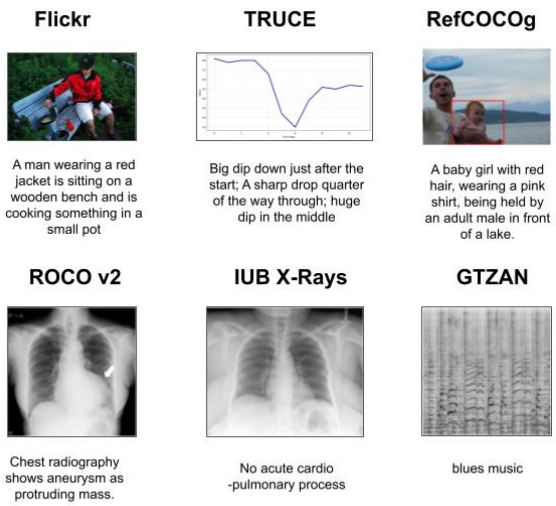

Figure 8: Representative examples from each dataset used in this study. These examples illustrate the diversity of visual modalities and caption styles across datasets.

Flickr Entities. Flickr Entities (Plummer et al., 2015) pairs natural images with region-level annotations and human-written captions. We select

a subset of 4,000 image-caption pairs for our experiments. Captions are descriptive, and visually grounded.

RefCOCOg. RefCOCOg (Yu et al., 2016) provides referring expressions aligned with explicit bounding box annotations. Each image is associated with one or more localized descriptions targeting specific objects or regions. We use a split of 2,000 samples and apply masking within the annotated bounding boxes, enabling evaluation of reconstruction effects under spatial grounding.

Medical Imaging (Indiana Chest X-ray and ROCOv2). We use two medical image captioning datasets to evaluate reconstruction effects in high-stakes domains. The Indiana University Chest X-ray dataset (Demner-Fushman et al., 2016) pairs chest radiographs with radiology reports; we extract the *Impression* section as the target caption. ROCOv2 (Rückert et al., 2024) extends this setting with medically curated image-caption pairs and structured clinical concepts. For both datasets, we use a randomly sampled subset of 1,000 images. The average caption length across medical datasets is approximately 149 characters.

GTZAN Spectrograms. GTZAN (Tzanetakis and Cook, 2002) is an audio dataset containing music recordings from 10 genre categories. We uniformly sample 500 audio files (50 per genre), resampled to 44.1 kHz. Each clip is truncated to the first 5.12 s and converted into a mel-spectrogram, which is treated as a 2D image. Genre labels serve as short textual descriptors.

TRUCE Time-Series Plots. TRUCE (Jhamtani and Berg-Kirkpatrick, 2021) consists of numeric time series paired with textual captions describing temporal properties such as trends, peaks, and anomalies. We render each series as a line plot prior to inpainting and captioning.

C.2 Masking and Degradation Strategies

To study how different forms of visual degradation affect downstream captioning and representation stability, we apply three controlled masking strategies to the input images prior to reconstruction. All masking operations are applied deterministically and only affect a localized region of the input, while the remainder of the image is left unchanged.

Masking Variants. We consider the following three degradation types:

- **Center Mask.** A rectangular region covering the target area is fully removed by setting all pixel values to zero. This produces a sharp spatial discontinuity and removes all visual information within the masked region.
- **Gaussian Blur.** The target region is degraded using a Gaussian blur with a fixed kernel size. Each pixel is replaced by a weighted average of its neighbors, resulting in a smooth attenuation of high-frequency details while preserving coarse structure.
- **Low-Dimensional Center Degradation.** Instead of fully removing the region, we apply an aggressive but structured degradation that preserves spatial layout while eliminating fine-grained semantic cues. Specifically, the masked region undergoes: (i) color quantization via k-means clustering with k=4 colors, (ii) spatial downsampling followed by upsampling to suppress high-frequency texture, and (iii) extremely low-quality JPEG compression. This produces a visually coherent patch that retains approximate shape and layout but loses texture, color fidelity, and detailed semantics.

Dataset-Specific Masking. For Flickr, medical image datasets and audio dataset, masking is applied to a fixed central region of the image. In contrast, RefCOCOg provides localized referring expressions paired with bounding boxes. For this dataset, we apply all masking operations directly within the annotated bounding box corresponding to the target object or region. As a result, the masked area may occur anywhere in the image and is semantically aligned with the referring expression. For TRUCE time-series plots, masking is applied along the temporal axis rather than a fixed spatial location. We select one of several informative contiguous segments of the plotted curve and degrade approximately 25% of the series length. This ensures that the degraded region corresponds to a meaningful portion of the temporal dynamics while preserving the overall structure of the plot. Gaussian blur and low-dimensional degradation use the same hyperparameters as in the natural image setting.

D Metrics

We evaluate the impact of inpainting artifacts using complementary metrics that quantify (i) reconstruction fidelity, (ii) caption quality, and (iii) represen-

tation and attention stability. All metrics compare outputs from reconstructed inputs against their original counterparts.

Reconstruction Fidelity. Reconstruction quality is evaluated within the degraded regions using a combination of pixel-level, signal-level, and perceptual metrics. We report MSE to quantify pixel-wise reconstruction error and PSNR to measure signal fidelity relative to reconstruction noise. Structural consistency between the original and reconstructed regions is assessed using the SSIM (Wang et al., 2004). To capture perceptual differences beyond low-level statistics, we additionally report LPIPS (Zhang et al., 2018), which measures distance in deep feature embedding space.

Caption Quality. Generated captions are evaluated against ground truths using standard lexical and semantic metrics. Lexical overlap is measured using BLEU-1 through BLEU-4 (Papineni et al., 2002), which capture n-gram precision at increasing orders, and ROUGE-L (Lin, 2004), which evaluates longest common subsequence recall. We also report METEOR (Banerjee and Lavie, 2005), an alignment-based metric that incorporates synonymy and stemming to better reflect semantic similarity. To assess semantic alignment beyond surface overlap, we additionally compute cosine similarity between sentence embeddings using SBERT (Reimers and Gurevych, 2019) and both supervised and unsupervised variants of SimCSE (Gao et al., 2021). For each generated caption, we report the maximum similarity to any reference caption for the corresponding input.

Representation Similarity. To quantify global visual drift, we extract CLS embeddings from a frozen ViT encoder and compute cosine similarity between embeddings obtained from original and reconstructed images.

Attention Drift and Entropy. To analyze how inpainting artifacts propagate through model internals, we quantify changes in attention behavior using divergence-based metrics computed from CLS-to-patch attention maps. We measure total variation distance to capture layer-wise divergence between attention distributions obtained from original and reconstructed inputs. In addition, we compute attention entropy to characterize the dispersion of CLS-to-patch attention within each layer, providing a measure of how concentrated or diffuse the model's visual focus becomes following reconstruction. TVD quantifies divergence between attention distributions as the sum of absolute differences across patch positions. For spatial analyses, TVD is computed separately over inpainted and non-inpainted regions using binary mask supervision.

E Settings and Configurations

All experiments are conducted under a unified inference-only evaluation protocol. Across all datasets, models are used in a frozen state with no fine-tuning or adaptation, ensuring that observed effects arise solely from input degradation and reconstruction rather than model updates.

Captioning Settings. Caption generation follows a shared decoding configuration across models, using beam search with six beams, top-p sampling with $p = 0.9$, temperature 0.8, and a maximum of 48-64 generated tokens depending on the task. For each image, three candidate captions are generated. Prompt conditioning is used where supported: RefCOCOg captions are explicitly constrained to describe the contents of the red bounding box, while BLIP operates in an unconditional captioning mode due to limited prompt adherence.

Diffusion-Based Inpainting. All models are applied using fixed hyperparameters across datasets. Inpainting is conducted with 50 inference steps in all cases. Guidance scales are fixed within the empirically stable value of 7.5 identified in preliminary analysis, and are held constant per diffusion model across all experiments. The strength parameter is set to 1.0 for SD1.5 and SD2. For SD3, strength is reduced to 0.6, as higher values were observed to either overwrite large portions of the original image or collapse to degenerate reconstructions in which the masked region remains blank.

Prompt Construction for Inpainting. Inpainting prompts are dataset-specific but strictly derived from existing annotations. For Flickr, captions associated with the original images are used directly as prompts. For RefCOCOg, multiple bounding-box captions are concatenated into a single prompt describing the target region, truncated to a maximum of 75 tokens to avoid prompt overflow. For TRUCE plots, prompts consist of a short reconstruction instruction followed by the raw numeric time-series and its description.

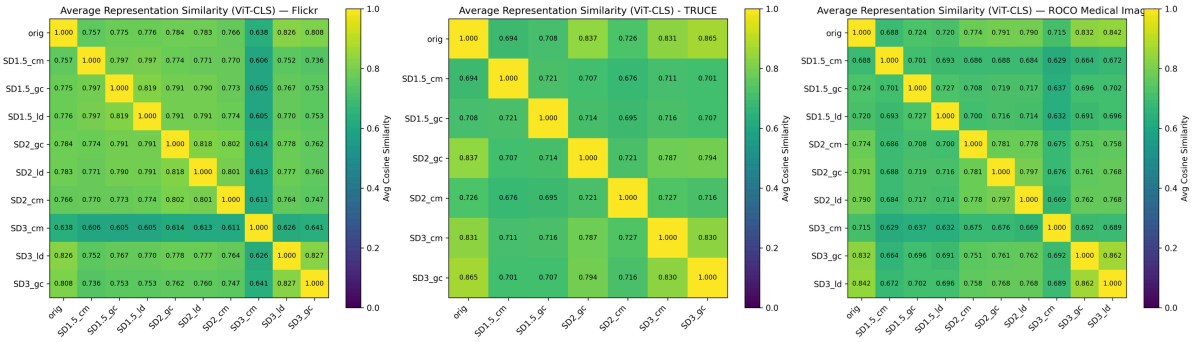

Figure 9: Cosine similarity between original and reconstructed visual representations across domains.

Table 1: SAM mask robustness check on Flickr (SD3, 50 images). Reconstruction fidelity and caption quality are reported for our three standard masking strategies and SAM object masks. Higher reconstruction fidelity shows stronger caption quality across both captioning models.

Mask Type	Reconstruction			BLIP			Qwen2.5-VL		
	PSNR ↑	LPIPS ↓	MSE ↓	BLEU ↑	METEOR ↑	ROUGE-L ↑	BLEU ↑	METEOR ↑	ROUGE-L ↑
Center Mask	14.974	0.256	0.034	0.389	0.303	0.441	0.337	0.318	0.419
Gaussian	15.253	0.247	0.032	0.397	0.307	0.456	0.351	0.328	0.438
Low-Dim	15.342	0.246	0.032	0.361	0.290	0.437	0.341	0.327	0.427
SAM	**19.989**	**0.214**	**0.013**	**0.398**	**0.311**	**0.479**	**0.403**	**0.347**	**0.477**

F SAM Mask Robustness Check

We conduct a robustness check comparing SD3 inpainting under our three standard masking strategies against SAM (Segment Anything Model) object masks (Kirillov et al., 2023) on 50 Flickr images to assess whether our findings generalize beyond fixed geometric masks. SAM masks are object-aligned and variable in shape, providing semantically targeted regions rather than fixed rectangular areas.

As shown in Table 1, SAM masks achieve higher reconstruction fidelity across all pixel-level metrics and correspondingly show improved caption quality under both BLIP and Qwen2.5-VL. We hypothesize this is because object-shaped masks provide more precise boundary information, making the inpainting task easier for the diffusion model. Crucially, the ordering of mask types by reconstruction fidelity continues to predict caption quality, consistent with our main findings. This shows that the reconstruction–caption relationship generalizes beyond fixed geometric masking.

G Shared Characteristics of Semantic and Geometric Masking

Both semantic object masks (e.g., RefCOCOg bounding boxes, SAM) and our geometric masks share a critical characteristic: they require reconstructing regions that often contain captioning-relevant visual semantics such as objects, attributes, and relations. This shared pressure on semantically critical regions explains why the same reconstruction-caption relationship is observed across both mask types.

One perspective to understand this is through the attention drift analysis (Figure 7): under center inpainting, attention drift is consistently higher within the reconstructed region than outside it, and increases with depth, suggesting that reconstruction artifacts preferentially perturb the visual processing of the edited region and amplify in later encoder layers. When reconstruction fidelity is low, captioning models produce more generic and incorrect outputs, consistent with reduced access to fine-grained visual evidence in the edited region. For SAM masks, the higher reconstruction fidelity yields reduced drift compared to geometric masks, further supporting this mechanism.

H Native Encoder Representations

To assess whether our representational findings generalize beyond the standalone ViT-B encoder, we conduct an additional analysis using the native vision encoders from LLaVA and Qwen2.5-VL on the Flickr test dataset. Table 3 reports embedding

Table 2: Caption error taxonomy on the 100 lowest-F1 Flickr cases (SD1.5). Each case may belong to multiple error categories.

Error Category	What the error represents	Frequency
Attribute-level detail loss	Missing fine-grained visual attributes such as color, clothing, background, and appearance details; scene structure often remains intact	98
Participant reassignment / specificity shift	Incorrect identity of entities (gender, age, number, or collapse into generic terms)	85
Object substitution	Core object is replaced with a semantically different category, altering the scene meaning	29
Activity substitution	Action verb is incorrectly predicted with a mutually incompatible activity, changing event semantics	22
Generic / degenerate output	Nonsense, repetition loops, or non-informative captions with no grounded semantics	9

drift across mask types for both models. We observe the same consistent trend as in our ViT-B study: harder center masking induces the largest embedding drift, while smoother masking strategies yield progressively smaller shifts. This confirms that our representation-level conclusions hold for the vision encoders actually used in the deployed VLMs, and are not specific to the standalone ViT-B. Layer-wise analysis using BLIP's native encoder further reveals the same depth-dependent attention drift observed in our standalone ViT-B experiments: earlier layers remain comparatively stable, while mid-to-late layers exhibit progressively larger drift as semantic features are formed.

Table 3: Native encoder embedding drift across mask types on Flickr. Drift is measured between original and reconstructed image representations using the native vision encoders of LLaVA and Qwen2.5-VL. Larger values indicate greater representational shift. Center masking consistently induces the highest drift, while smoother strategies preserve closer alignment.

Mask Type	LLaVA Shift	Qwen Shift
Center Mask	0.370	0.391
Gaussian Center	0.147	0.210
Low-Dim Mask	**0.126**	**0.215**

I Entity-Level Semantic Analysis

We conduct an entity overlap analysis on Flickr using SD1.5 inpainted images (center masking) to evaluate caption degradation at a finer granularity than global lexical metrics, Noun-based entities are extracted from generated and ground-truth captions using a spaCy-based pipeline, with precision, recall, and F1 computed per image. Across test images, we obtain mean entity precision = 0.58, recall = 0.32, and F1 = 0.40, showing that while coarse semantics are often preserved, fine-grained object details are frequently missed or distorted. Extending the analysis to action and attribute spaces shows lower overlap than noun-level entities (Verb F1 = 0.33; Adjective F1 = 0.16), confirming multi-level semantic impact beyond object identity.

Low F1 cases reveal clear hallucinations, semantic drift, and generation collapse at the object level:

- An image of two boys bouncing on a wet trampoline is described as brothers playing in an inflatable pool.
- A man reaching into a cigarette pack becomes a hallucinated scene involving a cell phone and a cup.
- A roller-skater on a railing is rewritten as a skateboarder on a ramp.
- A bicyclist at a race degenerates into repetitive nonsensical text (*"bale bale... "*).

To systematically map reconstruction artifacts to specific caption error types, we annotate the 100 lowest-F1 cases using a structured error taxonomy. We extend our spaCy-based noun-entity pipeline to capture actions and descriptive modifiers, and use embedding-based similarity thresholds to distinguish lexical variation from semantic text shifts. Table 2 reports the distribution of error types across annotated cases.

Attribute-level detail loss is near-universal (98%), reflecting that inpainting artifacts consistently erode fine-grained visual descriptors even when overall scene structure is preserved. Participant reassignment is also highly prevalent (85%),

with models frequently collapsing specific identities into generic terms or mis-assigning gender and age. Object substitution (29%) and activity substitution (22%) occur in a minority of cases but represent the most semantically disruptive failures. Generic or degenerate outputs (9%) indicate complete failure of language grounding. These results show that reconstruction artifacts induce structured, multi-level semantic failures rather than uniform caption degradation.

J Success Cases

J.1 Flickr dataset (Success Case)

Analysis Flickr exhibits a clear and stable reconstruction–caption relationship. MSE and LPIPS maintain strong negative correlations with all caption metrics across inpainting variants, while PSNR remains positively correlated. SSIM shows weak and variable behavior, reinforcing that structural similarity alone is not a reliable predictor of caption grounding. Gaussian-blurred and low-dimensional masking under SD2 and SD3 yield the most stable trade-offs between reconstruction fidelity and caption quality, whereas hard center masking introduces disproportionate degradation due to removal of salient semantic content (see Table 5 and 6).

Table 4: Leave-one-out correlation stability on Flickr. For each reconstruction metric, we report the range of full-data Pearson correlations with caption metrics, the mean and standard deviation of LOO correlations, and the number of sign reversals across all splits.

Metric	r_{full} range	μ_{LOO}	σ_{LOO}	Sign flips
MSE	$[-0.951, -0.906]$	≈ -0.88	≤ 0.19	0
LPIPS	$[-0.914, -0.857]$	≈ -0.85	≤ 0.11	0
PSNR	$[0.662, 0.761]$	≈ 0.69	≤ 0.13	0
SSIM	$[-0.166, -0.022]$	≈ -0.07	≥ 0.19	≥ 1

Correlation Robustness Results on Flickr We assess the robustness of reconstruction-caption correlations on Flickr using a leave-one-out (LOO) analysis, recomputing Pearson correlations after removing each inpainting configuration in turn. Table 4 summarizes the stability of each reconstruction metric across all caption quality measures.

MSE shows consistently strong negative correlations with all caption metrics, with full-data correlations in the range $[-0.951, -0.906]$ and no sign reversals under leave-one-out (LOO) analysis. LPIPS follows a similar pattern, exhibiting strong negative correlations, low LOO variance, and stable signs across splits. PSNR displays moderate positive correlations with caption quality; while its magnitude varies more across LOO splits, the correlation direction remains consistent. In contrast, SSIM exhibits weak and unstable behavior, with near-zero full correlations and frequent sign reversals under LOO analysis, indicating limited reliability for predicting semantic caption quality.

Table 5: Pixel-level reconstruction metrics on Flickr. Metrics are averaged over 4k samples. Lower MSE/LPIPS and higher PSNR/SSIM indicate better reconstruction fidelity.

Method	MSE	PSNR	SSIM	LPIPS
SD1.5-cm	0.0246	16.59	0.582	0.218
SD1.5-gc	0.0231	**16.88**	**0.589**	**0.204**
SD1.5-ld	**0.0231**	16.86	0.588	0.205
SD2-cm	0.0240	16.68	0.592	0.223
SD2-gc	**0.0225**	**16.98**	**0.599**	0.211
SD2-ld	0.0226	16.94	0.598	**0.212**
SD3-cm	0.0641	12.37	0.714	0.300
SD3-gc	**0.00629**	**22.47**	0.803	0.184
SD3-ld	0.00666	22.26	**0.808**	**0.143**

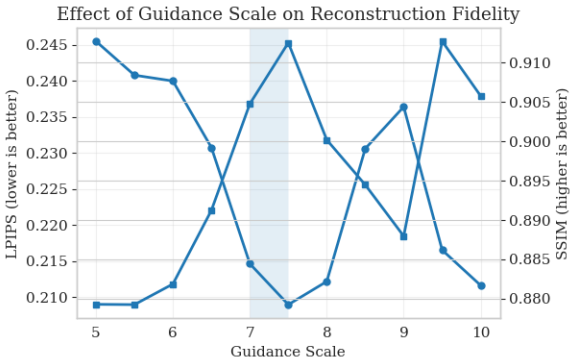

Figure 10: Effect of classifier-free guidance scale on reconstruction fidelity. LPIPS (left axis, lower is better) and SSIM (right axis, higher is better) are shown as a function of guidance over 100 samples.

Guidance Scale Analysis We analyze the effect of classifier-free guidance on reconstruction fidelity using LPIPS and SSIM. As shown in Figure 10, increasing guidance from low values improves perceptual and structural quality, with LPIPS decreasing and SSIM increasing up to a guidance range of approximately 7.0-7.5. Beyond this range, improvements saturate and become unstable, with no consistent gains at higher guidance scales. This indicates that moderate guidance achieves a stable trade-off between perceptual similarity and structural fidelity.

Table 6: Captioning performance (absolute and relative) across inpainting variants and original datasets for Flickr. Lexical metrics: BLEU-1–4, METEOR (MET.), ROUGE-L (R-L). Semantic metrics: supervised SimCSE (sup.), unsupervised SimCSE (unsup.), SBERT cosine similarity. Abbreviations — cm: center mask; gc: Gaussian center; ld: low-dimensional mask. $\%\Delta = (\text{Inpainted} - \text{Original})/\text{Original} \times 100$. Negative = performance drop.

Dataset	B1	%Δ	B2	%Δ	B3	%Δ	B4	%Δ	MET.	%Δ	R-L	%Δ	sup.	%Δ	unsup.	%Δ	SBERT	%Δ
BLIP																		
SD1.5-cm	0.602	-3.68	0.423	-5.79	0.289	-7.67	0.195	-10.14	0.303	-3.81	0.453	-3.63	0.753	-2.09	0.716	-2.16	0.653	-2.83
SD1.5-gc	0.604	-3.36	0.425	-5.35	0.291	-7.01	0.195	-10.14	0.305	-3.17	0.454	-3.40	0.756	-1.69	0.720	-1.64	0.658	-2.08
SD1.5-ld	0.608	-2.72	0.430	-4.23	0.296	-5.43	0.200	-7.83	0.306	-2.86	0.457	-2.81	0.756	-1.69	0.722	-1.37	0.659	-1.93
SD2-cm	0.610	-2.40	0.434	-3.34	0.298	-4.47	0.201	-7.37	0.312	-0.95	0.461	-1.87	0.765	-0.52	0.726	-0.82	0.668	-0.59
SD2-gc	0.610	-2.40	0.431	-3.99	0.296	-5.43	0.201	-7.37	0.309	-1.90	0.460	-2.13	0.764	-0.65	0.727	-0.69	0.669	-0.45
SD2-ld	0.614	-1.76	0.434	-3.34	0.297	-5.11	0.201	-7.37	0.310	-1.59	0.462	-1.70	0.763	-0.78	0.725	-0.96	0.668	-0.59
SD3-cm	0.546	-12.64	0.365	-18.67	0.239	-23.71	0.155	-28.57	0.261	-17.14	0.404	-14.04	0.670	-12.85	0.639	-12.72	0.552	-17.86
SD3-gc	0.608	-2.72	0.433	-3.56	0.299	-4.47	0.205	-5.53	0.307	-2.54	0.456	-2.98	0.757	-1.56	0.719	-1.78	0.656	-2.38
SD3-ld	0.623	-0.32	0.447	-0.45	0.311	-0.64	0.214	-1.38	0.316	+0.32	0.468	-0.43	0.774	+0.65	0.736	+0.55	0.680	+1.19
Orig.	0.625	—	0.449	—	0.313	—	0.217	—	0.315	—	0.470	—	0.769	—	0.732	—	0.672	—
QWEN																		
SD1.5-cm	0.555	-6.25	0.386	-9.59	0.263	-12.63	0.178	-14.83	0.322	-6.67	0.452	-5.98	0.781	-2.98	0.726	-2.94	0.674	-4.12
SD1.5-gc	0.562	-5.06	0.391	-8.43	0.268	-10.96	0.181	-13.40	0.326	-5.51	0.457	-4.81	0.785	-2.48	0.731	-2.27	0.678	-3.55
SD1.5-ld	0.564	-4.73	0.394	-7.72	0.270	-10.30	0.183	-12.44	0.328	-4.93	0.459	-4.56	0.786	-2.36	0.732	-2.14	0.680	-3.28
SD2-cm	0.565	-4.56	0.396	-7.26	0.270	-10.30	0.182	-12.92	0.329	-4.64	0.458	-4.77	0.788	-2.12	0.732	-2.14	0.683	-2.84
SD2-gc	0.575	-2.87	0.405	-5.15	0.281	-6.64	0.193	-7.66	0.332	-3.77	0.465	-3.32	0.791	-1.74	0.736	-1.60	0.688	-2.13
SD2-ld	0.569	-3.88	0.400	-6.32	0.275	-8.64	0.188	-10.05	0.331	-4.06	0.463	-3.74	0.791	-1.74	0.733	-2.01	0.687	-2.27
SD3-cm	0.431	-27.18	0.271	-36.54	0.169	-43.82	0.106	-49.28	0.250	-27.54	0.344	-28.48	0.667	-17.13	0.579	-22.57	0.530	-24.61
SD3-gc	0.561	-5.23	0.396	-7.26	0.274	-9.03	0.189	-9.57	0.334	-3.19	0.457	-5.00	0.789	-1.99	0.729	-2.54	0.682	-2.99
SD3-ld	0.579	-2.20	0.413	-3.28	0.290	-3.65	0.200	-4.31	0.346	+0.29	0.469	-2.50	0.803	-0.25	0.746	-0.27	0.703	0.00
Orig.	0.592	—	0.427	—	0.301	—	0.209	—	0.345	—	0.481	—	0.805	—	0.748	—	0.703	—
LLAVA																		
SD1.5-cm	0.729	-6.65	0.560	-9.53	0.412	-12.41	0.296	-15.69	0.302	-10.12	0.554	-6.27	0.796	-3.63	0.759	-3.81	0.711	-4.56
SD1.5-gc	0.737	-5.63	0.570	-7.91	0.425	-9.66	0.310	-11.79	0.306	-8.93	0.560	-5.34	0.800	-3.15	0.762	-3.42	0.715	-4.03
SD1.5-ld	0.737	-5.63	0.570	-7.91	0.426	-9.55	0.312	-11.12	0.305	-9.23	0.560	-5.34	0.801	-3.03	0.764	-3.17	0.715	-4.03
SD2-cm	0.738	-5.50	0.572	-7.60	0.426	-9.55	0.310	-11.69	0.309	-8.04	0.565	-4.39	0.806	-2.42	0.765	-3.04	0.723	-2.95
SD2-gc	0.746	-4.48	0.580	-6.30	0.435	-7.64	0.319	-9.10	0.313	-6.85	0.568	-3.90	0.809	-2.06	0.772	-2.16	0.725	-2.68
SD2-ld	0.742	-4.99	0.576	-6.94	0.432	-8.28	0.316	-9.97	0.311	-7.44	0.569	-3.73	0.807	-2.30	0.767	-2.79	0.722	-3.09
SD3-cm	0.563	-27.91	0.376	-39.26	0.248	-47.36	0.165	-52.99	0.232	-30.95	0.418	-29.26	0.645	-21.91	0.582	-26.21	0.525	-29.53
SD3-gc	0.715	-8.46	0.553	-10.67	0.409	-13.17	0.297	-15.37	0.305	-9.23	0.552	-6.60	0.791	-4.24	0.746	-5.44	0.695	-6.71
SD3-ld	0.755	-3.33	0.590	-4.68	0.444	-5.74	0.328	-6.55	0.318	-5.36	0.576	-2.54	0.819	-0.85	0.781	-1.01	0.737	-1.07
Orig.	0.781	—	0.619	—	0.471	—	0.351	—	0.336	—	0.591	—	0.826	—	0.789	—	0.745	—

J.2 RefCOCOg dataset (Success Case)

Analysis RefCOCOg exhibits very strong and consistent reconstruction–caption correlations, comparable to or stronger than those observed on Flickr. MSE and LPIPS show near-linear negative correlations with caption quality, while PSNR exhibits correspondingly strong positive correlations across BLEU_{avg}, METEOR, and SBERT. These trends indicate that reconstruction fidelity within the annotated region is critical for successful referring expression generation. In contrast, SSIM remains weakly correlated and unstable, suggesting limited sensitivity to semantically relevant degradation. Across diffusion variants, SD2 and SD3 with Gaussian or low-dimensional degradation perform best, while center-masked reconstructions consistently underperform due to complete removal of the referential target (see Table 7 and 8).

Table 7: Pixel-level reconstruction metrics across inpainting variants for RefCOCOg.Metrics are averaged over the evaluation set.

Variant	MSE	PSNR	SSIM	LPIPS
SD1.5-cm	0.03232	16.09	0.583	0.259
SD1.5-gc	0.02918	15.99	0.544	0.275
SD1.5-ld	0.02843	16.48	0.629	0.253
SD2-cm	0.02757	17.05	0.598	0.240
SD2-gc	0.02529	17.11	0.590	0.244
SD2-ld	0.02480	17.45	0.657	0.211
SD3-cm	0.02360	17.35	0.611	0.218
SD3-gc	**0.02082**	18.38	0.582	0.187
SD3-ld	0.02212	**18.56**	**0.614**	**0.188**

Table 8: Captioning performance across inpainting variants for RefCOCOg (QWEN). Reported metrics include BLEU average (BLEU-1–4), METEOR, and SBERT cosine similarity. Higher values indicate better caption quality and semantic alignment.

Variant	BLEU_{avg}	METEOR	SBERT
SD1.5-cm	0.516	0.240	0.595
SD1.5-gc	0.572	0.279	0.580
SD1.5-ld	0.594	0.313	0.590
SD2-cm	0.579	0.301	0.603
SD2-gc	0.613	0.304	0.623
SD2-ld	0.614	0.345	0.653
SD3-cm	0.642	0.337	0.657
SD3-gc	0.638	0.347	**0.718**
SD3-ld	0.640	**0.361**	0.707

J.3 TRUCE dataset (Success Case)

Analysis. TRUCE exhibits clear and consistent reconstruction–caption correlations, though lower strength than the two natural-image datasets (Figure 11. MSE and LPIPS maintain negative correlations with caption quality, while PSNR shows positive correlations, with the strongest relationships observed for SBERT similarity (up to $|r| \approx 0.8$). Lexical metrics such as BLEU_{avg} and METEOR show more moderate correlations. Across both reconstruction and captioning tasks, SD2-gc and SD3-gc perform best, achieving the strongest captioning scores (Table 10) and the lowest reconstruction error (Table 9) respectively. In contrast, center-mask variants consistently underperform. SSIM shows more stable but moderate correlations compared to observations on natural-image datasets. TRUCE confirms that the reconstruction–caption relationship generalizes beyond natural images, albeit with slightly reduced correlation strength.

Table 10: Captioning performance on the TRUCE dataset (QWEN). Reported metrics include BLEU-1–4, METEOR (MET.), ROUGE-L (R-L), and SBERT cosine similarity. Abbreviations — cm: center mask; gc: Gaussian center.

Variant	B1	B2	B3	B4	MET.	R-L	SBERT
SD1.5-cm	0.275	0.105	0.042	0.0260	0.195	0.190	0.47
SD1.5-gc	0.295	0.115	0.046	0.0280	0.210	0.205	0.48
SD2-cm	0.285	0.110	0.044	0.0270	0.205	0.200	0.49
SD2-gc	**0.315**	**0.125**	**0.049**	**0.0300**	**0.225**	**0.220**	**0.52**
SD3-cm	0.305	0.120	0.047	0.0290	0.220	0.215	0.51
SD3-gc	0.290	0.108	0.043	0.0265	0.208	0.202	0.50

Table 9: Pixel-level reconstruction metrics across inpainting variants for TRUCE. MSE ↓, PSNR ↑, SSIM ↑, and LPIPS ↓ are averaged over all samples. Lower MSE/LPIPS and higher PSNR/SSIM indicate better reconstruction fidelity.

Variant	MSE	PSNR	SSIM	LPIPS
SD1.5-cm	0.0544	15.24	0.871	0.177
SD1.5-gc	0.0439	15.97	0.879	0.171
SD2-cm	0.0360	17.20	0.892	0.161
SD2-gc	0.0235	19.67	0.925	0.128
SD3-cm	0.0158	20.10	0.945	0.096
SD3-gc	**0.0122**	**21.87**	**0.951**	**0.086**

J.4 ROCOv2 (Qualified Success Case)

Analysis. ROCOv2 exhibits a clear and internally consistent relationship between reconstruction fidelity and caption quality (Figure 12). Across inpainting variants, MSE and LPIPS show strong negative correlations with all captioning metrics (up to $|r| = 0.95$ for SBERT), while PSNR shows correspondingly strong positive correlations, indicating that relative improvements in reconstruction fidelity reliably translate to improved language alignment. SSIM also demonstrates strong positive correlations across all caption metrics, suggesting that structural preservation is more informative for semantic grounding in medical imagery than in natural-image datasets.

Table 11: Pixel-level reconstruction metrics across inpainting variants for ROCO V2. Metrics are averaged over the evaluation set.

Variant	MSE	PSNR	SSIM	LPIPS
SD1.5-cm	2217.34	16.74	0.677	0.275
SD1.5-gc	**2012.18**	**17.41**	**0.690**	**0.251**
SD1.5-ld	2132.21	17.25	0.685	0.256
SD2-cm	1417.41	17.97	0.711	0.209
SD2-gc	1336.09	18.34	0.715	0.195
SD2-ld	**1320.51**	**18.39**	**0.716**	**0.195**
SD3-cm	3442.00	14.13	0.698	0.298
SD3-gc	**245.75**	**25.15**	**0.830**	0.155
SD3-ld	276.92	24.64	0.815	**0.145**

The primary distinction from Flickr and RefCOCOg lies in the absolute scale of reconstruction

error, with substantially higher MSE values due to the low contrast and fine-grained structure of medical images rather than a breakdown in semantic alignment. Despite this scale difference, metric ordering and cross-metric agreement are preserved, and Gaussian-center and low-dimensional masking under SD2 and SD3 consistently achieve the strongest joint reconstruction–caption performance (Table 11 and Table 12).

Table 12: Captioning performance across inpainting variants for the ROCO V2 dataset. Reported metrics include BLEU_{avg} (average of BLEU-1 and BLEU-2), METEOR (MET.), ROUGE-L (R-L), supervised SimCSE (sup.), unsupervised SimCSE (unsup.), and SBERT cosine similarity.

Variant	BLEU_{avg}	MET.	R-L	sup.	unsup.	SBERT
BLIP						
SD1.5-cm	0.022	0.041	0.086	0.322	0.250	0.227
SD1.5-gc	0.024	0.044	0.090	0.326	0.257	0.228
SD1.5-ld	0.024	0.043	0.088	0.319	0.253	0.226
SD2-cm	0.026	0.047	0.094	0.366	0.296	0.260
SD2-gc	0.026	0.046	0.094	0.365	0.288	0.251
SD2-ld	0.027	0.047	0.096	0.363	0.291	0.252
SD3-cm	0.025	0.044	0.092	0.310	0.245	0.209
SD3-gc	0.027	0.047	0.098	0.397	0.329	0.268
SD3-ld	0.030	0.051	0.101	0.413	0.340	0.284
QWEN						
SD1.5-cm	0.086	0.120	0.177	0.589	0.585	0.409
SD1.5-gc	0.086	0.118	0.175	0.590	0.588	0.404
SD1.5-ld	0.087	0.121	0.179	0.592	0.585	0.405
SD2-cm	0.092	0.125	0.187	0.613	0.608	0.433
SD2-gc	0.092	0.126	0.190	0.611	0.606	0.432
SD2-ld	0.094	0.129	0.187	0.609	0.604	0.429
SD3-cm	0.094	0.124	0.186	0.597	0.595	0.409
SD3-gc	0.092	0.125	0.189	0.616	0.608	0.441
SD3-ld	0.093	0.127	0.191	0.619	0.610	0.440

K Failure Cases

Both the GTZAN and X-ray datasets fail to exhibit meaningful reconstruction to caption correlations despite large variations in reconstruction fidelity across inpainting variants (see Figure 13 and 14). This failure is not caused by limited reconstruction diversity. Metrics such as MSE, PSNR, and LPIPS span wide ranges across variants. Instead, the failure arises from caption impoverishment.

We report results using LP-MusicCaps (Doh et al., 2023) and LAION-CLAP (Wu et al., 2023) and for inpainting we use Riff-ControlNet (Shah et al., 2024) and Riffusion (Forsgren and Martiros, 2022). For GTZAN, captions are extremely short, generic, or effectively label-like, often consisting of genre names or minimal descriptors. As a result, captioning models generate nearly identical outputs for clean and heavily corrupted inputs. Caption quality metrics, therefore, remain almost constant, making them insensitive to reconstruction quality. Even large changes in reconstruction fidelity do not translate into measurable differences in caption performance.

A similar pattern is observed for X-ray images, where captions are repetitive and coarse-grained, typically describing global anatomical structures rather than localized visual evidence. The masked regions introduced during inpainting do not consistently overlap with the visual cues emphasized in the captions. Consequently, reconstruction differences have little influence on caption generation, leading to weak or unstable correlations across all metric pairs.

We include these datasets intentionally to test whether reconstruction to caption relationships persist under minimal linguistic variability. Their failure confirms that expressive and semantically rich captions are a necessary condition for reconstruction-aware caption evaluation. When captions lack modality-specific detail, reconstruction fidelity becomes largely irrelevant to downstream language outputs, even under substantial visual perturbations.

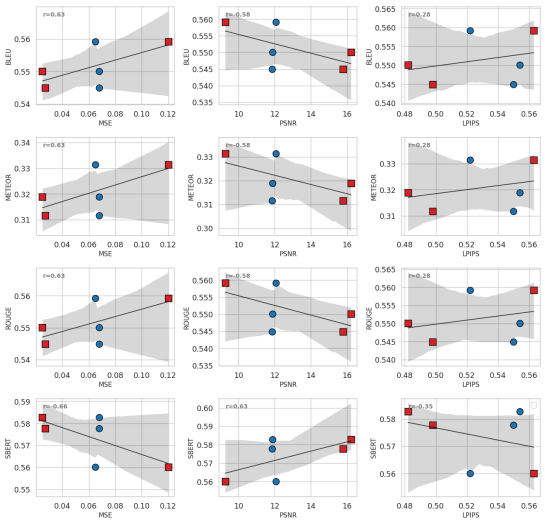

Figure 13: Relationship between reconstruction fidelity and captioning performance across Stable Diffusion variants and LAION model using the GTZAN dataset

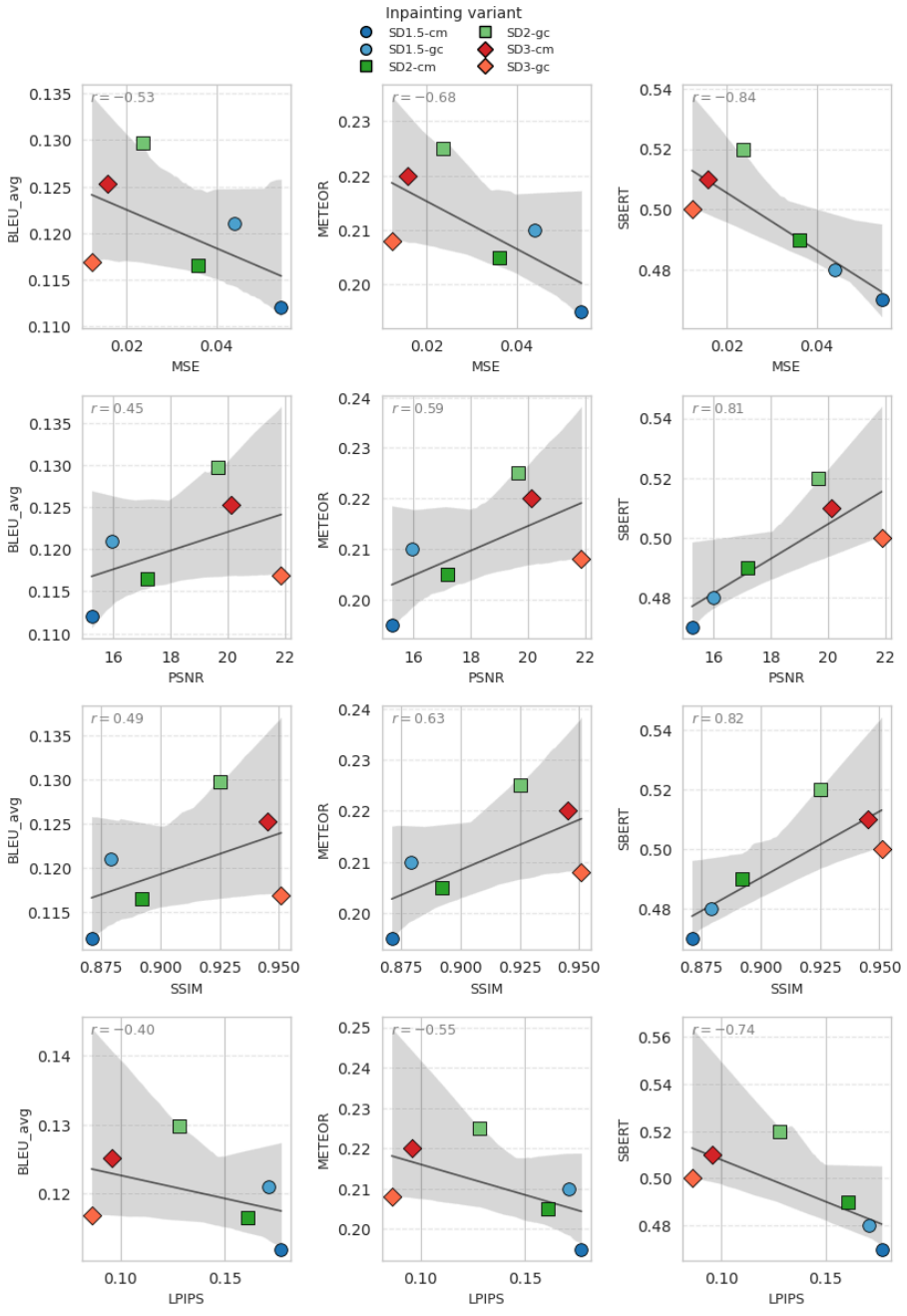

Figure 11: Relationship between reconstruction fidelity and captioning performance across Stable Diffusion variants and Qwen2.5-VL model using the TRUCE dataset

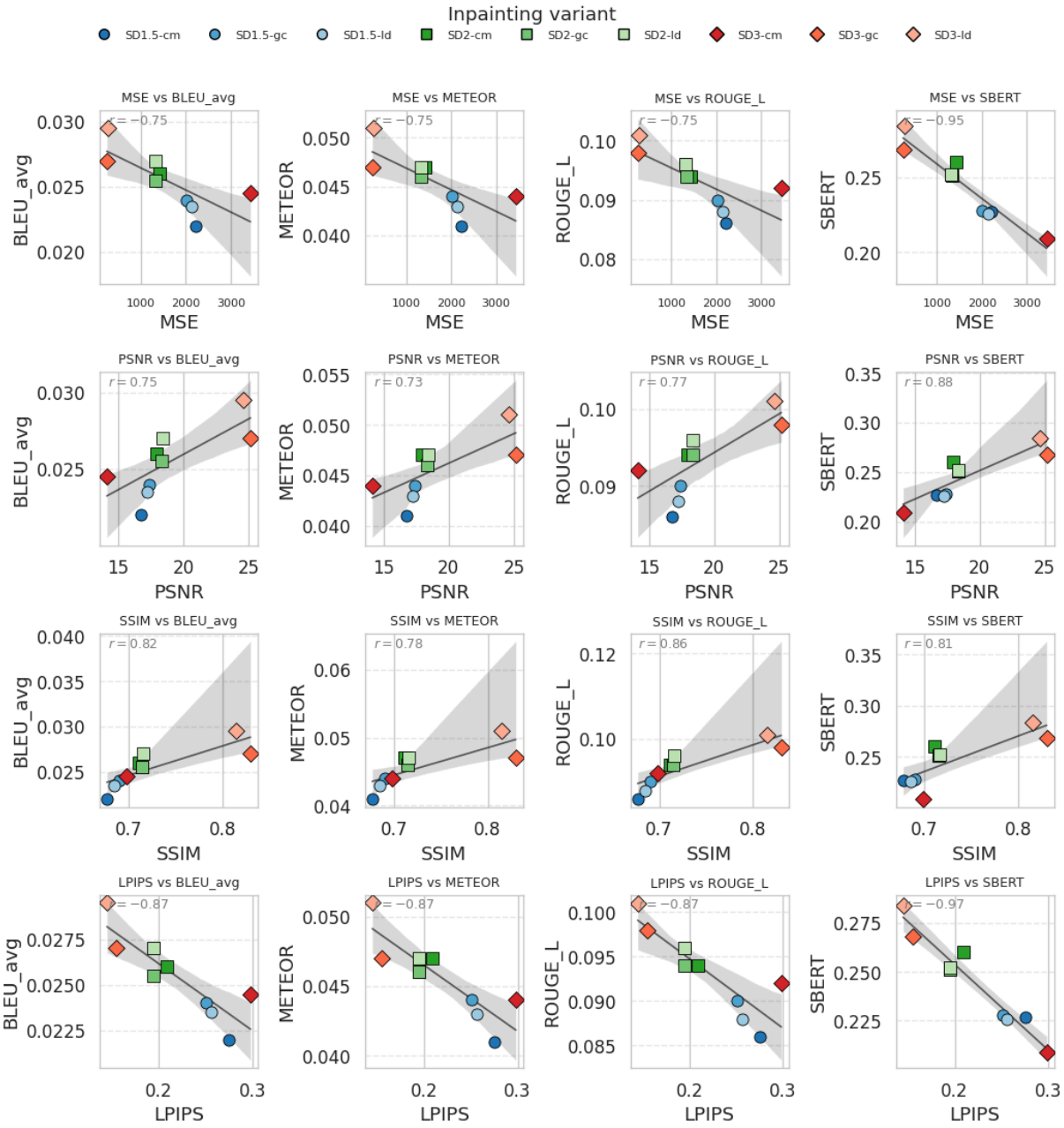

Figure 12: Relationship between reconstruction fidelity and captioning performance across Stable Diffusion variants and BLIP model using the ROCOv2 dataset

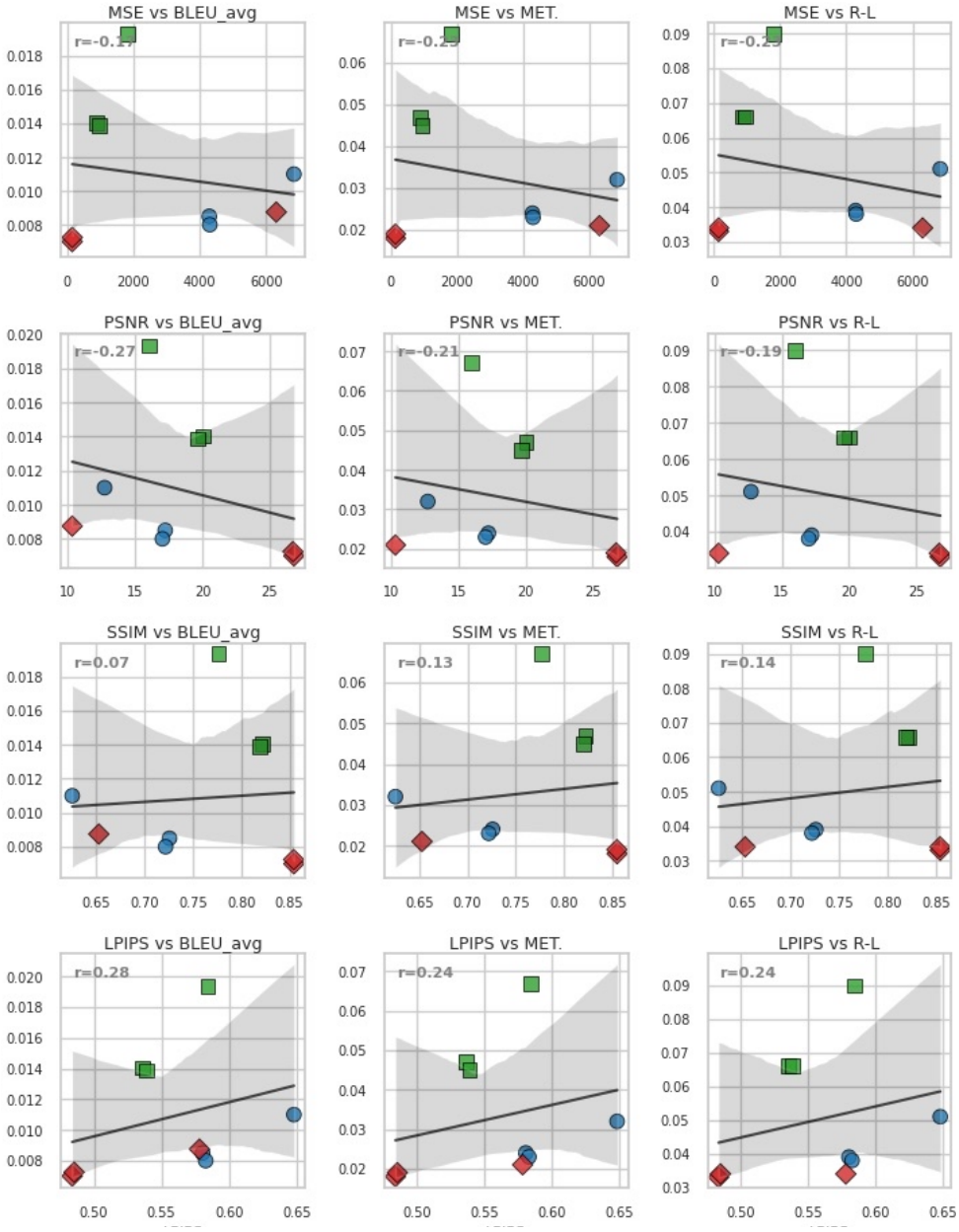

Figure 14: Relationship between reconstruction fidelity and captioning performance across Stable Diffusion variants and BLIP model using the X-ray dataset