MIMIC: Multimodal Inversion for Model Interpretation and Conceptualization

Animesh Jain, Alexandros Stergiou
University of Twente, The Netherlands
a.jain-2@alumnus.utwente.nl, a.g.stergiou@utwente.nl

Abstract

Vision Language Models (VLMs) encode multimodal inputs over large, complex, and difficult-to-interpret architectures, which limit transparency and trust. We propose a Multimodal Inversion for Model Interpretation and Conceptualization (MIMIC) framework that inverts the internal encodings of VLMs. MIMIC uses a joint VLM-based inversion and a feature alignment objective to account for VLM's autoregressive processing. It additionally includes a triplet of regularizers for spatial alignment, natural image smoothness, and semantic realism. We evaluate MIMIC both quantitatively and qualitatively by inverting visual concepts across a range of free-form VLM outputs of varying length. Reported results include both standard visual quality metrics and semantic text-based metrics. To the best of our knowledge, this is the first model inversion approach addressing visual interpretations of VLM concepts. Project page: https://anaekin.github.io/MIMIC

1. Introduction

Vision-Language Models (VLMs) have demonstrated impressive capabilities in numerous tasks. Despite their ability to encode multiple modalities, we still face difficulties in determining whether models' decisions are grounded on internal post-training reasoning [3, 5, 12] or are instead interpolated from memorized training examples [16, 22].

Uncovering visual explanations for model decisions has been the focus of many research works. Methods have visualized image region attributions [30, 34], saliency and hidden activations [26], traced information flow, or editing representations in LLMs [12, 14, 22, 42]. These methods, however, are primarily unimodal and rely on gradient access, auxiliary decoders, or architecture-specific modifications.

In this paper, we address this interpretability gap by optimizing visual inputs for VLM tokens with a Multimodal Inversion for Model Interpretation and Conceptualization (MIMIC) framework, shown in Fig. 1. MIMIC extends current unimodal inversion to autoregressive multimodal models. We use VLM logits as our optimization targets, along with a guidance objective to align the distributions of visual token encodings. Regularizers are added to promote smoothness, total variance, and alignment across distributions. We validate MIMIC's effectiveness using a diverse set of evaluation metrics for semantic alignment with textual prompts, perceptual quality, and embedding similarity.

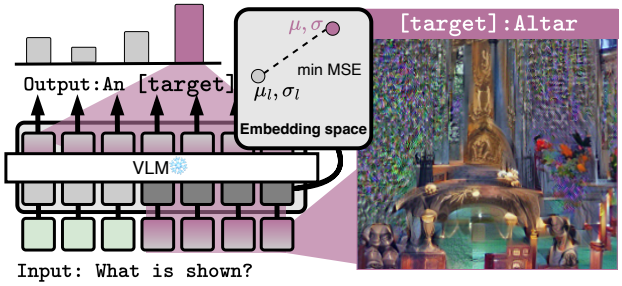

Figure 1. **Multimodal Inversion for Model Interpretation and Conceptualization (MIMIC)** inverts VLMs by synthesizing visual inputs that best correspond to VLM tokens and internal embeddings. The synthesized images represent the dominant visual features associated with predicted tokens.

Our contributions are: i) A model inversion objective that can optimize visual inputs from VLM logits. ii) MIMIC, a general visual interpretability approach, which, to the best of our knowledge, is the first attempt at inverting learned VLM visual features corresponding to tokens. iii) We show that MIMIC can invert VLM text semantics of different lengths to high-fidelity images.

2. Related Work

Feature attribution methods localize feature relevance. Class Activation Maps (CAM)-based methods [30, 37, 44] propagate local activations of class-relevant regions in vision models. Pixel-level attributions, such as LRP [1], DeepLIFT [31], and IG [34], visualize gradient-based signals given target outputs. Although attribution methods are widely used to explain CNN/ViT decisions on unimodal supervised tasks, they cannot be directly applied to autoregressive multimodal models. In contrast to these feature-localization approaches, we offer a general inversion method to discover visual encodings corresponding to (semantic information of) VLM tokens.

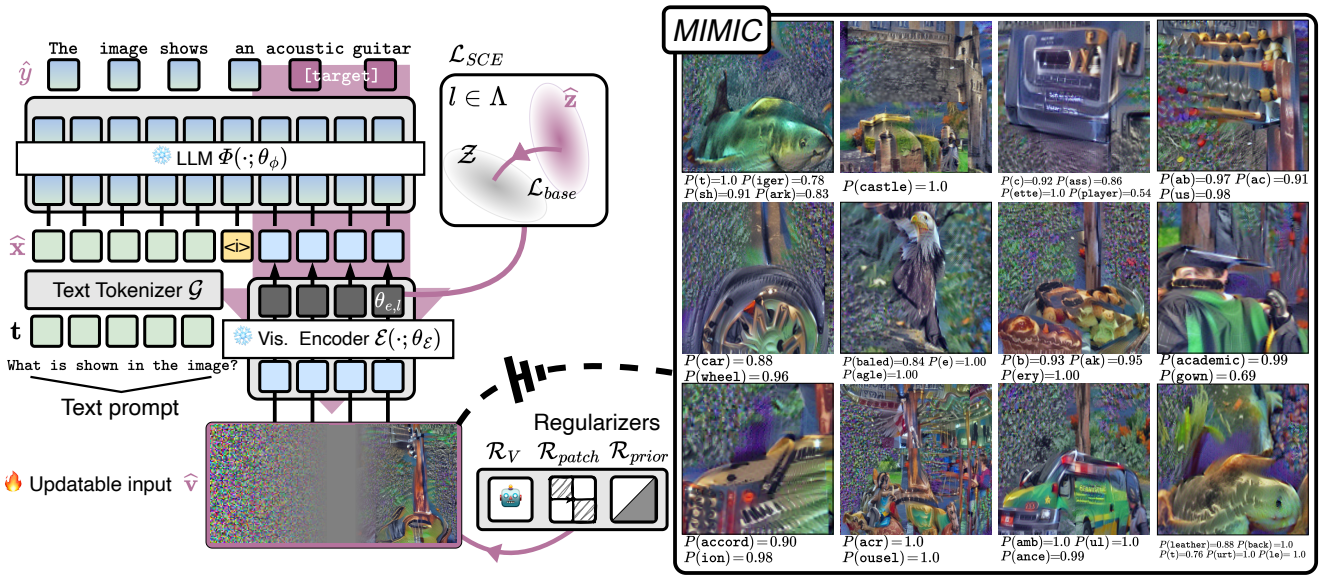

Figure 2. **MIMIC inversion** iteratively optimizes an updatable input $\hat{\mathbf{v}}$ with an adapted cross-entropy loss, $\mathcal{L}_{SCE}$, to maximize the probability distribution of `[target]` VLM token(s), and a base feature loss, $\mathcal{L}_{base}$, to match layer statistics to target mean and variance within the distribution manifold for vision tokens. Regularizers $\mathcal{R}$ are added to promote variance consistency, visual coherence across tokens, and perceptual quality. Optimized visual inputs are shown from inverting visual-instruct-tuned LLaMA3-8B alongside the per-`[target]`-token probability $P(\texttt{[target]})$.

Model inversion methods [13, 17, 32, 39] synthesize inputs to visually interpret encodings of models. Early approaches relied on Gradient Ascent (GA) by maximizing activations of specific neurons [40]. Later methods relied on generative priors [23]. More recently, [39] synthesized realistic images with CNN BN statistics, while [13, 17, 33] adapted gradient-based inversion to ViTs [8]. [10] defined hierarchies of visual concepts, and [36] discovered shared features across models and datasets. Although these inversion techniques have improved the interpretability of vision-only models, their applicability to VLMs remains limited, a problem this paper addresses.

V/LLM interpretability methods primarily aim to explain how model encodings relate to abstract concepts. [3] created text explanations for hidden LLM layers' states, later extended [12] by allowing direct modulations over hidden states. Other approaches [5, 22] focused on modulations in the model's internal structure with adapters. Works also explored attention and gradient signals. Grad-ECLIP [42] visualized image and text tokens contributions based on CLIP similarity scores. NOTICE [14] studied image-grounding and object-level reasoning properties within cross-attention heads. Second-Order Lens [11] ablated zero-shot accuracy of late layers in CLIP. [35] demonstrated the influence of intermediate layers towards semantic features, while [7, 24] focused on the alignment of linear layers with concept-relevant features. Despite these efforts, model explanations for VLMs are sparse. We thus take a step forward by inverting VLMs to visualize learned semantic features.

3. Method

In this section, we provide a detailed formulation for MIMIC. A conceptual overview of MIMIC is shown in Fig. 2. We initialize an updatable input $\hat{\mathbf{v}} \in \mathbb{R}^{C\times H\times W}$ with C channels, H height, and W width. As VLMs can use multimodal context, we include a text prompt template $\mathbf{t}$ alongside our updatable input. This takes the form of: `What is shown in the image?: a.[target]` concept, `or b.[negative]` concept. For target token `[tiger]`, this can be `a.tiger` and `b.dog`. Text is tokenized by $\mathcal{G}(\mathbf{t})$ to a sequence of embeddings. Similarly, $\hat{\mathbf{v}}$ is encoded by $\mathcal{E}$ with $\theta_\mathcal{E}$ parammeters, to embeddings tensor $\mathcal{E}(\hat{\mathbf{v}};\theta_\mathcal{E}) \in \mathbb{R}^{D\times\Omega}$ of D vision tokens of Ω channels. The concatenated text-image context used by the LLM is $\hat{\mathbf{x}} = [\mathcal{G}(\mathbf{t}), \mathcal{E}(\hat{\mathbf{v}};\theta_\mathcal{E})]$.

VLM Inversion. The backbone LLM $\Phi(\cdot;\theta_\phi)$, with θ_ϕ frozen params, infers $\hat{\mathbf{x}}$ and returns a probabilistic distribution of token logits. Each logit $\hat{\mathbf{y}}_i = \Phi(\hat{\mathbf{x}}, \hat{\mathbf{y}}_{<i};\theta_\phi)$, where $\hat{\mathbf{y}}_{<i}$ are the previously-generated $i-1$ logits. Due to LLM's autoregressive nature, generated logits sequences vary in length $|\hat{\mathbf{y}}|$. We define an adapted CE loss $\mathcal{L}_{SCE}$ given the token index i for `[target]`:

$$\mathcal{L}_{SCE}(\hat{\mathbf{y}}) = -\sum \mathbb{1}(\text{sg}(\hat{\mathbf{y}}), i, \texttt{[target]}) \log(\hat{\mathbf{y}}_i), \quad (1)$$

where $\mathbb{1}(\text{sg}(\hat{\mathbf{y}}), i, \texttt{[target]})$ is the indicator function given sequence index i and target token dictionary index `[target]`, stop gradient $\text{sg}(\hat{\mathbf{y}})$, and sequence length $|\hat{\mathbf{y}}|$.

Optimization Objectives	LLaMA3-8B				Mistral-7B				Vicuna-7B				Vicuna-13B			
	FID ↓	LPIPS ↓	IS ↑	CScr ↑	FID ↓	LPIPS ↓	IS ↑	CScr ↑	FID ↓	LPIPS ↓	IS ↑	CScr ↑	FID ↓	LPIPS ↓	IS ↑	CScr ↑
Partial objectives																
Base ($\mathcal{L}_{SCE}$)	421.63	1.71	1.21	21.84	469.42	1.88	1.03	20.64	404.51	1.68	1.10	21.44	389.45	2.14	0.95	22.38
Base+feat ($\mathcal{L}_{SCE}$+$\mathcal{L}_{base}$)	317.45	0.98	3.32	26.58	340.08	1.13	2.97	26.72	326.78	1.07	3.49	27.64	277.63	1.42	3.60	25.86
MIMIC	**178.42**	**0.73**	**5.38**	**29.43**	**184.56**	**0.75**	**5.18**	**30.51**	**162.92**	**0.83**	**5.21**	**29.17**	**145.19**	**0.64**	**5.77**	**32.50**

Table 1. **Image and semantic quality** comparisons. FID and LPIPS are computed to real images from [29] with results averaged over 1K `[targets]`. Each loss improves the results. The aggregated objective yields the best overall performance, while each component improves across semantic, perceptual, and distributional metrics. Best results are in **bold**.

We minimize this loss to enforce semantic alignment between the updatable image and the target token logits.

Base Feature Loss. Although the CE loss provides a strong signal for input updates, it does not guarantee that layer encodings will be within the exact manifold of target tokens' visual features. To align layer encodings $\hat{\mathbf{z}}_l = \mathcal{E}(\hat{\mathbf{v}}, \theta_{e,<l})$ with the internal representations of the vision encoder, we approximate the manifold's mean $\mu(\mathcal{Z}_l)$ and variance $\sigma(\mathcal{Z}_l)$ from `[target]` images given $\theta_{\mathcal{E},l}$ weights, across $l \in \Lambda = \{1, ..., L\}$ layers. Internal statistics $\mu(\mathcal{Z}_l)$,$\sigma(\mathcal{Z}_l)$ can also be approximated without sourced images by using a generator, as we demonstrate in §A-2. The distribution-matching feature loss takes the form:

$$\mathcal{L}_{base} = \sum_{l \in \Lambda} \left(\|\mu(\hat{\mathbf{z}}, \theta_{\mathcal{E},l}) - \mu(\mathcal{Z}_l)\|_2^2 + \|\sigma(\hat{\mathbf{z}}, \theta_{\mathcal{E},l}) - \sigma(\mathcal{Z}_l)\|_2^2 \right) \quad (2)$$

where Λ denotes ViT layers and $\mathcal{Z}_l$ are the approximated distribution for the `[target]` feature manifold at layer l.

Regularizers. VLMs encode multimodal inputs across large, complex feature spaces. We thus improve the optimization objective by including standard regularizers from seminal model inversion methods [13, 17, 39]. We use $\mathcal{R}_{patch}$ to smooth color signal variance across ViT tokens and improve the overall uniformity across image patches. We include $\mathcal{R}_{\text{prior}}$ as a combination of image priors for the total L1/2 feature variance and spatial smoothness $\mathcal{R}_{TV_1}$ $\mathcal{R}_{TV_2}$ with a ℓ_2-norm penalty $\mathcal{R}_{l_2}$ to regularize the range:

$$\mathcal{R}_{prior}(\hat{\mathbf{v}}) = \alpha_1 \mathcal{R}_{TV1}(\hat{\mathbf{v}}) + \alpha_2 \mathcal{R}_{TV_2}(\hat{\mathbf{v}}) + \alpha_3 \mathcal{R}_{\ell_2}(\hat{\mathbf{v}}) \quad (3)$$

where, $\alpha_1,\ \alpha_2,\ \alpha_3$ are the scaling factors used. Regularizer $\mathcal{R}_V$ constraints high-frequency noise (L2) inside each patch to improve realism. The aggregated regularizer $\mathcal{R}(\hat{\mathbf{v}})$ uses $\beta_1,\ \beta_2$ scaling factors and is defined as:

$$\mathcal{R}(\hat{\mathbf{v}}) = \beta_1\ \mathcal{R}_V(\hat{\mathbf{v}}) + \beta_2 \mathcal{R}_{patch}(\hat{\mathbf{v}}) + \mathcal{R}_{prior}(\hat{\mathbf{v}}) \quad (4)$$

MIMIC updates Our objective updates $\hat{\mathbf{v}}$ per s iterations:

$$\hat{\mathbf{v}}^{s+1} = \min_{\hat{\mathbf{v}}^s} \gamma_1 \mathcal{L}_{SCE}(\Phi([\mathcal{G}(\mathbf{t}), \mathcal{E}(\hat{\mathbf{v}}; \theta_e)]; \theta_\phi)) + \gamma_2 \mathcal{L}_{base} + \mathcal{R}(\hat{\mathbf{v}}) \quad (5)$$

it combines (1) and (2) losses and the regularizers from (4) with $\gamma_1,\ \gamma_2$ factors. The final MIMIC objective inverts VLM encodings relevant to `[target]` tokens. We note that the method is invariant to the length of `[target]` and can be used with varying prompted texts $\mathbf{t}$.

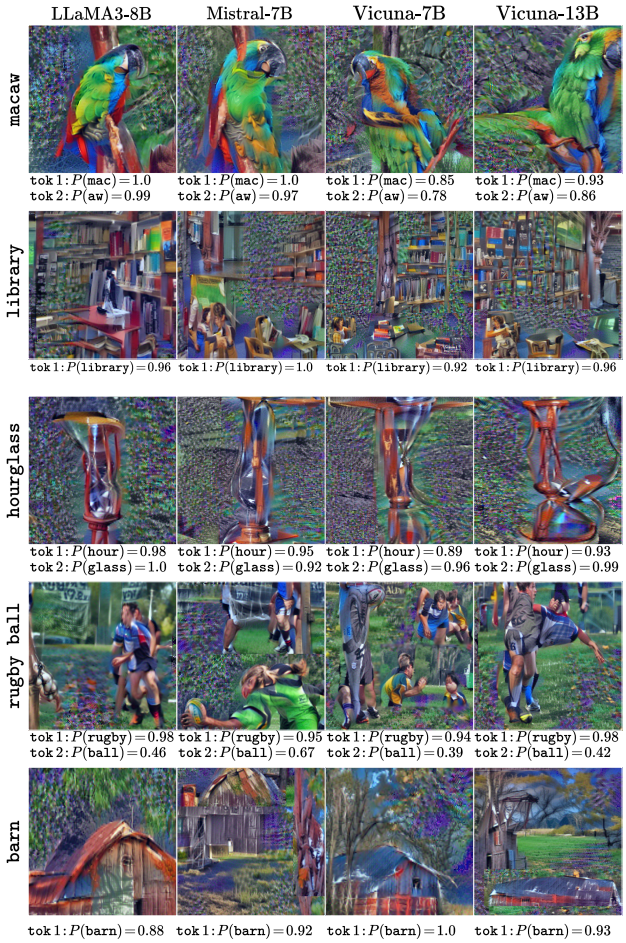

Figure 3. **Qualitative examples of synthesized visual inputs with MIMIC**. Each row optimizes different token logits. The number of tokens corresponding to a `[target]` semantic concept differs per row.

4. Results

Model Details We invert visual-instruct-tuned [21] LLaMA3-8B [15], Mistral-7B [20], and Vicuna-7/13B [43]. Models only run inference with the updatable input $\hat{\mathbf{v}} \in \mathbb{R}^{3\times448\times448}$ initialized from a Gaussian $\hat{\mathbf{v}} \sim \mathcal{N}(0, 1)$. We compute $\mathcal{L}_{base}$ statistics from [6]. We use a 100-iteration warmup exponential lr scheduler

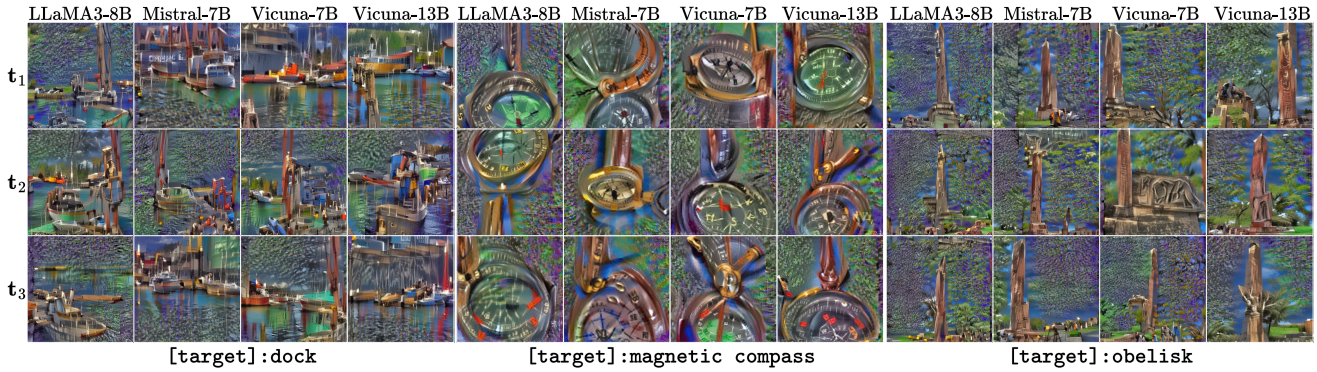

Figure 4. **Synthesized images over varying text prompts** for `dock`, `magnetic compass`, and `obelisk`. $\mathbf{t}_1$; `What is shown in the image?a.[target] or b.[negative]`, $\mathbf{t}_2$; `Does the image show an instance of [target] or [negative]?`, and $\mathbf{t}_3$; `The image depicts a scene that corresponds to [target] or [negative]?`

Length	BLEU↑	METEOR↑	ROUGE-L↑
$\lvert\hat{\mathbf{y}}\rvert \leq 2$	0.933	0.459	0.875
$\lvert\hat{\mathbf{y}}\rvert \leq 4$	0.928	0.465	0.880
$\lvert\hat{\mathbf{y}}\rvert \leq 6$	0.945	0.471	0.891
$\lvert\hat{\mathbf{y}}\rvert \geq 7$	0.936	0.464	0.886

Table 2. **Predicted to `[target]` VLM outputs** across text similarity metrics (BLEU, METEOR, ROUGE-L) on LLaMA3-8B. Results are grouped by text length.

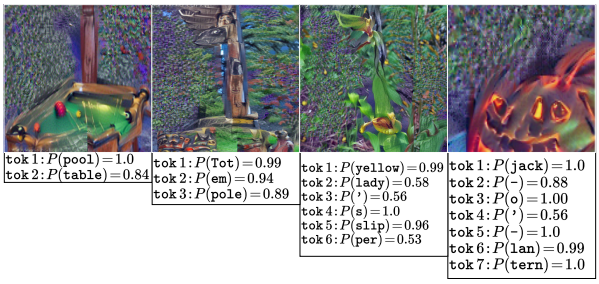

Figure 5. **Examples over varying lengths** with Vicuna-13B.

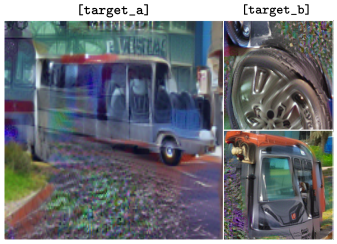

Figure 6. **Combined ablations by changing from `[target_a]` to `[target_b]`** on Mistral-7B.

starting from 0.05 over 5 gradient accumulation steps. A single run takes 15 min. on an L40s. Hyperparameters $\{a_1, a_2, a_3, \beta_1, \beta_2, \gamma_1, \gamma_2\}$ for all models are set with parallel hyperparameter tuning [28]. We run this once (3h) and use the same hyperparams throughout experiments.

Evaluation Metrics. We evaluate image and semantic quality using perceptual similarity [41], semantic alignment [18], and image quality [19, 27] metrics. Different `[target]` token lengths are quantitatively evaluated with BLEU [25], METEOR [2], and ROUGE-L [4] across $|\hat{\mathbf{y}}|$.

Main results. We metrically assess synthesized images' quality in Tab. 1 reporting Fréchet inception distance (FID) [19] and Learned Perceptual Image Patch Similarity (LPIPS) [41] between real [29] and synthesized images. The results show that the joint MIMIC objective is critical for obtaining understandable visualizations of features. MIMIC images are also better semantically aligned with respect to `[target]` tokens given the Inception Score (IS) [27] and CLIPScore (CScr) [18] performance.

Qualitative examples. Alignment is also shown with qualitative examples in Fig. 3. The images depict learned visual characteristics such as green-red feathers and black beak for `[macaw]`, bookshelves and study desks for `[library]`, and glass reflections for `[hourglass]`. The visualizations further allow exploration of learned correlations, such as the primary association of sport jerseys and shorts with `[rugby ball]` and prairies when optimizing `[barn]`.

Ablation studies. We further ablate template $\mathbf{t}$ across vision tokens $\hat{\mathbf{v}}$ in Fig. 4. MIMIC can robustly visualize the main learned features, such as water reflections in `[dock]` and the dial plate in `[magnetic compass]`. MIMIC also effectively inverts multi-tokens outputs of varying `[target]` lengths $|\hat{\mathbf{y}}|$, shown metrically in Fig. 6 with qualitative examples in Fig. 5. Our approach further allows the discovery of learned attributes through Chain-of-Through prompting [38]. In Fig. 6 (left side) we first use template $\mathbf{t}_a$: `What is shown in the image?` with `[minibus]` as target. Then (right side), we update to $\mathbf{t}_b$: `A minivan is shown in the image. What's the main characteristic?` and infer `[wheels]` and `[windshield]` from Mistral-7B. In turn, we use these responses as `[targets]`.

5. Conclusion

We propose MIMIC, a framework for visually inverting VLMs that combines inversion and feature alignment objectives with spatial alignment, image smoothness, and semantic realism regularizers. MIMIC can be applied across diverse models and settings to identify the main features learned by VLMs. We believe this first step for interpreting VLM representations is a promising direction towards understanding multi-modal encodings.

References

[1] Sebastian Bach, Alexander Binder, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 2015. 1

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACLw*, 2005. 4

[3] Haozhe Chen, Carl Vondrick, and Chengzhi Mao. Selfie: Self-interpretation of large language model embeddings. In *ICML*, 2024. 1, 2

[4] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *TSBOw*, 2004. 4

[5] Yu-Neng Chuang, Guanchu Wang, et al. Faithlm: Towards faithful explanations for large language models. *arxiv:2402.04678*, 2024. 1, 2

[6] Jia Deng, Wei Dong, et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[7] Teresa Dorszewski, Lenka Tětková, et al. From colors to classes: Emergence of concepts in vision transformers. *arxiv:2503.24071*, 2025. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arxiv:2010.11929*, 2021. 2

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1

[10] Thomas Fel, Agustin Picard, et al. Craft: Concept recursive activation factorization for explainability. In *CVPR*, 2023. 2

[11] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip. In *ICLR*, 2025. 2

[12] Asma Ghandeharioun, Avi Caciularu, et al. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *ICML*, 2024. 1, 2

[13] Amin Ghiasi, Hamid Kazemi, et al. Plug-in inversion: Model-agnostic inversion for vision with data augmentations. In *ICML*, 2022. 2, 3

[14] Michal Golovanevsky, William Rudman, et al. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *ACL*, 2025. 1, 2

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 3

[16] Roger Grosse, Juhan Bae, et al. Studying large language model generalization with influence functions. *arxiv:2308.03296*, 2023. 1

[17] Ali Hatamizadeh, Hongxu Yin, et al. Gradvit: Gradient inversion of vision transformers. In *CVPR*, 2022. 2, 3

[18] Jack Hessel, Ari Holtzman, et al. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLPS*, 2022. 4

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 4

[20] Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv. *arXiv:2310.06825*, 2023. 3

[21] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv:2407.07895*, 2024. 3

[22] Kevin Meng, David Bau, et al. Locating and editing factual associations in gpt. In *NeurIPS*, 2023. 1, 2

[23] Anh Nguyen, Jeff Clune, et al. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017. 2

[24] Tuomas Oikarinen and Tsui-Wei Weng. Linear explanations for individual neurons. In *ICML*, 2024. 2

[25] Kishore Papineni, Salim Roukos, et al. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 4

[26] Maithra Raghu, Thomas Unterthiner, et al. Do vision transformers see like convolutional neural networks? In *NeurIPS*, 2021. 1

[27] Tim Salimans, Ian Goodfellow, et al. Improved techniques for training gans. *NeurIPS*, 2016. 4

[28] Sandeep Singh Sandha, Mohit Aggarwal, Igor Fedorov, and Mani Srivastava. Mango: A python library for parallel hyperparameter tuning. In *ICASSP*, 2020. 4

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 3, 4

[30] Ramprasaath R Selvaraju, Michael Cogswell, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 1

[31] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017. 1

[32] Alexandros Stergiou. The mind's eye: Visualizing class-agnostic features of cnns. In *ICIP*, 2021. 2

[33] Alexandros Stergiou and Nikos Deligiannis. Leaping into memories: Space-time deep feature synthesis. In *ICCV*, 2023. 2

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 1

[35] Mingxu Tao, Quzhe Huang, et al. Probing multimodal large language models for global and local semantic representations. In *LREC-COLING*, 2024. 2

[36] Harrish Thasarathan, Julian Forsyth, et al. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *ICML*, 2025. 2

[37] Haofan Wang, Zifan Wang, et al. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPRw*, 2020. 1

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 4

[39] Hongxu Yin, Pavlo Molchanov, et al. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, 2020. 2, 3

[40] Jason Yosinski, Jeff Clune, et al. Understanding neural networks through deep visualization. In *ICMLw*, 2015. 2

[41] Richard Zhang, Phillip Isola, et al. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[42] Chenyang Zhao, Kun Wang, et al. Grad-eclip: Gradient-based visual and textual explanations for clip. In *ICML*, 2024. 1, 2

[43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023. 3

[44] Bolei Zhou, Aditya Khosla, et al. Learning deep features for discriminative localization. In *CVPR*, 2016. 1

MIMIC: Multimodal Inversion for Model Interpretation and Conceptualization

Appendix

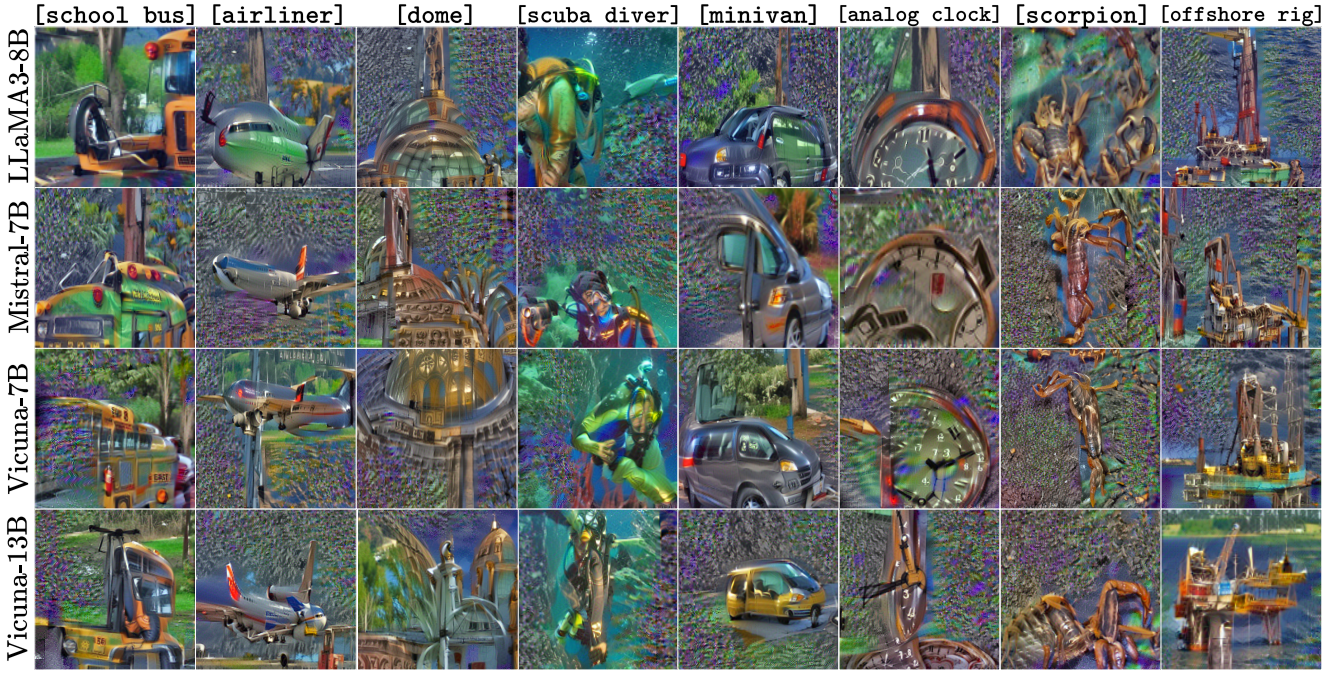

Figure A-1. **Qualitative examples of synthesized features** across `[target]` tokens.

A-1. Additional qualitative results

We demonstrated qualitative results of inverted `[target]` VLM tokens in Sec. 4. Supplementary to Fig. 3, we further visualize additional examples across VLMs in Fig. A-1. As shown, MIMIC synthesizes coherent features across models with various target tokens. Descriptive VLM features learned for target semantics are often based on distinct shapes such as the examples for `[airliner]` and `[offshore rig]`. Positive correlations between materials and colors are also learned for instances such as `[school bus]`, `[dome]`, and `[minivan]`. VLMs also learn environment associations, such as oxygen bubbles, for `[scuba diving]`.

A-2. Generated base feature priors

As described in Sec. 3, internal vision encoder layer statistics $\mu(\mathcal{Z}_l)$,$\sigma(\mathcal{Z}_l)$ can be approximated without sourced images. For this, we instead approximate `[target]` manifold's distribution by sourcing images generated with SD3-M [9] with prompt template `A photorealistic image of [target]` as the condition. Fig. A-2 demonstrates that statistics from generated images do not result in a significant quality drop, with images still including visually distinct features.

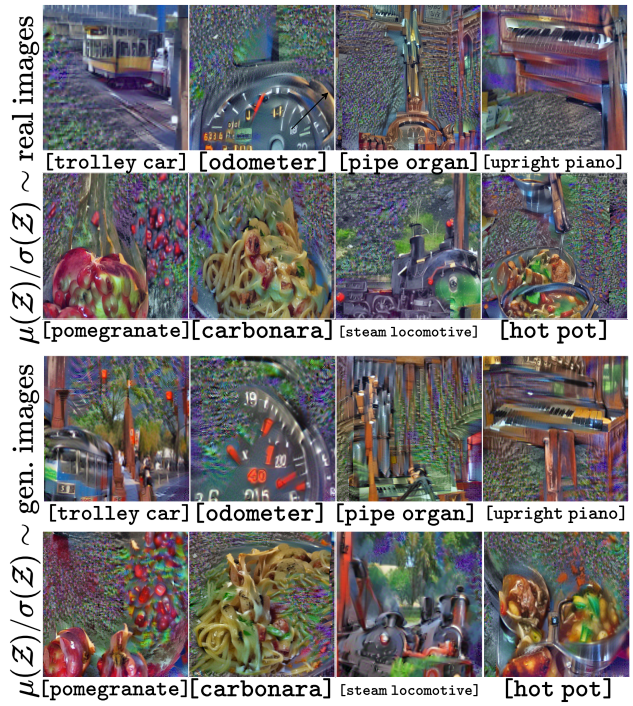

Figure A-2. **Comparisons between statistics from real and generated images**. (top) includes base feature loss as in Sec. 3. (bottom) computes base loss from SD3-M [9] image embeddings.