

# CAST: Cross-modal Alignment Similarity Test for Vision Language Models

**Gautier Dagan**  
University of Edinburgh  
gautier.dagan@ed.ac.uk

**Olga Loginova**  
University of Trento  
olga.loginova@unitn.it

**Anil Batra**  
University of Edinburgh  
a.k.batra@sms.ed.ac.uk

## Abstract

Vision Language Models (VLMs) are typically evaluated with Visual Question Answering (VQA) tasks which assess a model’s understanding of scenes. Good VQA performance is taken as evidence that the model will perform well on a broader range of tasks that require both visual and language inputs. However, scene-aware VQA does not fully capture input biases or assess hallucinations caused by a misalignment between modalities. To address this, we propose a Cross-modal Alignment Similarity Test (CAST) to probe VLMs for *self-consistency* across modalities. This test involves asking the models to identify similarities between two scenes through text-only, image-only, or both and then assess the truthfulness of the similarities they generate. Since there is no ground truth to compare against, this evaluation does not focus on objective accuracy but rather on whether VLMs are internally consistent in their outputs. We argue that while not all self-consistent models are capable or accurate, all capable VLMs must be self-consistent.

## 1 Introduction

Vision Language Models (VLMs) integrate vision and language modalities to learn image-text correspondences from large-scale image-text pairs (Zhang et al., 2023; Radford et al., 2021a; Kwon et al., 2022). Given image-text pairs, VLMs combine a text encoder and an image encoder to extract image and text features and then learn to align vision and language through generative objectives, such as Visual Question Answering (VQA). As a result, VLMs pose a unique challenge in ensuring consistent outputs across different input types – be it text, images, or a combination of both.

Consistency in AI models is essential for their reliability and trustworthiness (Ji et al., 2023). Self-consistency refers to a model’s ability to produce stable, coherent outputs across similar inputs and conditions (Elazar et al., 2021). If a VLM exhibits

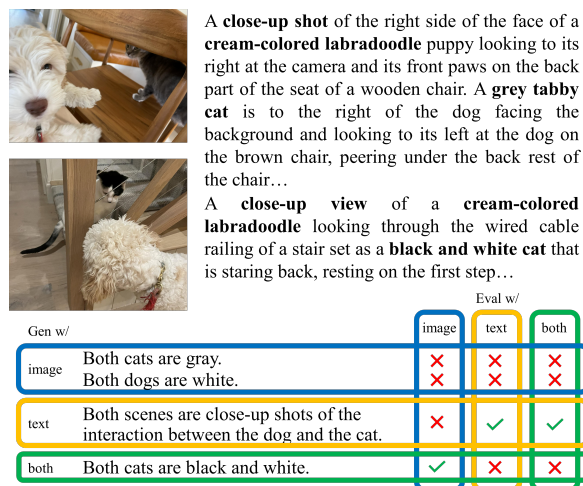


Figure 1: Example of paired scenes and statements from the CAST dataset. Horizontal blocks show generated statements, while vertical blocks are evaluations for each modality: **image-only**, **text-only**, and **image+text**. Red crosses indicate where each model disagrees with its own generation during the evaluation step. Similarity topics are highlighted in **bold**. Note that VLMs may produce hallucinations, as the CAST method checks for consistency rather than correctness.

inconsistent behavior when given the same input across different modalities, it could raise concerns about its robustness and internal reasoning. So while models might perform well on major VQA benchmarks, such as MMMU (Yue et al., 2023) or MME (Fu et al., 2023), we argue that they must also be evaluated for self-consistency.

We propose to evaluate self-consistency through the absence of contradictions between a model’s generated output and the evaluation of this output by different modalities. To this end, we introduce the two-step Cross-modal Alignment Similarity Test (CAST).<sup>1</sup> We apply CAST to different VLMs and find that despite strong performance on many other downstream tasks, the majority of VLMs exhibit a lack of internal self-consistency

<sup>1</sup>We publicly release all our code and dataset [here](#).

and modality alignment (see examples in Figure 1). CAST provides a more nuanced understanding of VLMs’ reasoning capabilities and potential biases, which is critical for real-world applications.

## 2 Related Works

Traditionally, self-consistency has been tested through meaning-preserving alternations to model’s inputs, such as adding illogical statements, filler tokens, or paraphrasing (Elazar et al., 2021; Parcalabescu and Frank, 2024; Yue et al., 2024). Logical consistency ensures that the model’s outputs remain coherent and non-contradictory throughout multiple iterations (Yang et al., 2024; Zhang et al., 2024). We instead design CAST to evaluate the **cross-modal** consistency of VLMs through a comparison task.

Several work have proposed image comparison benchmarks for VLMs (Fu et al., 2024; Zhao et al., 2024; Dunlap et al., 2023). They focus on contrastive pairs where similar images differ in key features. For example, VisDiffBench (Dunlap et al., 2023) uses human-annotated differences between two sets of images. In contrast, we focus on similarities to capture the semantic overlap between example pairs.

Self-consistency in VLMs and LLMs is closely tied to uncertainty in predictions, resulting in noisier outputs (Chen et al., 2024). Consequently, it is used as a metric to detect hallucinations caused by misalignment (Manakul et al., 2023; Mündler et al., 2024; Li et al., 2024). CAST can also reveal logical hallucinations where a model’s uncertainty causes it to be inconsistent.

## 3 Method

We propose CAST (shown in Figure 3) as a fully automated two-step approach to evaluate multi-modal self-consistency in VLMs.

### 3.1 Generating Similarities

CAST leverages similarities between two scenes to assess a model’s ability to evaluate its own outputs. In our case, a scene is an image paired with its high-quality description (see Section 4.1). By focusing on shared features, the model is less likely to rely on surface-level distinctions or superficial strategies. For instance, if tasked with finding differences between two images, the model might only attend to one image or highlight minor details like color changes. Emphasizing similarities

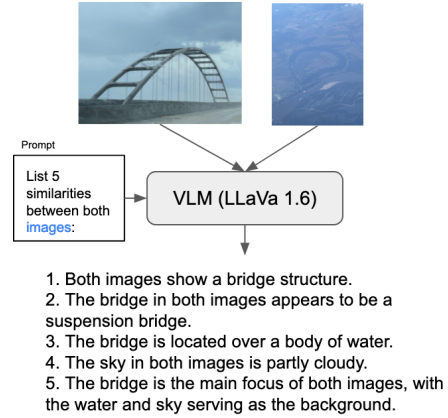


Figure 2: Example of the first step of CAST: **Generation**. We pass a pair of related examples (either as images, text descriptions, or both) and prompt the VLM to generate a set of similarity statements. In the second step, the truthfulness of each statements is evaluated by the same VLM for each sets of modality.

encourages a deeper evaluation of each input.

The first step of CAST is to prompt the VLM to generate a number of statements about the similarities between two input scenes  $S_A$  and  $S_B$ . Since we generate a list of similarities using the VLM, each subsequent similarity statement is conditioned on all previously generated ones. We can view the generation of a given similarity as following:

$$sim_0 = VLM(S_A, S_B, P^{gen}) \quad (1)$$

$$sim_i = VLM(sim_{i-1}, \dots, sim_0; S_A, S_B, P^{gen}), \quad (2)$$

where  $P^{gen}$  are the instructions. Similarity statements are generated for different modalities: scenes can be represented as two images ( $S^{img}$ ), two text descriptions ( $S^{txt}$ ), or two images combined with the corresponding descriptions ( $S^{img+txt}$ ). In Figure 2 we show an example of a set of similarity statements generated by  $S^{img}$ . We obtain the similarity statements conditioned on a pair of scenes for each modality stream. We restrict the input pairs to the same modality and generate all statements using greedy sampling ( $t = 0$ ).

### 3.2 Evaluating Similarities

The second step of our approach is to evaluate each similarity statement and test whether a model remains consistent under different modalities. Since we focus on self-consistency, we use the same model for both generation and evaluation. The evaluation step can be represented as following:

$$s = VLM(S_A, S_B, P^{eval}), \quad (3)$$

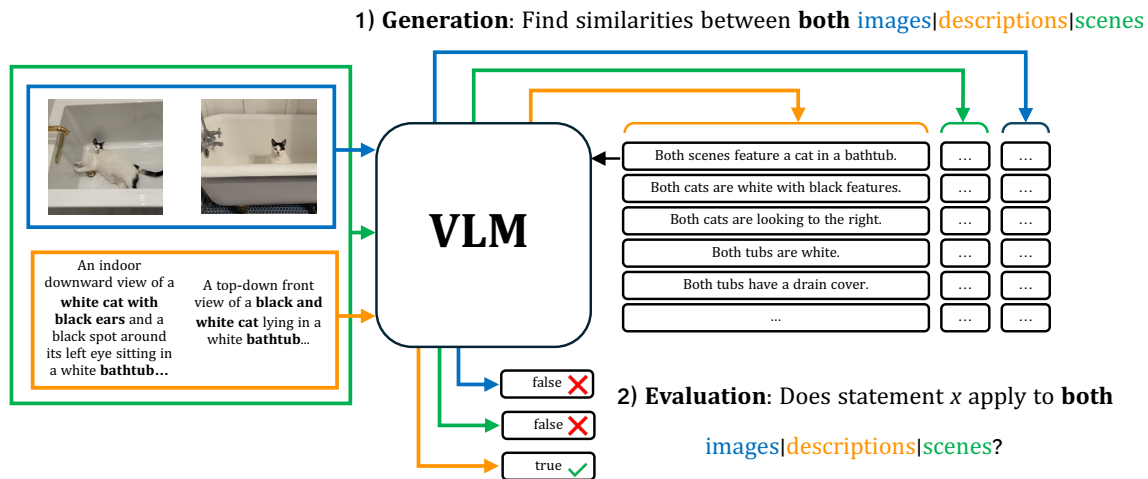


Figure 3: CAST is two-fold. In the first step, we ask the model to *generate* a set of similarity statements conditioned on different modality input types (image-only, text-only, both). In the second step, the model *validates* the truthfulness of the generated statements with respect to each modality. This allows us to measure whether the VLM is self-consistent within a modality and across different modalities.

where  $s$  is 1 if the model confirms that the statement is true and 0 otherwise. We filter out the generations that cannot be parsed (see Appendix A for details). To mitigate bias towards a certain prompt or phrasing (Pezeshkpour and Hruschka, 2024; Sclar et al., 2024), we use three different evaluation prompts. Thus, apart from the conventional Yes/No questions, we ask the model whether the statement applies to one or both scenes and whether the statement is true or false. To quantify self-consistency, we report the average  $s$  over all the evaluated pairs and prompts for each modality permutation (both generated and evaluated with).

## 4 Experiments and Results

### 4.1 Dataset

Since CAST relies on asking VLMs to find similarities between two scenes, we need a multi-modal set of pairs of aligned images/descriptions that contain similarities. To construct our evaluation dataset, we sub-sample example pairs from the DOCCI Dataset (Onoe et al., 2024). The dataset contains 15k images paired with human-annotated descriptions of 136 words on average. The images focus on spatial relations and world knowledge. Unlike popular captioning datasets, each description is comprehensively annotated to capture the differences between similar images.

We randomly sample 100 pairs of images from the DOCCI train dataset of 10k images. We threshold the CLIP (Radford et al., 2021b) cosine-

similarity and filter out the pairs of the  $< 0.75$  CLIP score (since the images might not have enough in common), or  $\geq 0.95$  CLIP score (to exclude near identical ones and duplicates). We also filter out images with captions of less than 500 characters to include only those that contain ample descriptive information about the scene.

### 4.2 Models

We test the following open-source and closed-source VLMs for self-consistency, each with distinct vision encoders, language models, and training dataset:

- **Bunny 1.1** (He et al., 2024)
- **LLaVA** (Liu et al., 2023a) in three configurations: LLaVA 1.5 (Vicuna), LLaVA 1.6 (Llama), and LLaVA 1.6 (Mistral). Additionally, we evaluate LLaVA 1.5 RLAIF (Yu et al., 2024), a version of LLaVA 1.5 aligned through AI feedback.
- **InternVL2** (Chen et al., 2023)
- **MiniCPM V2** (Yao et al., 2024)
- **Phi 3.5 Vision** (Abdin et al., 2024)
- **GPT4o-mini**

See Appendix B for more information on each model.

### 4.3 Results

Table 3 shows the CAST results for similarity statements generated and evaluated across different modalities. We average the CAST score over the

Model	Gen w/	Eval w/ text	Eval w/ image	Eval w/ both
Bunny	text	<b>0.93</b>	0.76	0.96
	image	0.71	<b>0.80</b>	0.85
	both	0.91	0.81	<b>0.96</b>
GPT4o-Mini	text	<b>0.94</b>	0.73	0.94
	image	0.79	<b>0.90</b>	0.87
	both	0.91	0.76	<b>0.91</b>
InternVL2	text	<b>0.67</b>	0.66	0.72
	image	0.57	<b>0.78</b>	0.75
	both	0.68	0.73	<b>0.77</b>
MiniCPM V2	text	<b>0.92</b>	0.90	0.93
	image	0.50	<b>0.91</b>	0.73
	both	0.84	0.89	<b>0.91</b>
Phi-3.5-V	text	<b>0.61</b>	0.60	0.63
	image	0.50	<b>0.72</b>	0.61
	both	0.60	0.63	<b>0.64</b>
LLaVA 1.5 (Vicuna)	text	<b>0.91</b>	0.87	0.82
	image	0.68	<b>0.91</b>	0.69
	both	0.86	0.87	<b>0.78</b>
LLaVA 1.6 (Llama)	text	<b>0.73</b>	0.57	0.74
	image	0.56	<b>0.69</b>	0.64
	both	0.68	0.62	<b>0.73</b>
LLaVA 1.6 (Mistral)	text	<b>0.81</b>	0.85	0.87
	image	0.52	<b>0.86</b>	0.72
	both	0.74	0.86	<b>0.84</b>
LLaVA 1.5 RLAIF	text	<b>0.57</b>	0.66	0.49
	image	0.58	<b>0.93</b>	0.70
	both	0.55	0.76	<b>0.58</b>

Table 1: CAST self-consistency scores (Top-3) averaged over the first three statements generated for each modality configuration. Bold cells show the performance when the evaluation is in the same modality as the generation.

first the first three statements generated (Top-3). The results indicate that models perform best when statements are generated and evaluated within the same modality. There is a noticeable drop in consistency during cross-modal evaluations, where statements generated from images are evaluated using text descriptions and vice versa. With the exception of GPT4o-Mini, the combination of image-generated and text-evaluated statements leads to the worst consistency. This is somewhat expected as the similarity statement generated by the model might have been about something not mentioned in the text description (see Appendix D).

Qualitatively, we find that most inconsistencies arise during generation, where models often produce incorrect statements, particularly about object attributes or relationships. Notably, the image modality shows the highest hallucination rates, with models emphasizing prominent features without verifying their relevance to both scenes. This suggests that while object recognition is strong in state-of-the-art VLMs, accurately describing attributes and relations remains a challenge.

MiniCPM exhibits high consistency when evaluating with images. To test whether this is due to its RLAIF (Yu et al., 2024) fine-tuning stage, we evaluate a version of LLaVA-1.5 specially trained with RLAIF. Overall, we find LLaVA-1.5 RLAIF to be significantly less consistent than its base-model LLaVA-1.5. We therefore fail to conclude that

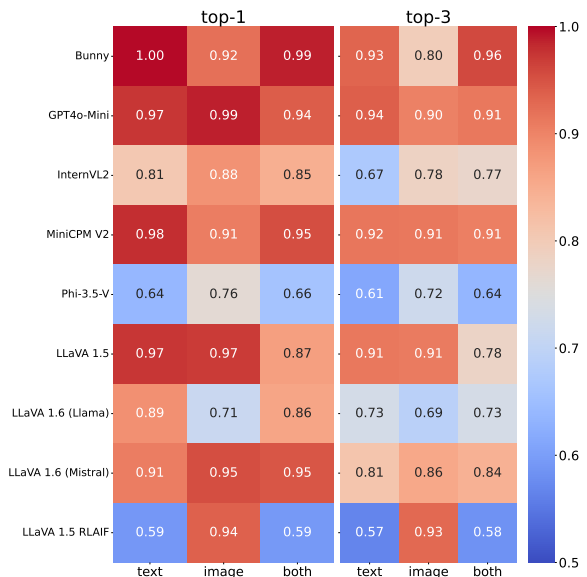


Figure 4: Average CAST self-consistency when multiple statements are generated and evaluated within the same modality. **Left:** Top-1 considers only the first statement generated. **Right:** Top-3 considers the first three statements generated, these are equivalent to the bolded results from Table 3.

RLAIF has a positive impact on consistency.

Figure 4 shows CAST scores for Top-1 and Top-3 generated statements. There is a slight decrease in CAST scores from Top-1 to Top-3, indicating that the quality of similarity statements typically declines with additional generations, as models become less reliable over longer generations.

We find GPT4o-Mini and MiniCPM to be the most consistent models overall. Both exhibit minimal drop with longer generations (9% for GPT4o-Mini and 6% for MiniCPM). In contrast, InternV2 and the LLaVA models experience a significant drop in consistency with additional generations. Overall, our single-modality CAST results highlight that VLMs fail to provide coherent and stable outputs as generations get longer.

Lastly, we can use CAST to evaluate how different modalities impact different VLMs. For instance, GPT4o-Mini and Bunny show a drop in image self-consistency when generation length increases, unlike MiniCPM and LLaVA-RLAIF which maintain more stability with generation length. Other models such as InterVL2 are more sensitive to the text modality.

## 5 Conclusion

We introduce CAST to evaluate the multi-modal self-consistency of VLMs by testing whether a

model applies consistent reasoning across text-only, image-only, or combined inputs. CAST uncovers cross-modal inconsistencies and goes beyond traditional accuracy metrics to assess the stability of a model’s logic across different modalities.

Our findings show that open-source VLMs still struggle with self-consistency across different modalities. CAST not only assesses self-consistency but also identifies modalities where the model may lack understanding. The strength of CAST lies in its lack of ground truth. The model’s self-consistency is evaluated only with respect to itself and not whether its generated statements are correct. As a result, CAST should be used in addition to traditional metrics, like accuracy, that capture the “correctness” capabilities of models. Additionally, because CAST does not rely on ground truth, it generalises well to different types of aligned inputs. It should therefore be easily extensible and applicable to a wide range of datasets and modalities.

CAST also provides a future direction for improving robustness in VLMs. For instance, using CAST during training, one could track self-consistency across modalities, which could provide insights into how VLMs align modalities over time. Using CAST as a task during instruction fine-tuning might also improve multi-modal alignment across different modalities.

Ultimately, given the method’s universality, CAST’s framework can be adapted to any domain or language dataset, provided there are sufficiently similar images and highly detailed descriptions.

## 6 Limitations

The main limitation is that our test does not guarantee the capability of a model. We make no claims about the correctness of the model, but focus solely on whether a model is self-consistent. This means a model that always predicts the similarity statement to match the scenes, regardless of the statement, would always be deemed consistent even though it would also likely be wrong. Our approach therefore needs to be taken in conjunction with the traditional evaluation methods. It is most useful for models trained and evaluated using standard correctness metrics.

Additionally, a potential limitation is that we only evaluate CAST on a sample of 100 selected pairs. However, we do not believe the sample size affects the validity of our benchmarking frame-

work, and we also release our code and sampling method to allow a greater evaluation set to be constructed. The primary objective of this paper is to introduce CAST as a flexible evaluation method rather than to establish a fixed dataset. To increase robustness, we also use three different prompts and assess each model across multiple modality combinations (text-only, image-only, and both). Increasing the sample size would also lead to a rise in inference costs without necessarily producing different insights.

Finally, there are also limitations with our VLM evaluations that follow directly from the brittle nature of these models. While we evaluated the generated statements using multiple prompts, we sample from each model using greedy sampling and therefore it is possible that some of our results are biased towards certain models. However, CAST could easily be expanded to include responses from different sampling mechanisms (temperature > 0) at the cost of increased computation.

## 7 Ethical Considerations

Our research relies on open-source and closed-source VLMs generating and evaluating text and image inputs and therefore carries the typical risks associated with open-ended text generation. The DOCCI dataset, which we sub-sample from, is licensed under the CC BY 4.0 license<sup>2</sup>. Overall, we hope that CAST leads to improvements in the trustworthiness and robustness of VLMs.

## Acknowledgments

The authors extend their gratitude to Amazon’s Development Centre Scotland (ADCS) for the challenge and AWS access to work with VLMs models of the Claude family. In particular, we wish to thank Christos Christodouloupoulos for his helpful feedback throughout this project. Olga Loginova also thanks Amazon Alexa for their support of her research through a grant.

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) at the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences and by the UKRI-funded TAS Governance Node (grant number EP/V026607/1).

<sup>2</sup><https://creativecommons.org/licenses/by/4.0/>

## References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. [INSIDE: LLMs’ internal states retain the power of hallucination detection](#). In *The Twelfth International Conference on Learning Representations*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, and Serena Yeung-Levy. 2023. [Describing differences in image sets with natural language](#). *ArXiv*, abs/2312.02974.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhishava Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). *ArXiv*, abs/2306.13394.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. [Blink: Multimodal large language models can see but not perceive](#). *ArXiv*, abs/2404.12390.
- Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yuezhe Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O’Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen Marcus McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2023. [Ai alignment: A comprehensive survey](#). *ArXiv*, abs/2310.19852.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefan O Soatto. 2022. [Masked vision and language modeling for multi-modal representation learning](#). *ArXiv*, abs/2208.02131.
- Qing Li, Chenyang Lyu, Jiahui Geng, Derui Zhu, Maxim Panov, and Fakhri Karray. 2024. [Reference-free hallucination detection for large vision-language models](#).
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. [Improved baselines with visual instruction tuning](#). *Preprint*, arXiv:2310.03744.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. [Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation](#). In *The Twelfth International Conference on Learning Representations*.
- Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. 2024. [Docci: Descriptions of connected and contrasting images](#). *arXiv preprint arXiv:2404.19753*.
- Letitia Parcalabescu and Anette Frank. 2024. [Do vision & language decoders use images and text equally? how self-consistent are their explanations?](#) *ArXiv*, abs/2404.18624.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. [Large language models sensitivity to the order of options in multiple-choice questions](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. [Learning transferable visual models from natural language](#)

- supervision. In *International Conference on Machine Learning*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations*.
- Qian Yang, Weixiang Yan, and Aishwarya Agrawal. 2024. [Decompose and compare consistency: Measuring vlms’ answer reliability via task-decomposition consistency comparison](#). *ArXiv*, abs/2407.07840.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. [Minicpm-v: A gpt-4v level mllm on your phone](#). *arXiv preprint 2408.01800*.
- Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. [Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness](#). *arXiv preprint arXiv:2405.17220*.
- Tongtian Yue, Jie Cheng, Longteng Guo, Xingyuan Dai, Zijia Zhao, Xingjian He, Gang Xiong, Yisheng Lv, and Jing Liu. 2024. [Sc-tune: Unleashing self-consistent referential comprehension in large vision language models](#). *arXiv preprint arXiv:2403.13263*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2023. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). *ArXiv*, abs/2311.16502.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.
- Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. 2023. [Vision-language models for vision tasks: A survey](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5625–5644.
- Yuan Zhang, Fei Xiao, Tao Huang, Chun-Kai Fan, Hongyuan Dong, Jiawen Li, Jiacong Wang, Kuan Cheng, Shanghang Zhang, and Haoyuan Guo. 2024. [Unveiling the tapestry of consistency in large vision-language models](#). *ArXiv*, abs/2405.14156.
- Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy M. Hospedales. 2024. [Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning](#). *ArXiv*, abs/2406.12742.

## A Prompts

### A.1 Generation

To generate a number of similarity statements, we use the prompt shown in Figure 5. We slightly modify the prompt to fit each modality input.

Given two `scenes` | `side-by-side images` | `scenes and their corresponding images`, find up to five similarities between each `scene`|`image`|`scene`. Output each similarity in a numbered list.

Figure 5: Generation Prompt: For each model and each of the three modalities, we generate a list of similarity statements using the above prompt.

### A.2 Evaluation

To reduce variance in our results and potential biases that might exist towards certain prompt phrasing (Pezeshkpour and Hruschka, 2024; Sclar et al., 2024), we opt to use three different evaluation prompts shown in Figure 6.

### A.3 Parsing the Evaluation Output

To parse the resulting evaluation from the model we use a simple post-processing step:

```
1 def parse_validator(x):
2     x = x.strip("*").lower().split("\n")
3     [0]
4     if x.startswith(positive):
5         return 1
6     elif x.startswith(negative):
7         return 0
8     else:
9         return None
```

Note that we ignore generations that we cannot parse from the evaluation score. This is typically rare for most models and prompts.

## B Additional model details

## C Additional Results

### C.1 Results for each Prompt types

### C.2 Results for position of generated statement

## D Information Flow in Image Similarity

Since the human annotators of DOCCI (Onoe et al., 2024) are given the image from which to write the description, from an information content, we can view the textual description of an image as a subset of the overall information content contained within the image.

1. Given two `scenes`|`side-by-side images`|`scenes and their corresponding images`, does the following statement apply to only one of the `scenes` | `images` | `scenes`? Answer with ‘one’ or ‘both’.
2. Given two `scenes`|`side-by-side images`|`scenes and their corresponding images`, is the following statement true for both of the `scenes`|`images`|`scenes`? Answer with ‘true’ or ‘false’ if the statement is untrue or only true for one of the `scenes`|`images`|`scenes`.
3. Given two `descriptions`|`side-by-side images`|`descriptions and their corresponding images`, does the following statement describe both of the `descriptions`|`images`|`descriptions`? Answer with ‘yes’ or ‘no’ if the statement is not applicable to one of the `descriptions`|`images`|`descriptions`.

Figure 6: Evaluation Prompts: For each model and each of the three modalities, we generate validate a similarity statement from the generation step. We use three different evaluation prompts to reduce potential bias of models towards a particular prompt format.

If denote information content as entropy  $H$ , then:

$$H(S^{img}) \geq H(S^{txt})$$

And since the text description should not be introducing new information, the union of both the Image and Description should be equal in entropy to that of the image:

$$H(S^{img}) = H(S^{img+txt})$$

Unfortunately, it is the case that text can introduces new information through subjective interpretation, and the obvious fact that a photograph can very rarely be fully described in language. However it might still be useful to model the annotation of images as conditional on the images and not independent. This might lead to further inter-modal consistency analysis which we leave open as direction for future work.



Model	Vision Encoder	LLM	Additional Design Choices
Bunny 1.1 (He et al., 2024)	SigLip-400M (Zhai et al., 2023)	Llama-3-8B Ins (Dubey et al., 2024)	-
MiniCPM V 2.5 (Yao et al., 2024)	SigLip-400M (Zhai et al., 2023)	Llama-3-8B Ins (Dubey et al., 2024)	Adaptive Visual Encoding and RLAIF-V
InternVL2 (Chen et al., 2023)	InternViT	InternLM 2.5 7B	-
Phi 3.5 Vision (Abdin et al., 2024)	CLIP ViT (Radford et al., 2021b)	phi-3-mini-128K-instruct LM	-
LLaVa-Next 1.5 (Liu et al., 2023b)	CLIP-ViT (Radford et al., 2021b)	Vicuna-7B	-
LLaVa-Next 1.6 (Liu et al., 2023b)	CLIP-ViT (Radford et al., 2021b)	Mistral-7B (Jiang et al., 2023)	-
LLaVa-Next 1.6 (Liu et al., 2023b)	CLIP-ViT (Radford et al., 2021b)	Llama-3-8B Ins (Dubey et al., 2024)	Image Slicing
LLaVa-Next 1.5 RLAIF (Yu et al., 2024)	CLIP-ViT (Radford et al., 2021b)	Vicuna-7B	Visual RLAIF Alignment

Table 2: Open-source VLMs tested along with a description of which Vision Encoder and LLM each model uses.

Model	Gen w/	Eval w/ text					Eval w/ image					Eval w/ both				
		yes/no	both/one	true/false	Avg.	Avg.	yes/no	both/one	true/false	Avg.	Avg.	yes/no	both/one	true/false	Avg.	
Bunny	text	0.97	0.99	0.84	0.93	0.73	0.88	0.66	0.76	0.76	0.99	0.96	0.93	0.96		
	image	0.72	0.86	0.55	0.71	0.75	0.90	0.74	0.80	0.80	0.85	0.89	0.80	0.85		
	both	0.94	0.98	0.80	0.91	0.77	0.92	0.74	0.81	0.98	0.96	0.94	0.96			
GPT4o-M	text	0.91	0.98	0.92	0.94	0.70	0.88	0.62	0.73	0.96	0.95	0.93	0.94			
	image	0.75	0.91	0.71	0.79	0.91	0.95	0.85	0.90	0.86	0.94	0.83	0.87			
	both	0.90	0.97	0.86	0.91	0.73	0.87	0.67	0.76	0.93	0.94	0.87	0.91			
InternVL2	text	0.80	0.36	0.86	0.67	0.86	0.17	0.94	0.66	0.90	0.32	0.95	0.72			
	image	0.65	0.34	0.71	0.57	0.98	0.39	0.99	0.78	0.88	0.42	0.95	0.75			
	both	0.85	0.35	0.85	0.68	0.96	0.25	0.99	0.73	0.94	0.40	0.98	0.77			
MiniCPM	text	0.96	0.79	0.99	0.92	0.96	0.78	0.97	0.90	0.98	0.83	0.99	0.93			
	image	0.48	0.40	0.61	0.50	0.96	0.82	0.96	0.91	0.73	0.61	0.84	0.73			
	both	0.89	0.69	0.95	0.84	0.93	0.78	0.95	0.89	0.96	0.78	0.98	0.91			
Phi-V	text	0.81	0.09	0.94	0.61	0.82	0.12	0.85	0.60	0.89	0.12	0.87	0.63			
	image	0.58	0.20	0.73	0.50	0.91	0.32	0.94	0.72	0.82	0.24	0.75	0.61			
	both	0.74	0.18	0.87	0.60	0.85	0.18	0.87	0.63	0.89	0.16	0.86	0.64			
LLaVA1.5	text	0.76	0.99	0.98	0.91	0.66	1.00	0.95	0.87	0.50	0.98	0.97	0.82			
	image	0.40	0.89	0.74	0.68	0.75	1.00	0.98	0.91	0.33	0.92	0.83	0.69			
	both	0.64	0.99	0.94	0.86	0.65	1.00	0.95	0.87	0.42	0.98	0.94	0.78			
LLaVA-1.6 (Llama)	text	0.69	0.99	0.51	0.73	0.48	1.00	0.24	0.57	0.71	0.99	0.52	0.74			
	image	0.42	0.93	0.33	0.56	0.72	0.99	0.36	0.69	0.54	0.96	0.40	0.64			
	both	0.65	0.98	0.43	0.68	0.58	0.99	0.29	0.62	0.71	0.99	0.50	0.73			
LLaVA1.6 (Mistral)	text	0.89	0.65	0.89	0.81	0.88	0.78	0.90	0.85	0.95	0.71	0.95	0.87			
	image	0.59	0.37	0.59	0.52	0.89	0.74	0.93	0.86	0.79	0.55	0.81	0.72			
	both	0.82	0.57	0.84	0.74	0.90	0.76	0.93	0.86	0.93	0.67	0.93	0.84			
LLaVA1.5 RLAIF	text	0.19	0.86	0.66	0.57	0.37	1.00	0.62	0.66	0.22	0.89	0.35	0.49			
	image	0.28	0.92	0.53	0.58	0.84	1.00	0.95	0.93	0.49	0.95	0.66	0.70			
	both	0.22	0.89	0.55	0.55	0.54	1.00	0.74	0.76	0.33	0.92	0.50	0.58			

Table 3: CAST self-consistency scores for the first three statements generated for each modality configuration.

## E Dataset Sub-sampling

As previously mentioned, we sample image pairs from the DOCCI (Onoe et al., 2024) train dataset (10k images), and reject pairs that do not exhibit a certain threshold of CLIP similarity. In particular, we use cosine-similarity between images to filter pairs which are either not similar enough ( $< 0.75$  CLIP score), or which contain near identical images or duplicates ( $\geq 0.95$  CLIP score). We decided on these boundaries through qualitative analysis of the DOCCI samples. Additionally, we filter pairs by description length to only select descriptions with at least 500 characters.

After sampling from our desired image CLIP similarity range, we plot our subset against the text CLIP similarity between each pairs of examples (shown in Figure 7).

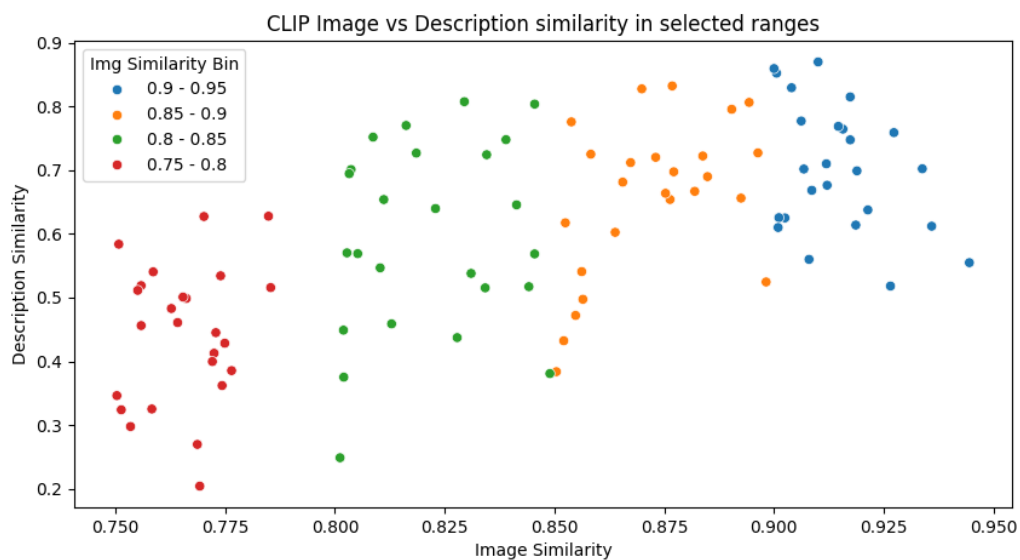
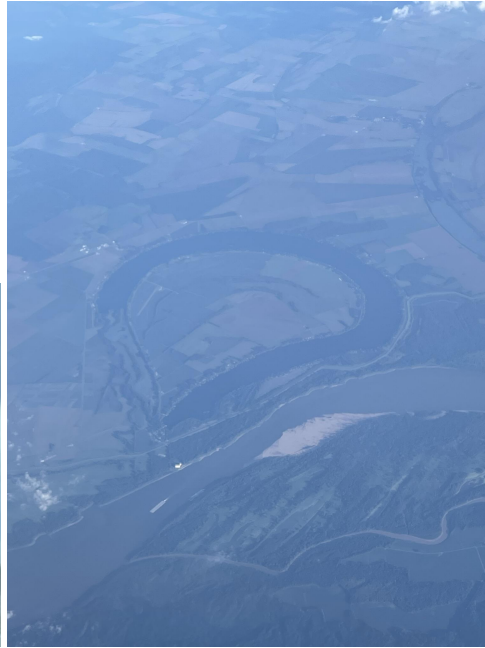


Figure 7: We plot the CLIP similarity between descriptions and images of the sampled example pairs. We find as expected that there is some positive correlation between the similarity of image pairs and the similarity of textual descriptions. However, we can also observe that some description similarity can be low even for images pairs which are predicted to be similar. This is because some of the descriptions are short and/or the annotators decided to focus on different aspects of the image.

<p>Outdoor medium shot view of the General W. K Wilson Jr. Bridge from a 3/4 view from behind the glass of a motor vehicle on the opposite road. There are droplets of water on the glass that are out of focus. There are dark gray rain clouds outside. A bridge arch over a portion of a highway road with suspension cords coming down from the left and right sides of the bridge where the beams cross horizontally on the arch. The archway resembles a silver ladder. Yellow reflective bumpers border the road on the lower right. A cement barrier rises between the arch and just before it.</p>	<p>An aerial view of a dark green and gray blue landscape with a river running through it. The image is low resolution and not in focus. The river is wide and runs from the bottom left corner to one third of the way up and out on the right edge of the frame. One tanker ship is traveling in the center of the river to the left and angled to the bottom left corner. A large sand bank bows out from the lower river bank as the river bends to the right. Below the lower river bank is a forested area with many thick trees. A tributary river feeds the main river from the right, and meanders down to the left. Above the far side of the river, a forest makes a large loop and fills the center of the frame. Farm land fills the top half of the frame beyond.</p>
---	---



Bunny 1.1	<ol style="list-style-type: none"> <li>Both images feature a bridge structure.</li> <li>The bridge in both images has a curved arch design.</li> <li>The color of the bridge is gray.</li> <li>The sky in both images is blue.</li> <li>The clouds in both images are white.</li> </ol>
LLaVA 1.5 (Vicuna)	<ol style="list-style-type: none"> <li>Both images feature a bridge.</li> <li>Both images show a body of water.</li> <li>Both images depict a cloudy sky.</li> <li>Both images have a similar color palette, with blues and grays dominating.</li> <li>Both images show a landscape with natural elements.</li> </ol>
MiniCPM V2.5	<ol style="list-style-type: none"> <li>Both images are taken during the daytime.</li> <li>They both feature man-made structures.</li> <li>The sky is visible in both images.</li> <li>There is a sense of motion or travel suggested by the road in the left image and the aerial view in the right image.</li> <li>Both images have a color palette dominated by shades of blue and gray.</li> </ol>
InternVL2	<ol style="list-style-type: none"> <li>Both images feature a bridge.</li> <li>Both images have a cloudy sky.</li> <li>Both images show a landscape with water bodies.</li> <li>Both images have a sense of depth and perspective.</li> <li>Both images have a similar color palette, with blues and grays dominating the scene.</li> </ol>
GPT4o-Mini	<p>Here are five similarities between the two images:</p> <ol style="list-style-type: none"> <li>Curved Structures: Both images feature prominent curved shapes; the left image shows an arch bridge, while the right image depicts a river with a winding curve.</li> <li>Natural Elements: Each image incorporates natural elements; the right image prominently features a river and land, while the left image is set against a cloudy sky.</li> <li>Man-Made vs. Natural: The left image represents a man-made structure (the bridge), while the right image showcases a natural landscape, highlighting the contrast between human engineering and nature.</li> <li>Perspective: Both images are</li> </ol>
Phi 3.5 Vision	<ol style="list-style-type: none"> <li>Both images feature man-made structures.</li> <li>The first image shows a bridge, while the second image shows a river with a bridge.</li> <li>Both images have a blue color palette.</li> <li>The first image has a cloudy sky, while the second image has a clear sky.</li> <li>Both images are taken from a ground perspective.</li> </ol>
LLaVA 1.6 (Llama)	<ol style="list-style-type: none"> <li>Both images show a bridge structure.</li> <li>The bridge in both images appears to be a suspension bridge.</li> <li>The bridge is located over a body of water.</li> <li>The sky in both images is partly cloudy.</li> <li>The bridge is the main focus of both images, with the water and sky serving as the background.</li> </ol>

Figure 8: Generated statements for each model when given image inputs

<p>An eye-level view of a tree trunk that has been ripped out of the ground laying on its side. The tree trunk is facing away from view, only the bottom and very top of the tree trunk is visible. The bottom of the tree trunk is hollow, there is a hole visible through the bottom that allows you to see a small sliver of the ground in the distance. The ground is sloped toward the bottom left corner of the image, it is a dirt surface that is covered mostly with gray discolored leaves and brown leaves scattered throughout the image. There are thin tree trunks and trees in the background behind the tree trunk.</p>	<p>An overhead view of a group of nine California pipevine swallowtail butterflies sitting on a dirt surface. The butterflies are all facing different directions. The front of their wings are dark blue and fade into a lighter shade of blue as they go back. There are white dots lining the edge of each butterfly's wings. There are two large gray rocks visible in the bottom right and bottom left corner of the image. A large concentration of dry leaves and sticks are covering the dirt surface at the top half of the image. There are sticks and dry leaves scattered more sparingly in the middle of the image where the butterflies are standing.</p>
---	---



Bunny 1.1	<ol style="list-style-type: none"> <li>Both images feature a fallen tree stump.</li> <li>The tree stump is surrounded by a mix of leaves and twigs.</li> <li>There are butterflies present in both images.</li> <li>The butterflies are black with yellow spots.</li> <li>The background of both images is a forest with green foliage.</li> </ol>
LLaVA 1.5 (Vicuna)	<ol style="list-style-type: none"> <li>Both images show a natural setting with a forest floor.</li> <li>There is a fallen tree trunk in both images.</li> <li>The fallen tree trunks are surrounded by leaves and debris.</li> <li>The ground appears to be damp and muddy.</li> <li>The overall color palette is earthy and muted.</li> </ol>
MiniCPM V2.5	<ol style="list-style-type: none"> <li>The presence of butterflies is a common element in both images.</li> <li>Both images depict a natural outdoor setting with trees and foliage.</li> <li>The ground in both images is covered with leaves and twigs, indicating a forest floor environment.</li> <li>The lighting in both images suggests they were taken during the daytime.</li> <li>The overall composition of both images focuses on the interaction between the butterflies and their natural habitat.</li> </ol>
InternVL2	<ol style="list-style-type: none"> <li>The presence of a large tree stump in the center of each image.</li> <li>The ground is covered with fallen leaves and branches.</li> <li>The images show a forested area with trees and foliage.</li> <li>The lighting in both images appears to be natural, suggesting they were taken during the day.</li> <li>The images have a similar color palette, with earthy tones dominating the scene.</li> </ol>
GPT4o-Mini	<p>Here are five similarities between the two images:</p> <ol style="list-style-type: none"> <li>Natural Setting: Both images are set in a natural environment, showcasing elements of the outdoors.</li> <li>Organic Materials: Each image features organic materials; the left image has a decaying tree stump, while the right image includes butterflies interacting with the ground.</li> <li>Vegetation: Both images contain vegetation, with leaves and plants visible in the surroundings.</li> <li>Earthy Tones: The color palette in both images includes earthy tones, such as browns, greens, and muted colors.</li> <li>Wildlife Presence: Each image depicts wildlife</li> </ol>
Phi 3.5 Vision	<p>Both images depict natural outdoor scenes.</p> <ol style="list-style-type: none"> <li>Each image contains organic debris, such as leaves and branches.</li> <li>The color palette is dominated by earthy tones like browns and greens.</li> <li>Both images have a ground-level perspective, looking upwards towards the sky.</li> <li>There is a sense of decay and natural processes at work in both images. = 0.000000000000000000000000</li> </ol>
LLaVA 1.6 (Llama)	<ol style="list-style-type: none"> <li>Both images show a natural outdoor setting with a focus on a fallen tree or stump.</li> <li>The ground appears to be covered with leaves and possibly moss, indicating a forested or wooded area.</li> <li>There are insects or small creatures on the ground near the tree stump in both images.</li> <li>The lighting suggests it might be daytime with natural sunlight.</li> <li>The overall composition of the images is similar, with the tree stump as the central point of interest.</li> </ol>

Figure 9: Generated statements for each model when given image inputs

<p>A group of four Tiger sharks under water in an aquarium, the sharks appear to be near some man made stones and a small school of fish to the bottom left corner. The sharks have grey skin with white pale underbellies, majority of the sharks are facing towards the left. In the center is a rock where there is a shark in front of the rock and one behind it, the shark to the front of the rock is facing to the right. The water is dark blue with a light shining to the left side of the photo and some reflections on the surface below.</p>	<p>A low-angle view of three sharks swimming in an aquarium among a large number of small gray fish scattered throughout the image. One of the sharks is on the top left side of the image swimming toward the top right corner of the image. There is another shark further away on the right side of the image facing the left side of the image. The front half of a shark is visible extending from the bottom right side of the image facing the left side of the image. There is a light blue hue throughout the image and the water in the distance fades into blue. Light from above the surface of the water is visible at the top of the image shining through the ripples on the surface of the water. The light is shining on the top shark and on the fish at the top of the image.</p>
--	--



Bunny 1.1	<ol style="list-style-type: none"> <li>Both scenes take place in an aquarium.</li> <li>Both scenes feature sharks as the main subjects.</li> <li>Both scenes have a school of fish present.</li> <li>Both scenes have a light source that is shining on the water surface.</li> <li>Both scenes have a dark blue hue to the water.</li> </ol>
LLaVA 1.5 (Vicuna)	<ol style="list-style-type: none"> <li>Both scenes feature sharks in an aquarium environment.</li> <li>Both scenes depict sharks swimming near rocks and fish.</li> <li>The sharks in both scenes have grey skin with white underbellies.</li> <li>The majority of the sharks in both scenes are facing towards the left.</li> <li>There is a light shining in the water in both scenes, creating reflections on the surface.</li> </ol>
MiniCPM V2.5	<ol style="list-style-type: none"> <li>Both scenes take place in an aquarium.</li> <li>Both scenes feature sharks swimming in the water.</li> <li>Both scenes have a school of fish present.</li> <li>Both scenes have a light source shining on the water surface.</li> <li>Both scenes have reflections on the water surface.</li> </ol>
InternVL2	<ol style="list-style-type: none"> <li>The sharks are in an aquarium.</li> <li>The sharks are swimming.</li> <li>The sharks are surrounded by fish.</li> <li>The water is blue.</li> <li>There is light shining on the sharks.</li> </ol>
GPT4o-Mini	<p>Here are five similarities between Scene 1 and Scene 2:</p> <ol style="list-style-type: none"> <li>Aquarium Setting: Both scenes take place in an aquarium environment, showcasing sharks swimming in water.</li> <li>Presence of Sharks: Each scene features multiple sharks; Scene 1 has four Tiger sharks, while Scene 2 has three sharks.</li> <li>Interaction with Fish: In both scenes, the sharks are surrounded by small fish. Scene 1 has a small school of fish in the bottom left corner, while Scene 2 has a large number of small gray fish scattered throughout the image.</li> <li>Lighting Effects: Both scenes exhibit</li> </ol>
Phi 3.5 Vision	<ol style="list-style-type: none"> <li>Both scenes feature sharks in an aquarium setting.</li> <li>Both scenes feature a group of sharks.</li> <li>Both scenes feature small fish in the background.</li> <li>Both scenes feature a light source shining on the sharks.</li> <li>Both scenes feature a dark blue color palette.</li> </ol>
LLaVA 1.6 (Llama)	<ol style="list-style-type: none"> <li>Both scenes take place in an aquarium.</li> <li>Both scenes feature sharks.</li> <li>Both scenes feature fish.</li> <li>Both scenes have a water setting.</li> <li>Both scenes have a light source shining on the water.</li> </ol>

Figure 10: Generated statements for each model when given text inputs

A close up view of a large glass sphere light fixture and a yellow light, hanging freely from a silver metal pole. To the right of the light pole, is a gold metal pole with a "M" next to it on the left. The light pole is creating a shadow behind it. There is a green and yellow wall behind the light fixtures, the green wall is on the left, and the yellow wall is on the right. There is a black wire and push bottom hanging behind the light fixture. To the left of the sphere light fixture, is a screen glass door leading to outside. Plants can be seen on the bottom left and right corners.

A low-angle shot of a creative light blue pendant lamp on a gray concrete ceiling. In the center is the lamp made with light blue rattan material, creating a woven spherical shape with a view of a white cylinder case with a bright white light in the center. The lamp is connected to a silver metal rod and a silver metal cylinder adjacent to the ceiling above. Behind the lamp is a gray concrete ceiling made with cement. The ceiling is vaulted diagonally from left to right. On the bottom of the frame is a gray tubing system of pipes connected to the ceiling.



Bunny 1.1	<ol style="list-style-type: none"> <li>Both scenes feature a light fixture hanging from a pole.</li> <li>Both scenes have a wall in the background.</li> <li>Both scenes have a shadow being cast by the light fixture.</li> <li>Both scenes have a wire hanging behind the light fixture.</li> <li>Both scenes have a door leading to outside.</li> </ol>
LLaVA 1.5 (Vicuna)	<ol style="list-style-type: none"> <li>Both scenes feature a light fixture hanging from a pole.</li> <li>Both scenes have a yellow light.</li> <li>Both scenes have a silver metal pole.</li> <li>Both scenes have a green and yellow wall.</li> <li>Both scenes have a black wire and push bottom hanging behind the light fixture.</li> </ol>
MiniCPM V2.5	<ol style="list-style-type: none"> <li>Both scenes have a light fixture as the main focus.</li> <li>Both scenes have a metal pole or rod supporting the light fixture.</li> <li>Both scenes have a shadow created by the light fixture.</li> <li>Both scenes have a wall in the background.</li> <li>Both scenes have a door or opening leading to the outside.</li> </ol>
InternVL2	<ol style="list-style-type: none"> <li>Both scenes have a light fixture.</li> <li>Both scenes have a metal pole.</li> <li>Both scenes have a shadow.</li> <li>Both scenes have a green wall.</li> <li>Both scenes have a black wire.</li> </ol>
GPT4o-Mini	<p>Here are the similarities between Scene 1 and Scene 2:</p> <ol style="list-style-type: none"> <li>Light Fixtures: Both scenes feature prominent light fixtures that are central to the composition—Scene 1 has a large glass sphere light fixture, while Scene 2 showcases a light blue pendant lamp.</li> <li>Metal Components: Each scene includes silver metal elements—Scene 1 has a silver metal pole, and Scene 2 has a silver metal rod and cylinder.</li> <li>Spherical Shape: Both light fixtures have a spherical design; Scene 1's glass sphere and Scene 2's woven spherical lamp.</li> <li>Background Elements:</li> </ol>
Phi 3.5 Vision	<ol style="list-style-type: none"> <li>Both scenes feature a light source hanging from a pole or rod.</li> <li>Both scenes have a metallic element in the form of a pole or rod.</li> <li>Both scenes have a wall in the background.</li> <li>Both scenes have a spherical element in the form of a light fixture or pendant lamp.</li> <li>Both scenes have a green element in the form of a plant or foliage.</li> </ol>
LLaVA 1.6 (Llama)	<ol style="list-style-type: none"> <li>Both scenes feature a light fixture hanging from a pole.</li> <li>Both scenes have a wall in the background.</li> <li>Both scenes have a shadow behind the light fixture.</li> <li>Both scenes have a wire and push button hanging behind the light fixture.</li> <li>Both scenes have plants in the bottom left and right corners.</li> </ol>

Figure 11: Generated statements for each model when given text inputs

<p>A Lamborghini showroom is seen from outside the window, which shows reflections of lights outside. A bright blue Lamborghini is seen in the front, with a silver convertible to the right and a white SUV in the background. The blue car in the front is seen from the front passenger's side, with a white note on the windshield and a large white tag hanging from the rearview mirror. Several small lights are reflected on this vehicle, while the silver next to it shows one large reflection over the front passenger fender. Behind the white SUV is a doorway leading to a lit office. Behind the other two vehicles is a large black window on the left, indicating nighttime, two small racks of t-shirts, and a set of stairs leading up to the right. The flooring in the showroom is made of large concrete slabs.</p>	<p>A front view of a red Lamborghini Aventador parked in the middle of a grey concrete show room floor. The Lamborghini has a black grill and the Lamborghini logo below the hood of the car. Reflections of light are on the hood of the Aventador and left headlight. A shadow of the Lamborghini encircles the front end of the car on the ground. A neon Lamborghini sign hangs on the wall in the background in the right corner behind a white Lamborghini car. A white couch and clothes hanging on a rack are behind the red Lamborghini on the left. A black staircase is behind the red Lamborghini in the background.</p>
--	--



Bunny 1.1	<ol style="list-style-type: none"> <li>1. Both scenes feature a Lamborghini showroom.</li> <li>2. The showrooms have a concrete floor.</li> <li>3. There are reflections of lights on the cars.</li> <li>4. There is a Lamborghini logo visible on the cars.</li> <li>5. There is a staircase in the background of both scenes.</li> </ol>
LLaVA 1.5 (Vicuna)	<ol style="list-style-type: none"> <li>1. Both scenes feature a Lamborghini showroom.</li> <li>2. Both scenes have a variety of Lamborghini cars on display.</li> <li>3. Both scenes have a mix of luxury and sports cars.</li> <li>4. Both scenes have a modern and sleek interior design.</li> <li>5. Both scenes have a focus on the cars, with the cars being the main subject.</li> </ol>
MiniCPM V2.5	<ol style="list-style-type: none"> <li>1. Both scenes feature Lamborghini vehicles prominently displayed in showrooms.</li> <li>2. The showrooms have a modern and sleek design with concrete flooring.</li> <li>3. Reflections of lights are visible on the vehicles in both scenes.</li> <li>4. There is a white Lamborghini in the background of both scenes.</li> <li>5. Both scenes include a black staircase in the background.</li> </ol>
InternVL2	<ol style="list-style-type: none"> <li>1. Both scenes feature a Lamborghini showroom.</li> <li>2. Both scenes have a concrete floor.</li> <li>3. Both scenes have a Lamborghini car prominently displayed.</li> <li>4. Both scenes have a staircase in the background.</li> <li>5. Both scenes have a large window indicating nighttime.</li> </ol>
GPT4o-Mini	<p>Here are five similarities between the two scenes:</p> <ol style="list-style-type: none"> <li>1. Showroom Environment: Both scenes take place in a Lamborghini showroom with a modern design and concrete flooring.</li> <li>2. Car Models: Each scene features a Lamborghini as the focal point, showcasing different models (a blue Lamborghini in Scene 1 and a red Lamborghini in Scene 2).</li> <li>3. Lighting and Reflections: Both images display reflections of lights on the cars, enhancing the visual appeal and highlighting the vehicles.</li> <li>4. Additional Vehicles: Each scene includes multiple vehicles; Scene 1 has a silver convertible and a white SUV, while Scene 2 features</li> </ol>
Phi 3.5 Vision	<ol style="list-style-type: none"> <li>1. Both scenes feature a Lamborghini car as the main subject.</li> <li>2. Both images show a showroom setting with a clean, modern interior.</li> <li>3. Each image has a car parked in the foreground with another car in the background.</li> <li>4. Both images have a reflective surface on the car, indicating a polished finish.</li> <li>5. The flooring in both images is made of large, flat concrete slabs.</li> </ol>
LLaVA 1.6 (Llama)	<ol style="list-style-type: none"> <li>1. Both scenes feature a Lamborghini showroom.</li> <li>2. Both scenes have a Lamborghini car as the central focus.</li> <li>3. Both scenes have a showroom floor with a concrete finish.</li> <li>4. Both scenes have a car with a large grille and the Lamborghini logo.</li> <li>5. Both scenes have reflections of light on the cars.</li> <li>6. Both scenes have a neon sign with the Lamborghini logo.</li> <li>7. Both scenes have a staircase in the background.</li> <li>8. Both scenes have a white couch in the background.</li> <li>9. Both scenes have a car with a large windshield.</li> <li>10. Both scenes have a car with a black headlight.</li> </ol>

Figure 12: Generated statements for each model when given both inputs

<p>Paper lanterns are seen hanging from a tent-type ceiling with a metal frame. Four of the lanterns are globe-shaped, while the fifth one that hangs behind the bottom left globe is a flattened shape. Both of the globe lanterns on the right are lit up. All but one of the globes have a symmetrical frame. The top right globe has a swirled frame and has two tears on the bottom of the paper. The right bottom lantern has a tear on the side, and the bottom left lantern has a tear on the bottom. The ceiling is black cloth with large metal frame beams. A small blue sign hangs from the metal beam on the left and can only partially be seen in the bottom corner. The sign is a circular shape and reads "MENUS" in white.</p>	<p>A very large light fixture is seen from a low angle, attached to wood beams on a wood plank ceiling. It has a large black circle base with numerous black cords of different lengths attached. The black cords hold iridescent glass globes with open bottoms. Each globe has a warm-toned round light bulb lit up inside. The ceiling is made of dark wood with small flush-mounted lights. Behind the large light fixture is another large light fixture that is slightly different, with a hanging hollow circular base that has round glass globes attaching to the base with warm light bulbs lit up. Two white brick pillars are seen in the distance in the bottom right of the image.</p>
--	--



Bunny 1.1	<ol style="list-style-type: none"> <li>1. Both scenes have a ceiling with a unique design.</li> <li>2. Both scenes have light fixtures that are the main focus.</li> <li>3. Both scenes have multiple light sources.</li> <li>4. Both scenes have a warm-toned light bulb.</li> <li>5. Both scenes have a circular shape in the light fixtures.</li> </ol>
LLaVA 1.5 (Vicuna)	<ol style="list-style-type: none"> <li>1. Both scenes feature hanging light fixtures.</li> <li>2. Both scenes have a combination of globe-shaped and flattened light fixtures.</li> <li>3. Both scenes have light bulbs lit up inside the fixtures.</li> <li>4. Both scenes have a symmetrical frame design for the globe-shaped light fixtures.</li> <li>5. Both scenes have a black ceiling or background.</li> </ol>
MiniCPM V2.5	<ol style="list-style-type: none"> <li>1. Both scenes feature hanging light fixtures with multiple globes.</li> <li>2. The globes in both scenes are illuminated.</li> <li>3. The ceiling in both scenes is dark, with the first being cloth and the second being wood.</li> <li>4. Both scenes have a circular base for the light fixtures.</li> <li>5. The globes in both scenes have an open bottom design.</li> </ol>
InternVL2	<ol style="list-style-type: none"> <li>1. Both scenes feature hanging light fixtures.</li> <li>2. Both scenes have a ceiling with a dark color.</li> <li>3. Both scenes have light bulbs lit inside the fixtures.</li> <li>4. Both scenes have a variety of shapes for the hanging elements.</li> <li>5. Both scenes have a sense of a cozy, warm atmosphere.</li> </ol>
GPT4o-Mini	<p>Here are five similarities between the two scenes:</p> <ol style="list-style-type: none"> <li>1. Hanging Light Fixtures: Both scenes feature light fixtures that are suspended from the ceiling, creating a focal point in the space.</li> <li>2. Globe Shapes: Each scene includes globe-shaped elements: Scene 1 has globe-shaped paper lanterns, while Scene 2 has iridescent glass globes.</li> <li>3. Illumination: Both scenes showcase light sources that are lit up, contributing to the ambiance of the environment.</li> <li>4. Ceiling Design: The ceilings in both scenes are designed with a dark color scheme—black cloth in Scene 1</li> </ol>
Phi 3.5 Vision	<ol style="list-style-type: none"> <li>1. Both scenes feature hanging light fixtures.</li> <li>2. The light fixtures in both scenes are made of glass globes.</li> <li>3. Both light fixtures have a warm-toned light bulb inside each globe.</li> <li>4. The light fixtures are suspended from a ceiling structure.</li> <li>5. Both light fixtures have a modern and artistic design.</li> </ol>
LLaVA 1.6 (Llama)	<ol style="list-style-type: none"> <li>1. Both scenes feature light fixtures with multiple glass globes.</li> <li>2. The glass globes in both scenes have light bulbs inside.</li> <li>3. Both scenes have a mix of round and non-round glass globes.</li> <li>4. Both scenes have a ceiling with a combination of wood and metal elements.</li> <li>5. Both scenes have a warm lighting effect from the bulbs inside the glass globes.</li> </ol>

Figure 13: Generated statements for each model when given both inputs