Edu Komputika 13 (1) (2026)

Edu Komputika Journal

https://journal.unnes.ac.id/journals/edukom/

Semantic Specificity Degradation in Zero-Shot Generative Vision-Language Models

Nahumi Nugrahaningsih✉ and Felicia Sylviana

Department of Informatics, Faculty of Engineering, Universitas Palangka Raya, Kampus UPR Jl. Yos Sudarso, Palangka Raya, Kalimantan Tengah, 73111, Indonesia

Article Info

Article History:
Received: 1 March 2026
Revised: 12 August 2026
Accepted: 31 August 2026

Keywords:
Hierarchical Semantics, Multimodal Evaluation, Semantic Specificity, Vision-Language Models, Zero-Shot Learning

Abstract

Vision-Language Models (VLMs) have shown strong zero-shot performance in generating free-form image descriptions. However, most evaluations focus on hallucinated content, while less attention is given to whether models preserve specific object identities. This study investigates whether zero-shot generative VLMs retain fine-grained semantic information or tend to produce broader category labels. We introduce the Generic Downgrade Rate (GDR), a metric that measures predictions that are correct at a general category level but incorrect at the specific class level. A controlled zero-shot experiment was conducted using BLIP-2 (OPT-2.7B) on a balanced subset of Food-101 containing 101 classes and 10 images per class (N = 1,010). Model outputs were evaluated using Fine Accuracy, Coarse Accuracy, and GDR. Fine Accuracy was 0.0099, while Coarse Accuracy reached 0.3069. The resulting GDR of 0.2970 indicates that many predictions captured the general food category but failed to preserve the specific class identity. Bootstrap confidence intervals showed that the difference between fine- and coarse-level accuracy was stable across resampled data. Overall, the results indicate a clear loss of semantic specificity in zero-shot generative outputs. By distinguishing general semantic correctness from exact class identification, this evaluation provides a more detailed view of how VLMs preserve different levels of semantic information.

✉ Correspondence Address:
Kampus UPR Jln. Yos Sudarso Palangka Raya
Kalimantan Tengah, 73111
E-mail: nahumi@it.upr.ac.id

ISSN 2252-6811
E-ISSN 2599-297X

INTRODUCTION

Vision-Language Models (VLMs) such as CLIP, BLIP-2, LLaVA, and MiniGPT-4 rapidly advance multimodal reasoning by aligning visual and textual representations within a shared embedding space. These models demonstrate strong zero-shot generalization and generate image descriptions without task-specific fine-tuning. In many applications, they operate in what is commonly called open-ended caption generation. In this setting, the model is not restricted to a predefined list of labels. Instead, it freely produces a natural language sentence describing the image content. The output can vary in wording and structure, as long as it reflects the visual input.

Recent empirical studies show that reliability problems still appear, especially in zero-shot settings. Two issues often reported are hallucination and misidentification (Guan et al., 2024; Jing et al., 2024; Wu et al., 2024). Hallucination happens when the model mentions objects or attributes that are not actually in the image. Misidentification happens when the object is present, but the model gives it the wrong name. The sentence may sound natural and fluent, but this does not mean the model has correctly identified what is truly in the image.

Large-scale benchmarks such as HallusionBench, AUTOHALLUSION, BEAF, and FAITHSCORE show that hallucination persists across architectures and prompting strategies (Guan et al., 2024; Jing et al., 2024; Wu et al., 2024; Ye-Bin et al., 2025). However, these evaluations primarily examine whether generated content is faithful to the visual input rather than whether a model preserves the semantic specificity of the object being described.

Beyond hallucination, recent literature notes that evaluation frameworks for VLMs are fragmented and often assess only limited dimensions of semantic quality (Qiu et al., 2024). Some recent metrics attempt to measure faithfulness, reasoning quality, and uncertainty (Farquhar et al., 2024). However, less attention is given to hierarchical semantic specificity. It remains unclear whether models preserve detailed object identity or shift toward broader category names during generation.

To clarify this distinction, we refer to two levels of semantic representation: coarse semantic and fine-grained semantic. Coarse semantics refers to general or higher-level categories in a concept hierarchy. At this level, the model recognizes the overall type of object but does not provide a specific identity. Fine-grained semantics refers to more precise and detailed labels that distinguish between closely related classes. For example, if an image depicts "gnocchi," a prediction such as "pasta" is correct at the coarse level but not at the fine-grained level. Only the prediction "gnocchi" preserves full categorical precision. This difference between general category recognition and specific identity preservation is central to our study.

The issue also appears in domain-specific settings. Studies in biomedical and specialized benchmarks show that VLMs may recognize general categories while failing to identify specific subtype (Marzullo et al., 2025; Patel et al., 2024). These findings further indicate that maintaining detailed object identity remains challenging for current multimodal models.

Recent studies propose mitigation strategies such as adjusting the model's internal signals during generation, asking the model to check its own answer before finalizing it, and comparing multiple possible outputs to choose the most consistent one (Lu et al., 2025; Su et al., 2025). Most of these studies measure simple outcomes, such as whether the answer is correct or whether it contains hallucinated content. They mainly focus on counting errors. Few studies examine semantic downgrade behavior, where the object in the image is correctly recognized but described using a more general name instead of the exact label.

In open-ended caption generation, because the model produces free-form natural language rather than selecting from fixed labels, semantic abstraction can occur naturally during wording. A model may correctly recognize that an image depicts a food-related object, yet replace the exact dish name with a broader term such as "cake" or "curry." This output is semantically plausible and does not introduce nonexistent objects. However, it reduces specificity. We refer to this phenomenon as semantic specificity degradation. It differs from classical hallucination because the model does not invent new objects. Instead, it abstracts upward within the semantic hierarchy.

Traditional caption metrics such as BLEU and CIDEr (Papineni et al., 2001; Vedantam et al., 2015) were not designed to tell the difference between naming an object exactly and describing it with a broader term. They mainly look at how many words overlap between the generated sentence and the reference. But similar words do not always mean the same level of detail. A general term can still sound correct while losing important specificity. For example, calling "gnocchi" simply "pasta" keeps the broad category, but the exact identity is no longer preserved. FAITHSCORE evaluates hallucination by assessing the faithfulness of generated content to visual input, while VALOR-

EVAL evaluates coverage and faithfulness across multiple dimensions (Qiu et al., 2024). However, neither explicitly measures cases in which a prediction remains correct at a broader semantic level while losing its fine-grained identity. GDR is designed to directly measure this semantic specificity downgrade.

For this reason, this study focuses on the degradation of semantic specificity in zero-shot generative Vision-Language Models. We evaluate BLIP-2 (OPT-2.7B) on a balanced subset of Food-101 to see whether the model keeps the exact class name or instead shifts its predictions toward broader semantic categories. To make this analysis measurable, we introduce the Generic Downgrade Rate (GDR). This metric captures how often the model produces an answer that is correct in general but incorrect at the specific class level. We report GDR together with Fine Accuracy (FA), which counts exact class matches, and Coarse Accuracy (CA), which reflects correct recognition at a broader category level. Taken together, these three metrics allow us to clearly quantify how often detailed object identity is reduced to a more general label. By separating semantic downgrade behavior from hallucination, this study offers a more focused way to evaluate multimodal reliability. The results demonstrate that degradation of semantic specificity can be quantitatively measured in zero-shot generative settings.

RESEARCH METHODS

A. Research Design

This study followed a quantitative experimental design to examine semantic specificity in zero-shot generative Vision-Language Models (VLMs). Our aim was analytical, not to improve performance. We wanted to see whether the model kept the exact object label or instead replaced it with a broader category when producing free-form descriptions.

Unlike supervised classification studies that focus on fine-tuning to boost accuracy, we evaluated the model in its original pretrained form. We did not perform task-specific adaptation, modify parameters, or add further training. The model was tested exactly as it was released.

This choice was deliberate. By keeping the model unchanged, the prediction patterns reflected its internal representation and natural generation behavior. In simple terms, the results show how the model behaves in a zero-shot setting, not how it performs after being optimized for a specific dataset.

B. Dataset

We conducted the experiment using a balanced subset of the Food-101 dataset (Bossard et al., 2014). Food-101 contains 101 food categories at a fine-grained level. Many classes look very similar, and the differences between them are often small. This makes the task difficult. For example, several dishes share similar ingredients, colors, or presentation styles. Because of this, the dataset is suitable for testing whether a model can keep detailed object identity instead of relying only on general visual patterns.

For the evaluation, we randomly selected ten images from each of the 101 classes, resulting in 1,010 images. This sample size was chosen to keep the evaluation computationally manageable while maintaining equal representation across all classes. Because each class contributed equally, the evaluation metrics were not influenced by class frequency. As a result, Fine Accuracy, Coarse Accuracy, and Generic Downgrade Rate reflected how the model made its predictions, not how the data were distributed across classes.

All images were processed using the standard preprocessing pipeline provided with BLIP-2. The pipeline included standard resizing and normalization steps that match the original pretrained setup of the model. We did not add any extra preprocessing. This ensured that the inputs followed the same conditions used during pretraining.

C. Zero-Shot Inference Procedure

All images were evaluated in a strict zero-shot setting using BLIP-2 (OPT-2.7B) with its original pretrained weights. No fine-tuning, gradient updates, or parameter changes were applied. Using a single model provided a controlled setting for this initial evaluation, although it limits the generalizability of the findings across different VLM architectures.

The same prompt was used for every image: "Describe the main food dish in the image. Provide its name if you can." The prompt was kept unchanged to avoid variation caused by different wording. It was designed to direct the model toward identifying the main dish rather than generating a long, open-ended description. This made it possible to examine whether the generated output preserved the specific class identity or shifted toward a broader food category. Because the prompt is more focused than those typically used for open-ended image captioning, the findings should be interpreted within this controlled prompting setting.

We used deterministic decoding to reduce variation in the generated outputs. At each decoding step, the model selected the most

probable next token, and the output length was limited to avoid unnecessarily long responses. Sampling methods such as nucleus and top-k sampling were not used. Each image was processed independently, and the generated text was stored in its original form for subsequent evaluation.

D. Output Normalization

Before evaluation, all generated texts were normalized to reduce minor differences in formatting and word form. Each prediction was converted to lowercase, punctuation was removed, and extra spaces were standardized. Simple singular–plural variations were also treated as equivalent when they referred to the same dish, such as "burger" and "burgers." These steps ensured that fine-level evaluation reflected differences in semantic content rather than minor variations in formatting or word form.

E. Hierarchical Semantic Mapping

To distinguish exact class identification from broader but semantically related predictions, we constructed a hierarchical semantic mapping. Each of the 101 Food-101 classes was mapped to a broader food category based on its general semantic meaning. For example, "gnocchi," "dumplings," and "ravioli" were mapped to the broader category "pasta," while "steak" was mapped to "meat". This structure allowed us to evaluate two levels of correctness. We could measure whether the model predicted the exact class name. At the same time, we could assess whether it at least identified the correct general type of food.

The coarse-level categories were generated automatically using WordNet. For each fine-grained class name, we searched for its noun entry. If the full phrase was not found, we searched using the main word in the phrase. If no suitable entry was available, we assigned a default coarse label, "dish."

When a valid entry was found, we moved upward in the semantic hierarchy by selecting a hypernym, typically two levels above the original term. To avoid overly general categories such as "entity" or "object," we applied a simple stop list. If the selected term was too broad, we chose a closer hypernym instead. When no appropriate broader category could be identified, we again assigned "dish". This procedure produced a single coarse category for each fine-grained label, stored as a dictionary. Using this mapping, each prediction was classified into one of three mutually exclusive outcomes: fine-correct, coarse-correct but not fine-correct, or incorrect. Before evaluation, the mapping was checked to ensure that all 101 Food-101 classes were included and that each class was assigned to exactly one coarse category. The complete mapping used in the experiment is provided in Supplementary Table S1.

F. Evaluation Metrics

Fine Accuracy (FA) measured the proportion of predictions that matched the fine-grained ground-truth label after normalization. A prediction was considered correct only when the generated dish name matched the reference class label. FA was calculated as:

$$FA = \frac{N_{fine}}{N} \quad (1)$$

where (N) is the total number of evaluated samples and N_{fine} is the number of predictions that were correct at the fine-grained level.

Coarse Accuracy (CA) measured correctness at a broader semantic level. A prediction was considered correct if it matched either the exact class label or its corresponding coarse category defined in the hierarchical mapping. CA was calculated as:

$$CA = \frac{N_{fine}+N_{coarse}}{N} \quad (2)$$

where N_{coarse} is the number of predictions that were correct at the coarse level but not at the fine-grained level.

The generic Downgrade Rate (GDR) measures how often the model produced a correct coarse-level prediction without preserving the fine-grained class identity. GDR was calculated as:

$$GDR = \frac{N_{coarse}}{N} \quad (3)$$

Together, FA, CA, and GDR evaluated model predictions at different levels of semantic specificity. FA measured exact class identification, CA measured correctness at the broader category level, and GDR measured how often a correct broader category was produced without preserving the specific class identity.

G. Statistical Stability Analysis

To assess the stability of the reported metrics, we applied non-parametric bootstrap resampling at the image level with 1,000 iterations. In each iteration, (N) samples were drawn with replacement from the full set of predictions, and Fine Accuracy (FA), Coarse Accuracy (CA), and Generic Downgrade Rate (GDR) were recalculated. This produced an empirical distribution for each metric. The 95% confidence intervals were calculated using the percentile method. Bootstrap resampling was used because it provides uncertainty estimates

without requiring a normality assumption, making it suitable for the proportion-based metrics used in this study.

H. Reproducibility

All experiments were conducted in Google Colab using Python 3.12 and an NVIDIA T4 GPU. BLIP-2 (OPT-2.7B), using the Salesforce/blip2-opt-2.7b checkpoint, was run with the Hugging Face Transformers library using half-precision inference and automatic device mapping. The environment included pandas 2.2.2 and Pillow 11.3.0. Other dependencies, including PyTorch, torchvision, Transformers, Accelerate, bitsandbytes, NLTK, tqdm, SciPy, and NumPy, were used from the default Google Colab environment without explicitly pinned versions.

The Food-101 subset was randomly sampled using a fixed seed of 42, and WordNet processing was performed using NLTK. Bootstrap resampling used 1,000 iterations with the same fixed seed. No fine-tuning, hyperparameter search, or model modification was performed. The same model, prompt, preprocessing procedure, and evaluation settings were applied to all images.

RESULT AND DISCUSSION

A. Empirical Result

As shown in Table 1, Fine Accuracy (FA) was 0.0099, while Coarse Accuracy (CA) reached 0.3069. This large difference shows that the model rarely identified the exact fine-grained class, although it more often identified the correct broader category. The Generic Downgrade Rate (GDR = 0.2970) further indicates that most correct predictions at the broader level did not preserve the specific class identity.

This result shows a consistent shift from specific labels to more general terms. The model often recognized the overall type of dish but replaced the exact name with a broader category. In these cases, the prediction remained semantically related to the image but did not preserve precise category detail. Figure 1 illustrates this upward movement from fine-grained labels to broader hypernyms.

Table 1. Zero-Shot Performance on Food-101 Subset (N = 1010)

Metric	Value	95% CI
Fine Accuracy (FA)	0.0099	0.0040–0.0158
Coarse Accuracy (CA)	0.3069	0.2792–0.3366
Generic Downgrade Rate (GDR)	0.2970	0.2703–0.3248

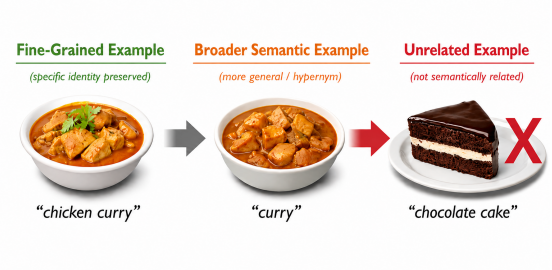

Figure 1. Conceptual illustration of semantic downgrade hierarchy from fine-grained label to broader hypernym category

To provide a qualitative illustration of semantic specificity changes, a small supplementary inference was performed on selected images using the question-answer prompt "Question: what food is shown in the image? Answer:". These examples are provided for illustration only and are not used in the calculation of FA, CA, or GDR.

Table 2. Supplementary qualitative examples of BLIP-2 predictions under a question-answer prompt

Food-101 label	BLIP-2 output
Prime rib	Roast beef
Spaghetti Bolognese	Spaghetti
Lobster bisque	Soup

The difference between FA and CA is informative. If visual grounding had failed, both metrics would likely have been similarly low. Instead, CA remained substantially higher, suggesting that the model usually captured the correct semantic family but did not retrieve or generate the exact label required for precise identification.

Bootstrap resampling with 1,000 iterations showed that the results were stable across resampled data. The 95% confidence intervals were narrow and remained close to the reported values. The clear separation between the confidence intervals for Fine Accuracy and Coarse Accuracy suggests that the observed difference was not driven by sampling variability alone.

B. Mechanisms and Interpretation

From a representational perspective, transformer-based and vision–language models organize concepts within structured embedding spaces. These spaces can show hierarchical patterns, with related concepts located near one another (Coenen et al., 2019). Similar patterns have been observed in image–text representations, where broader categories may form more stable regions in the embedding space

(Alper & Averbuch-Elor, 2024). This structure may make general concepts easier to access during generation than more specific ones. Cross-modal attention also supports alignment between visual features and textual tokens (Ilinykh & Dobnik, 2022). While such alignment helps connect visual and textual information, it may not always preserve fine-grained distinctions.

Representation geometry may also contribute to this pattern. During training, deep networks can develop increasingly structured class representations (Hong & Ling, 2024; Jiang et al., 2024; Súkeník et al., 2023). This structure can support class separation, but it does not necessarily preserve small differences between closely related categories. As a result, fine-grained distinctions may be more difficult to maintain.

Studies on hierarchy-aware training further suggest that preserving detailed category boundaries may require explicit modeling of semantic hierarchies (Liang & Davis, 2023). Without such modeling, broader semantic groupings may be easier to preserve than fine-level distinctions.

Embedding anisotropy provides another possible explanation. In anisotropic spaces, representations are unevenly distributed and tend to align along dominant directions (Rajaee & Pilehvar, 2021; Zhou & Srikumar, 2022). Such concentration may reduce the separation of some specific concepts and make broader semantic information easier to retrieve. However, the present study did not directly examine the internal embedding geometry of BLIP-2. Anisotropy should therefore be considered a possible explanation rather than a mechanism demonstrated by our results.

Decoding behavior may further contribute to the observed pattern. During generation, token selection is influenced by the probability assigned to candidate words, which can favor high-probability choices over more specific alternatives (Josifoski et al., 2023). Models may also rely on common lexical patterns learned during training (Jiang et al., 2024). Consequently, a model may generate a familiar general term instead of a more specific dish name. The output can therefore remain semantically relevant while losing some of its specificity. This interpretation is consistent with the large difference between Fine Accuracy and Coarse Accuracy observed in this study.

This pattern should be distinguished from object hallucination (Li et al., 2023; Rohrbach et al., 2018). Hallucination occurs when a model generates an object that is not supported by the image, reflecting a failure of visual grounding. Recent studies have examined this problem in terms of grounding errors and possible causal mechanisms in multimodal systems (Huang et al., 2025; Shang et al., 2025). The pattern observed in our study is different. Many predictions remained correct at the broader semantic level but failed to preserve the specific class identity. The main issue was therefore reduced semantic specificity rather than the generation of an unrelated object.

C. Model Evaluation and Practical Implications

Traditional captioning metrics such as BLEU (Papineni et al., 2001) and CIDEr (Vedantam et al., 2015) mainly evaluate lexical similarity between generated and reference texts. They do not directly distinguish between exact class identification and correctness at a broader semantic level. For example, a prediction such as "curry" instead of "beef rendang" may be semantically related to the reference, but the specific dish identity is lost. This distinction becomes important when precise category identification is required. VALOR-EVAL provides a broader assessment of coverage and faithfulness in VLM outputs (Qiu et al., 2024), but it does not directly measure whether a prediction shifts from a fine-grained label to a broader semantic category. GDR addresses this specific aspect by measuring the loss of semantic specificity.

Previous domain-specific evaluations have shown that model robustness can vary across different contexts (Gilal et al., 2025; Koddenbrock et al., 2025). More broadly, studies of robustness and uncertainty have also shown the limitations of relying on a single performance measure to describe model behavior (Manvi et al., 2024; Wenzel et al., 2020). These findings support the use of multiple evaluation measures that capture different aspects of model performance. In zero-shot generative settings, one aspect that remains less explored is whether incorrect predictions retain the correct broader meaning or shift to an unrelated concept. Measuring semantic specificity provides additional information about this behavior.

This distinction also has practical implications. In applications that require exact category labels, a broader but semantically related prediction may still be insufficient. An e-commerce system, for example, may require a specific product category rather than a general product type. Similarly, some medical imaging tasks may require identification of a specific condition rather than a broad diagnostic category. In such settings, loss of specificity can affect retrieval, indexing, or structured

annotation. Systems that require strict taxonomic alignment may therefore need additional validation or post-processing to preserve fine-grained accuracy.

In our experiment, the substantially higher Coarse Accuracy compared with Fine Accuracy shows that BLIP-2 often retained broader semantic information even when it failed to identify the exact Food-101 class. The high GDR captures this difference directly. Rather than treating all incorrect predictions in the same way, the proposed evaluation distinguishes predictions that preserve the broader category from those that fail at both levels. This provides a more detailed view of semantic errors in zero-shot generative output and complements existing approaches that focus on lexical similarity, general correctness, or hallucination.

D. Limitation

Several limitations should be considered. First, the evaluation was limited to a balanced subset of Food-101, so the findings may not generalize to domains with different visual and semantic characteristics. Second, only one model, BLIP-2 (OPT-2.7B), was evaluated. Comparisons across different VLMs are needed to determine whether semantic downgrade is specific to this model or occurs more broadly in zero-shot generative VLMs. Third, the hierarchical mapping was based on WordNet hypernyms and substring matching. This simple approach may miss some relationships between the generated text and the reference labels. Finally, we used only one English prompt and kept the generation settings fixed throughout the experiment. Different prompts, languages, or decoding settings may produce different levels of semantic specificity. Future work should evaluate additional VLMs, prompting strategies, and WordNet hierarchy depths to examine whether GDR remains consistent across different evaluation settings.

CONCLUSION

This study examined semantic specificity in zero-shot VLM outputs using a balanced subset of Food-101. By separating fine-grained correctness from broader semantic correctness, the evaluation revealed a clear and statistically stable gap between Fine Accuracy and Coarse Accuracy. The high GDR shows that many BLIP-2 outputs were not completely unrelated to the Food-101 labels. Instead, they were correct at the broader category level but failed to identify the specific dish. This distinction would be missed if evaluation considered only whether the generated output was generally semantically correct. Reporting fine- and coarse-level performance separately therefore helps show where semantic information is lost, particularly when the exact class label is required. Future work should evaluate other VLM architectures and datasets, test different prompting strategies and hierarchy levels, and examine whether the same pattern occurs across different domains.

REFERENCES

Alper, M., & Averbuch-Elor, H. (2024). Emergent Visual-Semantic Hierarchies in Image-Text Representations. Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LII, 220–238. https://doi.org/10.1007/978-3-031-72943-0_13

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101 – Mining Discriminative Components with Random Forests. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), Computer Vision – ECCV 2014 (pp. 446–461). Springer International Publishing.

Coenen, A., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and Measuring the Geometry of BERT. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). https://papers.nips.cc/paper/9065-visualizing-and-measuring-the-geometry-of-bert

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Gilal, N. U., Zegour, R., Al-Thelaya, K., Özer, E., Agus, M., Schneider, J., & Boughorbel, S. (2025). PathVLM-Eval: Evaluation of open vision language models in histopathology. Journal of Pathology Informatics, 18, 100455. https://doi.org/10.1016/j.jpi.2025.100455

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2024). Hallusionbench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large

Vision-Language Models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14375–14385. https://doi.org/10.1109/CVPR52733.2024.01363

Hong, W., & Ling, S. (2024). Neural Collapse for Unconstrained Feature Model under Cross-entropy Loss with Imbalanced Data. Journal of Machine Learning Research, 25(192), 1–48.

Huang, P.-H., Li, J.-L., Chen, C.-P., Chang, M.-C., & Chen, W.-C. (2025). Who Brings the Frisbee: Probing Hidden Hallucination Factors in Large Vision-Language Model via Causality Analysis. 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 6125–6135. https://doi.org/10.1109/WACV61041.2025.00597

Ilinykh, N., & Dobnik, S. (2022). Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022 (pp. 4062–4073). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.320

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., Taylor, C. J., & Roth, D. (2024). A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 4722–4756). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.272

Jing, L., Li, R., Chen, Y., & Du, X. (2024). FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 5042–5063). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.290

Josifoski, M., Peyrard, M., Rajič, F., Wei, J., Paul, D., Hartmann, V., Patra, B., Chaudhary, V., Kiciman, E., & Faltings, B. (2023). Language Model Decoding as Likelihood–Utility Alignment. In A. Vlachos & I. Augenstein (Eds.), Findings of the Association for Computational Linguistics: EACL 2023 (pp. 1455–1470). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.107

Koddenbrock, M., Hoffmann, R., Brodmann, D., & Rodner, E. (2025). On the Domain Robustness of Contrastive Vision-Language Models. KI 2025: Advances in Artificial Intelligence: 48th German Conference on AI, Potsdam, Germany, September 16–19, 2025, Proceedings, 62–76. https://doi.org/10.1007/978-3-032-02813-6_5

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., & Wen, J.-R. (2023). Evaluating Object Hallucination in Large Vision-Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 292–305. https://doi.org/10.18653/v1/2023.emnlp-main.20

Liang, T., & Davis, J. (2023). Inducing Neural Collapse to a Fixed Hierarchy-Aware Frame for Reducing Mistake Severity. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 1443–1452. https://doi.org/10.1109/ICCV51070.2023.00139

Lu, Y., Zhang, Z., Yuan, C., Gao, J., Zhang, C., Qi, X., Li, B., & Hu, W. (2025). Mitigating Hallucinations in Large Vision-Language Models by Self-Injecting Hallucinations. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025 (pp. 13861–13877). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.746

Manvi, R., Khanna, S., Burke, M., Lobell, D. B., & Ermon, S. (2024). Large Language Models are Geographically Biased. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 34654–34669). PMLR. https://proceedings.mlr.press/v235/manvi24a.html

Marzullo, A., Cappa, N., Morellini, M., & Ranzini, M. B. M. (2025). Generalist

Models in Specialized Domains: Evaluating Contrastive Language-image Pre-training for Zero-shot Anomaly Detection in Brain MRI. Journal of Medical Systems, 49(1), 156. https://doi.org/10.1007/s10916-025-02272-2

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 311. https://doi.org/10.3115/1073083.1073135

Patel, T., El-Sayed, H., & Sarker, M. K. (2024). Evaluating Vision-Language Models for hematology image Classification: Performance Analysis of CLIP and its Biomedical AI Variants. 2024 36th Conference of Open Innovations Association (FRUCT), 578–584. https://doi.org/10.23919/FRUCT64283.2024.10749850

Qiu, H., Hu, W., Dou, Z.-Y., & Peng, N. (2024). VALOR-EVAL: Holistic Coverage and Faithfulness Evaluation of Large Vision-Language Models. Findings of the Association for Computational Linguistics ACL 2024, 1783–1805. https://doi.org/10.18653/v1/2024.findings-acl.105

Rajaee, S., & Pilehvar, M. T. (2021). How Does Fine-tuning Affect the Geometry of Embedding Space: A Case Study on Isotropy. Findings of the Association for Computational Linguistics: EMNLP 2021, 3042–3049. https://doi.org/10.18653/v1/2021.findings-emnlp.261

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object Hallucination in Image Captioning. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4035–4045). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1437

Shang, Y., Zeng, X., Zhu, Y., Yang, X., Fang, Z., Zhang, J., Chen, J., Liu, Z., & Tian, Y. (2025). From Pixels to Tokens: Revisiting Object Hallucinations in Large Vision-Language Models. Proceedings of the 33rd ACM International Conference on Multimedia, MM '25, 10496–10505. https://doi.org/10.1145/3746027.3755728

Su, J., Chen, J., Li, H., Chen, Y., Qing, L., & Zhang, Z. (2025). Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 12964–12974. https://doi.org/10.18653/v1/2025.acl-long.634

Súkeník, P., Mondelli, M., & Lampert, C. H. (2023). Deep neural collapse is provably optimal for the deep unconstrained features model. Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23. The 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA.

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

Wenzel, F., Snoek, J., Tran, D., & Jenatton, R. (2020). Hyperparameter ensembles for robustness and uncertainty quantification. Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20.

Wu, X., Guan, T., Li, D., Huang, S., Liu, X., Wang, X., Xian, R., Shrivastava, A., Huang, F., Boyd-Graber, J. L., Zhou, T., & Manocha, D. (2024). AutoHallusion: Automatic Generation of Hallucination Benchmarks for Vision-Language Models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 8395–8419). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.493

Ye-Bin, M., Hyeon-Woo, N., Choi, W., & Oh, T.-H. (2025). BEAF: Observing BEfore-AFter Changes to Evaluate Hallucination in Vision-Language Models. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), Computer Vision – ECCV 2024 (pp. 232–248). Springer Nature Switzerland.

Zhou, Y., & Srikumar, V. (2022). A Closer Look at How Fine-tuning Changes BERT. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1046–1061). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.75