ODE: Open-Set Evaluation of Hallucinations in Multimodal Large Language Models

Yahan Tu   Rui Hu   Jitao Sang
Beijing Jiaotong University
Beijing, China
{yahan.tu, rui.hu, jtsang}@bjtu.edu.cn

Abstract

Hallucination poses a persistent challenge for multimodal large language models (MLLMs). However, existing benchmarks for evaluating hallucinations are generally static, which may overlook the potential risk of data contamination. To address this issue, we propose ODE, an open-set, dynamic protocol designed to evaluate object hallucinations in MLLMs at both the existence and attribute levels. ODE employs a graph-based structure to represent real-world object concepts, their attributes, and the distributional associations between them. This structure facilitates the extraction of concept combinations based on diverse distributional criteria, generating varied samples for structured queries that evaluate hallucinations in both generative and discriminative tasks. Through the generation of new samples, dynamic concept combinations, and varied distribution frequencies, ODE mitigates the risk of data contamination and broadens the scope of evaluation. This protocol is applicable to both general and specialized scenarios, including those with limited data. Experimental results demonstrate the effectiveness of our protocol, revealing that MLLMs exhibit higher hallucination rates when evaluated with ODE-generated samples, which indicates potential data contamination. Furthermore, these generated samples aid in analyzing hallucination patterns and fine-tuning models, offering an effective approach to mitigating hallucinations in MLLMs. Our code are available at https://github.com/Iridescent-y/ODE.

1. Introduction

Multimodal Large Language Models (MLLMs) [1, 3, 16, 30, 33, 36] have rapidly advanced in recent times, enabling detailed image descriptions (i.e., image captioning) and responses to image-related queries (i.e., visual question answering). However, these models face the persistent challenge of "hallucination" [8, 15, 27], where generated re-
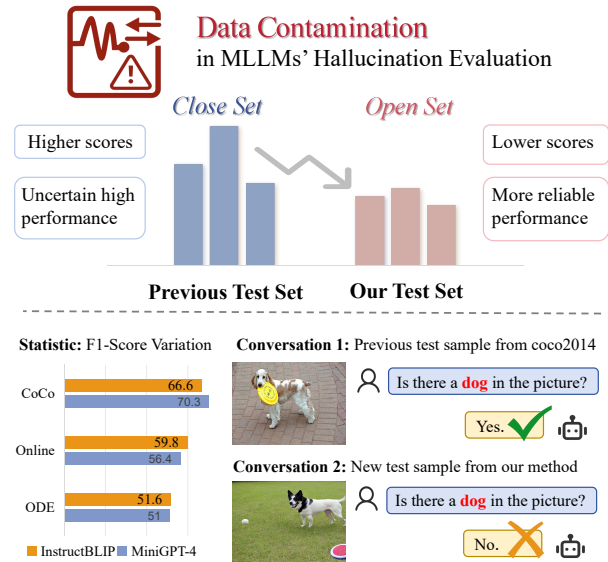

Figure 1. Comparison of closed-set and open-set evaluations for MLLMs, showing that open-set testing reduces data contamination and provides a more reliable assessment of hallucination rates.

sponses appear plausible but lack fidelity to the actual image content. This issue can lead to harmful consequences, limiting the utility of MLLMs.

Therefore, the evaluation of hallucinations in MLLMs is crucial to improve model reliability and practical application. Prior studies have proposed various benchmarks to evaluate hallucinations in MLLMs, focusing on different types (e.g., existence hallucinations[14, 26] and relational hallucinations [34]) or levels of difficulty [6, 31, 32]. However, these benchmarks are predominantly static, using fixed test data with limited distributions, which increases the risk of data contamination. Contamination occurs when test data overlaps with training data, leading to inflated performance metrics. For example, we find that, under the same semantic distribution, models perform better on the COCO2014 image subset than on the latest Internet images (as shown

in Fig. 1), with the latest Internet data being less likely to be contaminated by training overlap, thus providing a more reliable evaluation. This raises concerns about whether correct responses on COCO reflect genuine understanding or result from data contamination.

Recent studies[13, 35] highlight the issue of data contamination in Large Language Models (LLMs). Some LLM studies introduce dynamic evaluation methods to address this risk. For instance, DyVal [37] dynamically synthesizes test samples using a directed acyclic graph, though its application is limited to specific algorithms. Similarly, MSTemp[18] generates semantically equivalent evaluation samples based on the SST-2 dataset, yet its scope remains confined to the distribution of that dataset. To the best of our knowledge, there is currently no evaluation method existing to specifically mitigate data contamination in MLLMs.

We argue that an effective evaluation benchmark should be open-set, meaning that evaluation data are novel to the model at both sample and distribution levels. Inspired by contamination studies in the LLM domain and guided by insights into the unique challenges of multimodal models, we outline three distinctive features that characterize our approach: (1) Out-of-distribution (OOD) evaluation at a broad distributional level rather than modifying existing datasets; (2) dynamic sample generation across different modalities; and (3) a multi-tiered dynamic structure extending from concept-level to attribute-level and distribution-level granularity.

Building on these characteristics, we introduce the Open-Set Dynamic Evaluation (ODE) protocol. ODE automatically generates datasets to evaluate object hallucinations in MLLMs, covering both existence-level and attribute-level hallucinations. Initially, we model real-world object concepts, object attributes, and concept-attribute combinations as a graph. Then, from this graph, we extract concept nodes and their associated attributes, designing diverse semantic scenes and queries for each test sample and synthesizing high-quality images. To guide the selection of concept node pairs, we design four frequency-based criteria in the following order: Standard, Long-tail, Random, and Fictional. Each criterion reflects a unique distribution pattern of object combination frequencies (see Section 2.2 for detailed explanation).

The ODE protocol enables iterative dynamic updates to the dataset, generating new content based on selected concepts and their distribution combinations. The automated protocol enhances controllability while reducing human intervention. The phenomenon in Fig. 1, where the performance of our synthetic images is comparable to internet-sourced images, along with subsequent experiments, validates the effectiveness of synthetic images. We conducted extensive evaluations on multiple MLLMs under varied criteria, finding that hallucination rates were more pronounced

than on existing static benchmarks, with performance differences observed across distributions, tasks, and models. Further analysis revealed varying hallucination tendencies linked to specific concepts. Finally, we conduct fine-tuning experiments using ODE-generated samples, demonstrating that selectively refining models on ODE-identified error samples or directly fine-tuning on ODE-generated data effectively reduces hallucinations. Our findings primarily illustrate ODE's utility in general scenarios; additionally, ODE can also be applied to domains with data scarcity or imbalanced distributions.

In summary, this paper makes the following contributions:

- We introduce an open-set dynamic evaluation protocol that generates novel samples through dynamic target concept combinations, mitigating potential data contamination in MLLM hallucination evaluation.
- We perform extensive hallucination evaluations across multiple models, showing the effectiveness of our protocol.
- The generated evaluation data aid model debugging, and targeted fine-tuning on ODE-identified errors or general ODE-generated samples effectively enhances model performance.

2. Related Work

2.1. Hallucination Evaluation in MLLMs

To evaluate the degree of hallucination in MLLMs, various hallucination benchmarks have been proposed. The earliest proposed CHAIR [22] measures the accuracy of object references in captions by calculating the precision of hallucinatory objects. POPE [14] improves CHAIR by evaluating the presence of hallucinations based on object detection, suitable for discriminative tasks. AMBER [26] evaluates object hallucinations in both discriminative and generative tasks from three dimensions: existence, attributes, and relationships. Expanding beyond object hallucinations, examples include HallusionBench [6], which focuses on visual common sense and reasoning; Hal-Eval [10], which examines event hallucinations; and CorrelationQA [7], which studies the impact of false visual inputs and other cues like counting and OCR. Some studies further evaluate complex hallucinations. VHTest [9] compares embedding similarity to find potential VH instances, while HaloQuest [31] uses real and synthesized generated images to focus on more complex scenes. Despite the proliferation of benchmarks, from simple to complex tasks, and small to large scales, testing within closed sets remains unsolved. We dynamically update the dataset by combining target concepts to achieve an open set of samples and distributions.
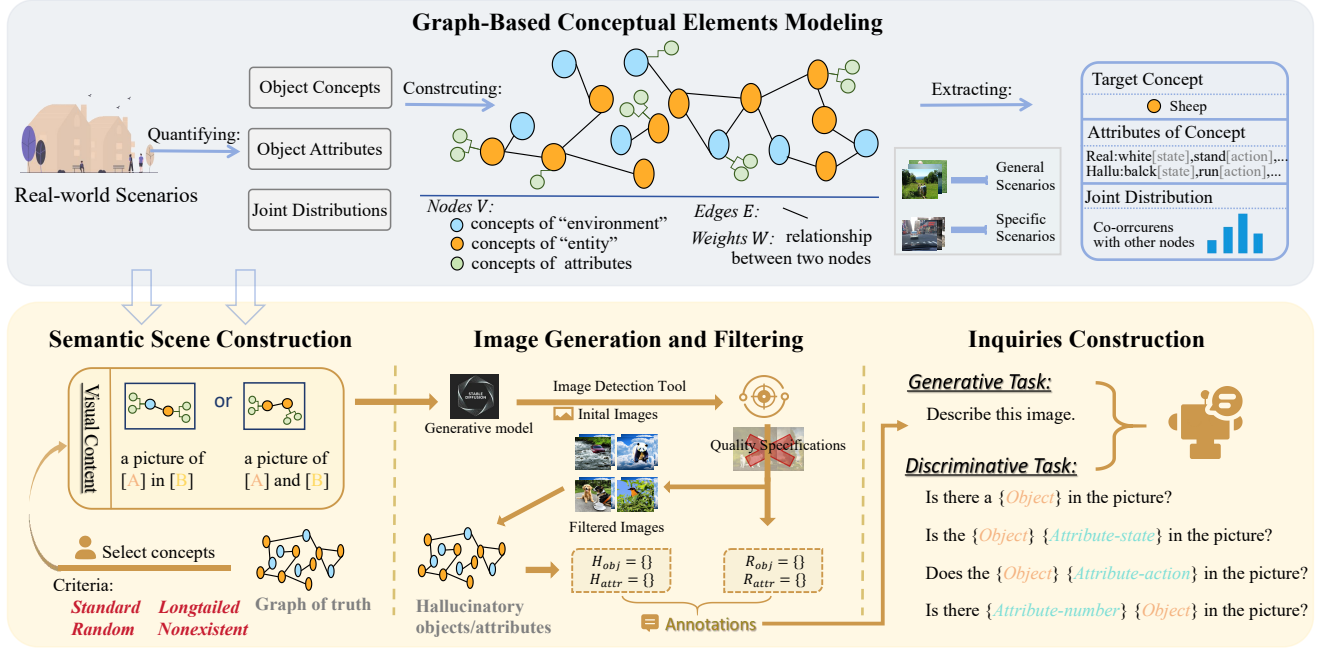
Figure 2. Pipeline of the Open-Set Dynamic Evaluation Protocol. The workflow involves constructing a graph and generating test samples based on the graph, with four distinct steps.

2.2. Data Contamination

Data contamination has attracted considerable attention in LLMs. The GPT-4[19] and LLama [25] reports highlight this phenomenon. Zhou et al. [35] discussed the risks and impacts of benchmark data contamination in evaluating LLMs; similarly, Ni et al. [21] explored related concerns. Several researchers have developed methods for detecting data contamination. Fan[5], Lei [12], Zhu [37], and others introduced dynamic evaluation strategies through different algorithms to reduce data pollution. The dataset generated through the ODE protocol is dynamically distributed and can be continuously updated, which avoids the problem of data pollution that may exist in existing static datasets for illusion evaluation.

2.3. Dynamic Evaluation

Dynamic Evaluation refers to dynamically generating evaluation data for testing. The key to early works such as DynaBoard [20], DynaTask [23], and Dynabench [11] was to utilize the intelligence of the crowd to challenge the design of evaluation sets, with the main effort focused on crowdsourcing systems and interfaces rather than algorithms, which required significant costs. Recent methods automate dynamic evaluation by rewriting and expanding samples based on task scenarios or datasets. Zhu et al. [37] used mathematical reasoning tasks to reduce data pollution in LLM by dynamically generating test samples from directed acyclic graphs. MsTemp [17], DyVal2 [38],

and Benchmark self evolving[29] dynamically reconstruct samples from existing datasets. DeNEVIL [4] iteratively updates and improves prompts based on values, and dynamically mines samples. Our approach reduces reliance on fixed datasets by dynamically combining meta-concepts and attributes, adapting visual content and textual prompts.

3. Methodology

This section outlines the ODE protocol for dynamically generating image content and prompt text for test data. As shown in Fig. 2, the workflow consists of four steps: modeling real-world scenarios using a graph structure, conceptual design of semantic scene, image generation and filtering, and template design for inquiries.

3.1. Graph-Based Conceptual Modeling

Aiming to cover a broader range of target object concepts in our evaluation of object existence hallucination, ODE employs a weighted graph G to model real-world scenes, facilitating the generation of more diverse scenarios in subsequent stages. Specifically, the graph $G = (V, A, E, W)$ provides an abstract representation of real-world objects, their interrelations, and attributes. Common object categories are extracted from existing datasets to form the nodes V, termed meta-concepts. In subsequent processes, a subset of these nodes will be selected as target concepts to test hallucination cases in the model. The attribute nodes A of the main nodes V represent the properties of each object

concept, including state and action attributes. Additionally, quantity attributes are considered during testing. The edge weights W denote the strength of the relationships between nodes, based on the co-occurrence quantity of object concepts in real-world scenes. We consider that co-occurrence frequency can reflect the semantic association and distributional characteristics between the conceptual entities V_i and V_j. For example, the high co-occurrence frequency between "table" and "chair" reflects their common pairing in indoor scenes. An edge E is established between nodes if a connection exists (i.e., the edge weight W is non-zero). To facilitate a more detailed analysis of the mutual influence of hallucinations among these concepts, we categorize the concepts into environment-level V_{env} and entity-level V_{ent}, such as "grass" and "frisbee". We particularly focus on two coexistence patterns: entity-environment and entity-entity coexistence patterns, ensuring the content generated in subsequent stages targeted.

ODE performs concept extraction from real-world scenarios and then focuses on hallucinatory scenarios by generating a graph that captures associations related to hallucinations, facilitating the creation of a comprehensive range of test samples for general hallucination evaluation. This modeling approach is applicable to both general and specific domains. We further emphasize the customization of hallucination detection tailored to specific fields.

3.2. Construction of Semantic Scenes

After obtaining a scene graph with object concepts, we select two concept nodes at each step to form a pair, which is used as the content for the test image. This image is then generated using a text-to-image model.

3.2.1. Selection Criteria

Through extensive learning from multimodal data, the model gradually acquires semantic representations of concepts, however, it may show varying comprehension of concept combinations, especially with diverse distributions, potentially causing hallucinations. Based on this observation, we design four distributional levels of concept combination standards—**Standard**, **Long-tail**, **Random**, and **Fictional**—to evaluate the model's grasp of different association patterns.

In implementation, we first select two object concepts V_i and V_j from the graph structure and assign each object specific attributes, including state A_{state}, action A_{action}, and quantity $A_{\text{number}} = 1$ attributes. For Standard and Long-tail pairs, attributes are chosen from frequently observed categories, reflecting realistic scenarios, while attributes for Random and Fictional pairs are randomly chosen to increase diversity. The four selection criteria are defined as follows:

• **Standard**: Pairs with the highest co-occurrence fre-

quency:
$$(V_i, V_j) \in \text{argmax}_{i,j}\, c_{i,j} \tag{1}$$
where $c_{i,j}$ represents the co-occurrence count of objects V_i and V_j in the existing dataset. These combinations, with attributes from standard categories, test the model's understanding of high-frequency associations.
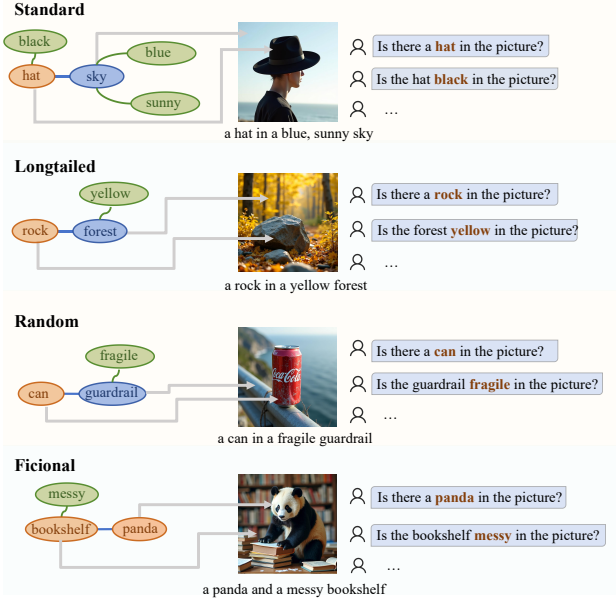

Figure 3. Examples of four distribution samples constructed by our method.

• **Long-tail**: Pairs with moderate co-occurrence, simulating long-tail distributions:
$$(V_i, V_j) \in \left\{ (V_k, V_l) \,\middle|\, \epsilon < c_{k,l} < \delta \right\} \tag{2}$$
where ϵ and δ set co-occurrence bounds. This standard evaluates the performance of the model in rare, associative pairs. Attributes are also chosen from commonly observed attributes to align with realistic contexts.

• **Random**: Randomly selected pairs without co-occurrence considerations:
$$(V_i, V_j) \sim \text{Uniform}(V \times V) \tag{3}$$
Attributes are also randomly chosen, evaluating the model's robustness to different semantic levels.

• **Fictional**: Pairs without any co-occurrence record:
$$(V_i, V_j) \notin \{(V_k, V_l) \,|\, c_{k,l} > 0\} \tag{4}$$
Fictional pairs assess the model's reasoning with novel distributions, assigning attributes at random.

Using these criteria, concept pairs (V_i, V_j) and attributes are dynamically selected to generate test images across distributions, ensuring varied semantic scenarios in the dataset and supporting a thorough evaluation of the model's semantic comprehension.

3.2.2. Scene Content

The semantic scene for each sample is constructed by extracting concept pairs and dynamically generating content, ensuring distinct test instances through inherent randomness. Our evaluation set includes two main object concepts and their attributes (or sometimes without specific attributes) in two combination types. The first type pairs two entity categories, requiring the model to capture direct interactions, such as "a black running dog and a yellow frisbee." The second type pairs an entity with an environmental category, prompting the model to understand contextual relationships, such as "a yellow deer and a lush forest" This approach broadens the range of concepts, facilitating detailed classification analysis of hallucination tendencies. Fig. 3 shows examples from the evaluation set.

3.3. Generation and Filtering of Images

To avoid model exposure to test data, we use text-to-image generation models (e.g., FLUX.1-dev and Stable Diffusion 1.5, as used in our experiments) to generate ODE test images from textual prompts like "a picture of attribute of A A and attribute of B B," where A and B represent specific visual concepts. Positive prompts include "clear and obvious entity," "photography," and "concise," while negative prompts use "bad anatomy," "incomplete body," and "mutated" to enhance quality.

Not all generated images meet quality standards due to generative model limitations. For each test case, we set different random seeds to produce multiple image variations with the same semantic scene, selecting higher-quality samples afterward. We assess image quality using an open vocabulary object detection model, discarding images lacking expected entities. For example, if an image described as "a picture of a dog and a frisbee" does not meet the required entities' confidence scores (below 0.65), it is filtered out. Our concept list includes all objects present in the images. By filtering low-confidence images, we retain high-quality samples, using the remaining concepts as "ground truth" data and annotating potential hallucinated objects based on object co-occurrence frequency.

3.4. Structuring of Inquiries

We develop an evaluation Inquiry template specifically designed to evaluate object-level and attribute-level hallucinations (state, action, and quantity) through automated generation.

For generative tasks, we use the inquiry "Please describe this image." to instruct the MLLM to identify the image's concepts. For discriminative tasks, object existence hallucinations are evaluated with inquiries like "Is there a {object} in the image?" Attribute hallucinations are evaluated with inquiries like "Is the {object} {state attribute} in the image?", "Does the {object} {action attribute} in the picture?", and "Is/Are there {quantity} {object(s)} in the picture?", expecting a "yes" or "no" response. To examine hallucinated objects, we include counterfactual prompts in the discriminative task, inquiring about nonexistent objects or attributes, such as "Is there a {hallucinated object} in the image?"

4. Experiments

4.1. Setup

Data Preparation. We extracted real-world target concepts through the AMBER benchmark[26], a multidimensional benchmark for evaluating MLLM hallucinations that includes 337 objects across 14 domains reflecting standard concepts. After concept modeling, we selected 40 concept combinations at four different distributional difficulty levels to generate images, resulting in 808 test images after sampling. Each image was paired with factual and hallucination questions, creating 8,786 test data pairs for both discriminative and generative tasks.

MLLMs Evaluated. We selected several state-of-the-art MLLMs for evaluation, including MiniGPT-4 [36], InstructBLIP [3], LLaVA-1.5 [16], CogVLM [30], mPLUG_Owl [33], Cambrian [24], InternVL-2.5 [2], and Qwen2-VL [28]. To ensure fair evaluation, we used each model's official hyperparameters to avoid length-based bias in response generation.

Evaluation Metrics. We adopted AMBER-derived evaluation metrics. For generative tasks, CHAIR measured hallucination frequency, Cover assessed content coverage, Hal represented the proportion of hallucinated responses, and Cog evaluated the similarity between generated hallucinations and human cognitive hallucinations. For discriminative tasks, we used standard classification metrics: Accuracy, Precision, Recall, and F1-Score. To evaluate hallucinations, ODE calculates Precision and Recall specifically for hallucination-related questions (ground truth "no") and uses Accuracy across all questions to prevent models from achieving high scores by simply rejecting answers.
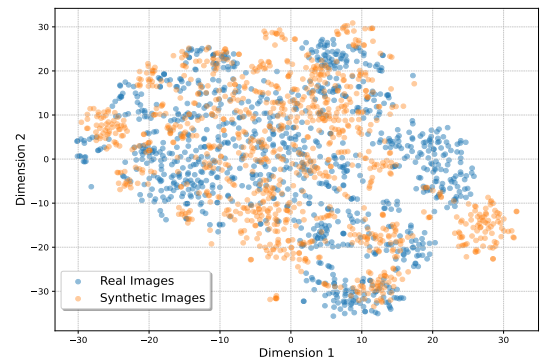

Figure 4. Visualization of synthetic and natural image features.

Table 1. Evaluation results of different models on both generative and discriminative tasks across various scenarios. Highlighted cells indicate optimal performance. Yellow highlights denote better performance with lower values, while blue highlights indicate better performance with higher values.

Criterion		Model	Generative Task				Discriminative-Existence Task				Discriminative-Attribute Task			
			CHAIR	Cover	Hal	Cog	Acc	P	R	F1	Acc	P	R	F1
AMBER		CogVLM	4.0	52.6	16.1	1.2	20.9	100.0	20.9	34.5	45.4	92.1	17.7	29.7
		LLaVA-1.5	7.7	50.4	34.4	3.7	71.0	100.0	71.0	83.0	72.5	89.8	50.7	64.8
		mPLUG	23.9	47.7	79.3	12.8	15.0	100.0	15.0	26.1	49.3	64.5	17.4	27.4
		MiniGPT-4	17.9	61.2	70.2	13.4	96.9	100.0	96.9	98.4	62.9	68.2	48.5	56.6
		InstructBLIP	12.5	57.5	63.2	8.5	67.4	100.0	67.4	80.5	70.6	88.8	47.2	61.6
Ours	Standard	CogVLM	51.9	76.5	89.1	12.2	50.7	100.0	26.2	41.5	55.2	46.8	55.6	50.8
		LLaVA-1.5	38.9	77.7	82.7	8.6	69.5	97.8	55.4	70.7	62.9	32.6	71.9	44.8
		mPLUG	50.8	77.2	96.0	11.5	41.7	94.7	13.4	23.5	72.5	41.7	28.6	33.9
		MiniGPT-4	49.4	76.0	93.6	14.2	64.5	97.5	48.0	64.3	69.5	39.7	12.5	19.0
		InstructBLIP	59.9	75.7	88.1	11.0	66.7	96.8	51.7	67.4	60.8	28.5	51.2	36.6
	Random	CogVLM	58.1	57.7	87.6	6.0	40.0	89.7	18.0	30.0	57.1	82.5	38.2	52.2
		LLaVA-1.5	45.2	57.7	84.2	4.7	74.7	89.7	69.6	78.3	75.7	90.1	65.9	76.1
		mPLUG	57.9	56.4	92.1	6.3	40.0	84.0	10.8	19.1	48.4	76.5	20.2	31.9
		MiniGPT-4	50.3	57.9	82.2	5.8	66.4	86.9	58.2	69.7	45.0	74.5	10.8	18.8
		InstructBLIP	55.9	58.4	83.2	5.6	64.6	87.6	54.6	67.2	72.3	91.5	58.6	71.4

4.2. Effectiveness of Synthetic Images

To validate the effectiveness of synthetic images in hallucination evaluation, we compared evaluation outcomes from three image sources under the same semantic distribution: a subset of COCO2014 images used in the POPE evaluation method, recent high-quality Internet images, and images generated via the ODE method with Stable-Diffusion 1.5. The results, presented in Fig. 1, indicate that test performance on COCO2014 images exceeds that on internet and ODE-generated images, suggesting potential data contamination, as the model may have encountered these images during prior training or fine-tuning. The observed difference in hallucination effects between ODE-generated images and internet images was minimal.

Additionally, we produced synthetic images containing the same visual information as 1,004 natural images. We extracted features using the CLIP model and reduced them to a two-dimensional space, revealing a high degree of similarity between synthetic and natural images in the feature space, as shown in Fig. 4. These findings suggest that synthetic images, within an acceptable margin of error, are a viable and sustainable alternative for constructing open-set datasets. Furthermore, we anticipate that improvements in text-to-image models will enhance the quality of generated images.

4.3. Main Results

Table 1 shows selected evaluation results (complete data are provided in the Appendix), and we found that:
- **Inconsistencies Between Static Benchmarks and ODE Performance:** Models like MiniGPT-4 and LLaVA perform well on static benchmarks, with F1-scores around 80-90. However, when tested with standard concept pairs in our evaluation, their scores declined. This suggests that these models may not fully understand fundamental concept relationships but rather rely on memorized correlations or noise from training data. This outcome highlights the limitations of static benchmarks, which can overestimate model capabilities due to data leakage and overfitting. Dynamic testing with diverse samples is necessary for a deeper evaluation of model learning and adaptability.
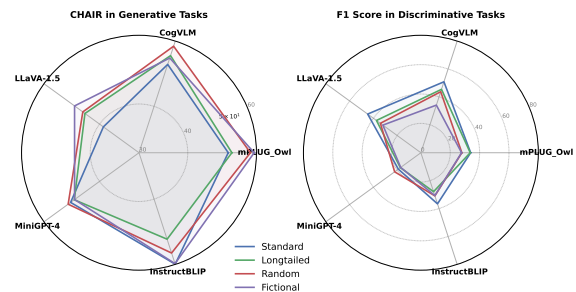

Figure 5. Performance across different distributions in generative tasks and discriminative tasks (existence-level).

- **Distribution Range and Hallucinations:** Performance varies significantly across different distributions, as shown in Fig. 5. On high-frequency concepts (e.g., Standard category), models perform well due to rich semantic associations. However, in object existence discrimination tasks, models like MiniGPT-4 exhibit instabil-

ity, likely due to an over-reliance on high-frequency patterns, which hampers generalization. In the Random and Fictional categories, hallucinations increase, particularly in attribute recognition, as the randomness of unrelated distributions amplifies generalization challenges and highlights the models' dependency on specific patterns. For low-frequency (Long-tailed) concepts, models perform adequately in generative tasks but struggle with object discrimination. This likely stems from insufficient low-frequency samples, limiting the foundational basis needed for generalized judgments and impairing the models' ability to handle combinations of low-frequency concept pairs.

- **Different Requirements of Tasks on Semantic Understanding:** Our results reveal distinct adaptive differences in model performance between high-frequency and low-frequency concepts in generative vs. discriminative tasks. During the training phase, models are repeatedly exposed to these high-frequency combinations, establishing robust semantic associations. This is reinforced in generative tasks which emphasize semantic consistency. This flexibility enables generative models to produce outputs that broadly align with concepts, even with incomplete understanding of details. Discriminative tasks, however, require precise identification of target concepts, making models sensitive to noise or biases from overrepresented pairs in training data. This over-memorization affects accuracy more in discriminative tasks, where even small misclassifications disrupt coherence.

- **Widening Performance Differences Among Models:** Different architectures and training methods, such as memory mechanisms and attention distributions, lead to performance variations. For example, mPLUG performs well in attribute discrimination, while MiniGPT-4 maintains balanced performance across tasks but lacks fine-grained attribute precision. Overall, LLaVA-1.5 exhibits robust performance across tasks and difficulty levels, while models like MiniGPT-4 and CogVLM highlight distinct strengths and weaknesses in data dependency and generalization capabilities.

4.4. Dynamic Benchmarking Updates

As a third-party benchmark evaluation, to ensure that models maintain robust performance in evolving application scenarios, we propose an ODE-based dynamic update mechanism for benchmark evaluations. This update mechanism involves the introduction of new pairs of target concepts, transformations in attribute combinations, and updates to the underlying knowledge base to include a broader range of object categories and their relationships. Such an iterative update process enables the benchmark to cover a wider range of distributional scenarios, ensuring both the accuracy and the comprehensiveness of model evaluation

Table 2. Performance of models on both generative and discriminative tasks (object existence hallucination) across two datasets, with the difference (Δ) between datasets.

Model	Dataset	Accuracy	Precision	Recall	F1Score
CogVLM	ODE-Flux	41.4	90.0	15.6	26.6
	ODE-SD	92.8	99.2	82.8	90.2
	Δ	+51.4	+9.2	+67.2	+63.6
LLaVA-1.5	ODE-Flux	51.3	84.7	32.9	47.4
	ODE-SD	94.3	98.6	87.3	92.6
	Δ	+43.0	+13.9	+54.4	+45.2
mPLUG_Owl	ODE-Flux	38.1	91.4	7.9	14.5
	ODE-SD	66.1	86.5	20.4	33.0
	Δ	+28.0	-4.9	+12.5	+18.5
MiniGPT-4	ODE-Flux	67.1	87.4	58.1	69.8
	ODE-SD	66.7	55.6	88.5	68.2
	Δ	-0.4	-31.8	+30.4	-1.6
InstructBLIP	ODE-Flux	51.5	91.1	30.4	45.6
	ODE-SD	72.1	97.8	56.1	71.3
	Δ	+20.6	+6.7	+25.7	+25.7

results. We conducted a second evaluation using data generated by Stable-Diffusion 1.5. Table 2 presents the model performance variations under a long-tail distribution. The significant discrepancies in performance indicate differing levels of model comprehension across various concepts.

5. Methodological Applications

5.1. Hallucination Tendencies in Evaluation Results

Our findings enable the analysis of hallucination tendencies within individual concepts and hallucination associations between different concepts. For example, we generated a frequency matrix of fact-hallucination concept pairs from LLaVA and performed clustering, identifying four groups, including indoor and traffic scene concepts. We found that hallucinations are more likely in scenarios with high contextual similarity or visual ambiguity. For instance, clusters containing both indoor and outdoor concepts (e.g., "car" and "chair") exhibit higher hallucination rates, as shown in Table 3, revealing potential weaknesses in the model's understanding of scene context and object differentiation.

Table 3. Cluster Analysis: Top Truth Concepts

Cluster	Top Truth Concepts
Indoor Concepts	table, chair, floor, person, cat
Mixed Concepts	car, person, bird, chair, cluster
Traffic & Outdoor	car, bench, bicycle, beach, road
Household Concepts	table, chair, cat, drink, lamp

5.2. ODE in Domain-Specific Scenarios

Our framework also addresses hallucination detection in specific scenarios and data-scarce fields. The input flexibil-

ity of ODE allows for the selection of particular concepts, combined with its ability to generate customizable images, opening new possibilities for tailored data. As multimodal models expand into fields such as autonomous driving and healthcare, constructing diverse, challenging samples becomes critical to overcome existing datasets' limitations, such as narrow distributions and small sample sizes, which limit adaptability to complex scenarios. Fig. 6 demonstrates our framework's capability to generate rare concept combinations for the transportation domain, supplementing real-world scene data. This generative capability exposes models to scarce data during training, thereby enhancing real-world adaptability and mitigating overfitting risks. Additionally, fine-tuning with ODE-generated, distributionally diverse images improves model reliability in specialized domains.

Figure 6. Examples of rare distribution samples constructed by our method in the transportation domain.

5.3. Fine-Tuning Enhancements

Furthermore, fine-tuning MLLMs with ODE-generated data effectively mitigates object hallucinations. The ODE-generated test set avoids data contamination while offering rich distributional diversity, uncovering "thinking patterns" that drive hallucinations in specific scenarios. Based on this, targeted fine-tuning using the erroneous sample set can effectively mitigate the model's shortcomings. We utilize a dynamic, iteratively generated test set, which helps prevent model overfitting and the potential for "score boosting" on a single dataset. In our experiments, we extracted the erroneous samples exhibiting hallucinations from the initial round of generated data in the LLaVA model tests and performed fine-tuning on LLaVA. Subsequently, we evaluated the fine-tuned model using the existing AMBER test set. The results presented in Tables 4 and 5 indicate that the fine-tuned model shows significant improvements in performance across both generative and discriminative tasks.

The dataset, comprising Standard, Long-tail, Random, and Fictional distributions, enabled an open-set analysis. For the object existence hallucination task, 800 sample pairs from each distribution type were used for fine-tuning.

Table 4. Generative Task Performance Comparison

Model	CHAIR ($\downarrow$)	Cover ($\uparrow$)	Hal ($\downarrow$)	Cog ($\downarrow$)
Initial	8.1	51.4	36.1	4.1
Fine-tuned	**6.5**	50.4	**28.5**	**2.9**
Δ	-1.6	-1.0	-7.6	-1.2

Table 5. Discriminative Task Performance Comparison

Model	Existence-level		Attribute-level	
	Recall($\uparrow$)	F1-Score($\uparrow$)	Recall($\uparrow$)	F1-Score($\uparrow$)
Initial	71.0	83.0	50.7	64.8
Fine-tuned	**96.0**	**97.9**	**90.7**	**68.0**
Δ	+25.0	+14.9	+40.0	+3.2

The fine-tuned model exhibited better hallucination control. Fig. 7 Results indicate that high-frequency distribution samples had the greatest positive effect, while long-tail samples contributed less than random samples. This discrepancy may arise from the regularity of high-frequency data, which facilitates model pattern recognition, whereas the imbalance in long-tail samples can lead to overfitting of frequent categories. Random samples provided better regularization, improving performance across diverse test sets. Incorporating diverse distributions in training data enhances robustness, generalization, and representation of low-frequency objects, boosting overall performance.
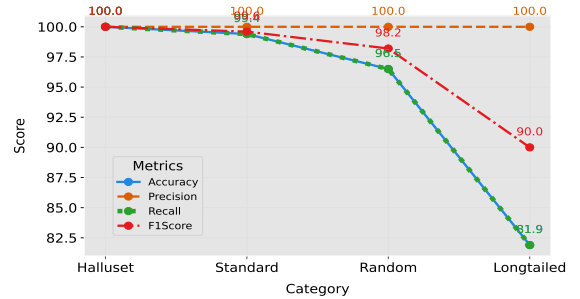

Figure 7. Performance of fine-tuned LLaVA models across distribution types, showing highest improvement with standard samples and limited impact from long-tail samples.

6. Conclusions

This paper addresses the issue of data contamination in the hallucination evaluation of multimodal large language models. We introduce a dynamic open-set evaluation protocol, initially applied to object hallucination at existence-level and attribute-level in visual question answering. The experimental results are more reliable than static benchmarks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 5

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1, 5

[4] Shitong Duan, Xiaoyuan Yi, Peng Zhang, Tun Lu, Xing Xie, and Ning Gu. Denevil: Towards deciphering and navigating the ethical values of large language models via instruction learning. *ArXiv*, abs/2310.11053, 2023. 3

[5] Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, Yongfeng Zhang, and Libby Hemphill. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*, 2023. 3

[6] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. pages 14375–14385, 2024. 1, 2

[7] Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. The instinctive bias: Spurious images lead to hallucination in mllms. *ArXiv*, abs/2402.03757, 2024. 2

[8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. 1

[9] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. pages 9614–9631, 2024. 2

[10] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. pages 525–534, 2024. 2

[11] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, 2021. Association for Computational Linguistics. 3

[12] Fangyu Lei, Qian Liu, Yiming Huang, Shizhu He, Jun Zhao, and Kang Liu. S3Eval: A synthetic, scalable, systematic evaluation suite for large language model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1259–1286, Mexico City, Mexico, 2024. 3

[13] Yucheng Li. An open source data contamination report for llama series models. *arXiv preprint arXiv:2310.17589*, 2023. 2

[14] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1, 2

[15] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 1

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 5

[17] Yachuan Liu, Liang Chen, Jindong Wang, Qiaozhu Mei, and Xing Xie. Meta semantic template for evaluation of large language models. *arXiv preprint arXiv:2310.01448*, 2023. 3

[18] Yachuan Liu, Liang Chen, Jindong Wang, Qiaozhu Mei, and Xing Xie. Meta semantic template for evaluation of large language models. *arXiv preprint arXiv:2310.01448*, 2023. 2

[19] B. Lovin. Gpt-4 performs significantly worse on coding problems not in its training data, 2023. 3

[20] Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Yu Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. In *Neural Information Processing Systems*, 2021. 3

[21] Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need. *arXiv preprint arXiv:2409.01790*, 2024. 3

[22] Anna Rohrbach, Lisa Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. pages 4035–4045, 2018. 2

[23] Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. Dynatask: A framework for creating dynamic AI benchmark tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 174–181, Dublin, Ireland, 2022. Association for Computational Linguistics. 3

[24] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 5

[25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[26] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 1, 2, 5

[27] Jiaqi Wang, Yifei Gao, and Jitao Sang. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*, 2024. 1

[28] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[29] Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*, 2024. 3

[30] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 5

[31] Zhecan Wang, Garrett Bingham, Adams Yu, Quoc Le, Thang Luong, and Golnaz Ghiasi. *HaloQuest: A Visual Hallucination Dataset for Advancing Multimodal Reasoning*, pages 288–304. 2024. 1, 2

[32] Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53553–53570, 2024. 1

[33] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 5

[34] Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *arXiv preprint arXiv:2408.09429*, 2024. 1

[35] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023. 2, 3

[36] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 5

[37] Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*, 2023. 2, 3

[38] Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. Dyval 2: Dynamic evaluation of large language models by meta probing agents. *arXiv preprint arXiv:2402.14865*, 2024. 3