

Blind Spot Navigation: Evolutionary Discovery of Sensitive Semantic Concepts for LVLMs

Zihao Pan^{1*}, Yu Tong^{2*}, Weibin Wu^{1†}, Jingyi Wang³, Lifeng Chen⁴,
Zhe Zhao⁵, Jiajia Wei¹, Yitong Qiao¹, Zibin Zheng¹

¹School of Software Engineering, Sun Yat-sen University, ²Wuhan University

³Tsinghua Shenzhen International Graduate School, Tsinghua University

⁴Computer Science, Beijing Jiaotong University

⁵University of Science and Technology of China

<https://github.com/Pan-Zihao/Blind-Spot-Navigation>

Abstract

Adversarial attacks aim to generate malicious inputs that mislead deep models, but beyond causing model failure, they cannot provide certain interpretable information such as “*What content in inputs make models more likely to fail?*” However, this information is crucial for researchers to specifically improve model robustness. Recent research suggests that models may be particularly sensitive to certain semantics in visual inputs (such as “wet,” “foggy”), making them prone to errors. Inspired by this, in this paper we conducted the first exploration on large vision-language models (LVLMs) and found that LVLMs indeed are susceptible to hallucinations and various errors when facing specific semantic concepts in images. To efficiently search for these sensitive concepts, we integrated large language models (LLMs) and text-to-image (T2I) models to propose a novel semantic evolution framework. Randomly initialized semantic concepts undergo LLM-based crossover and mutation operations to form image descriptions, which are then converted by T2I models into visual inputs for LVLMs. The task-specific performance of LVLMs on each input is quantified as fitness scores for the involved semantics and serves as reward signals to further guide LLMs in exploring concepts that induce LVLMs. Extensive experiments on seven mainstream LVLMs and two multimodal tasks demonstrate the effectiveness of our method. Additionally, we provide interesting findings about the sensitive semantics of LVLMs, aiming to inspire further in-depth research.

1 Introduction

Large Vision-Language Models (LVLMs) have achieved remarkable success in various multimodal downstream tasks, demonstrating great potential in understanding the real world—for example, in image captioning [7, 3, 33], visual question answering [2, 15, 36], and text-to-image generation [23, 28, 30]. Given the increasing demand for LVLm applications, the security and robustness of the model is increasingly important [27, 12].

Recent research has shown that powerful LVLMs still have security risks and potential vulnerabilities, especially when facing various adversarial attack methods [44, 21, 43, 20, 45, 42]. Adversarial attacks typically apply imperceptible perturbations to benign inputs, causing models to fail on various tasks and exposing the fragility of neural network-based models [13]. Existing attack methods targeting

*Equal Contribution

†Corresponding Author

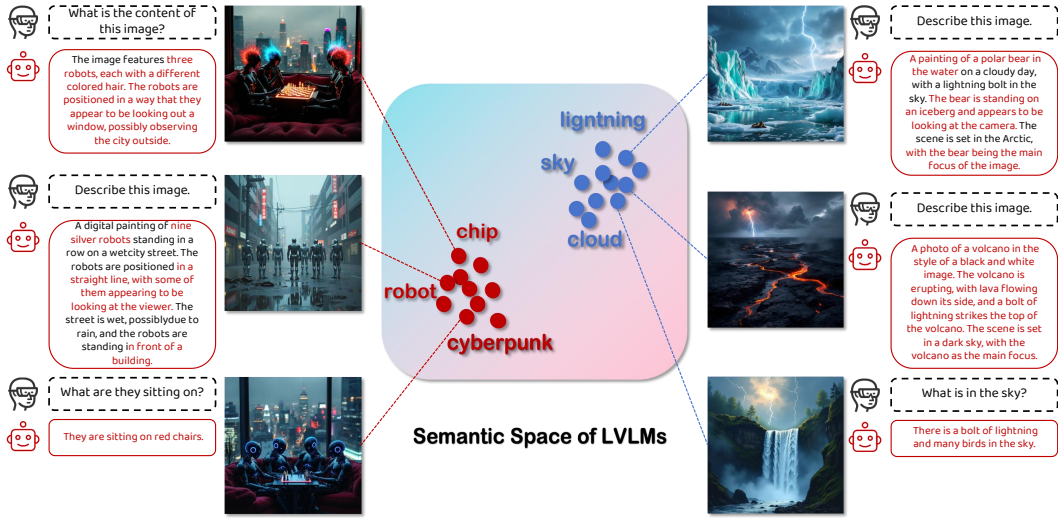


Figure 1: An overview of the core idea of our work. LVLMs may be particularly sensitive to certain semantic concepts, which can be viewed as blind spots in their knowledge space.

LVLMs usually mislead models by applying carefully designed noise perturbations to image inputs. However, when we want to understand what content in an input image makes LVLMs more prone to hallucinations or various errors, the above methods prove ineffective. This is because the artificially added perturbations in adversarial attacks do not contain semantic information, which leaves people at the point of causing model failure without understanding its failure patterns [46, 17]. Recent studies [46, 25] have found that classification models may be more sensitive to visual inputs containing specific semantic content. For example, when images contain semantic information such as “wet,” “foggy,” or “stretched,” the probability of model failure increases. Uncovering these sensitive semantic concepts is valuable because they represent the model’s “knowledge blind spots,” and understanding them helps researchers improve models in a targeted manner [17]. So we cannot help but ask: *Are LVLMs also sensitive to certain specific semantic concepts? How can we efficiently search for these semantic concepts?*

To address the above questions, we innovatively combined large language models (LLMs) and text-to-image (T2I) models to propose a semantic evolution algorithm framework to search for blind spots in LVLMs’ vast knowledge space. Existing limited researches [46, 25] attempt to artificially construct a word space to simulate the potential semantic space of image classification models, and subsequently optimize extraction and combination strategies for the word space using the classification model’s loss, to find a set of semantic concepts for synthesizing images that are most likely to mislead the model. However, their limited search space seems impractical compared to powerful LVLMs, resulting in poor effectiveness. Therefore, we choose to leverage the powerful prior knowledge of LLMs by carefully designing a series of prompts to drive the LLM to execute evolutionary algorithms for efficient searching. Specifically, the LLM first randomly initializes a large number of semantic concepts as the initial population, organized in the form of textual descriptions of image content. Subsequently, multiple sets of descriptions are repeatedly drawn from the initial population and input to the LLM to perform crossover and mutation operations, simulating searches in the semantic space. Finally, the LLM outputs the merged and modified image descriptions, which are then converted by the T2I model into corresponding images as visual inputs for LVLMs. We evaluate LVLMs’ sensitivity to the semantics involved in these images on multimodal tasks and quantify it as fitness scores for the corresponding image descriptions. These scores serve as reward signals to guide the LLM in exploring semantic concepts that better induce LVLM failures in the next round of evolution. After multiple iterations, we believe that the final batch of image descriptions represents the sensitive semantic concepts of LVLMs.

To verify the effectiveness of our method, we conducted extensive experiments on seven mainstream open-source and commercial LVLMs (LLaVA [19], LLaVA-NeXT [18], InternVL2 [8], Molmo [9], Qwen2-VL [38], Llama-3.2 [1], GPT-4o [24]), covering two multimodal tasks: image captioning (IC) and visual question answering (VQA). As shown in Table 1 and 2, the semantics we searched can effectively reduce the performance of LVLMs, particularly outperforming SOTA baselines by a large margin of **54%** on average. By analyzing the image descriptions obtained from our multiple rounds

of searching, we found that LVLMs are indeed particularly sensitive to certain special semantic concepts in images. This is manifested in frequently occurring similar concepts such as "robots," "cyberpunk," *etc.*, and the significant decrease in LVLMs' capabilities when facing them, as shown in Figure 1. Additionally, although the most sensitive concepts for each LVLM differ, the concepts found by searching for a particular LVLM can mislead all other LVLMs to varying degrees, similar to the transferability of adversarial examples. These findings can help us better understand the potential defects of LVLMs and inspire the design of targeted improvement strategies.

Our contribution can be summarized as follows:

- Our method fills the gap where existing adversarial attacks fail to explore sensitive semantic concepts in LVLMs, not only generating effective examples that mislead LVLMs, but also revealing which semantic concepts in visual inputs are more likely to induce model failures.
- We propose a novel evolutionary algorithm framework based on LLM and T2I Model. Through LLM-based crossover, mutation, and selection operations, randomly initialized semantic concepts gradually converge toward certain similar concepts after multiple rounds of evolutionary iterations.
- Extensive experiments demonstrate that our method significantly outperforms existing methods directly transferred from classification models. Our method also provides valuable findings about LVLM-sensitive semantics, aiming to promote in-depth research on LVLM robustness.

2 Related Work

2.1 Large Vision-Language Models

Benefiting from the success of LLMs, recent large vision-language models (LVLMs) have achieved significant advancements across various multimodal downstream tasks, such as Image Captioning and Visual Question Answering. These LVLMs built upon the achievement of LLMs train a modality connector to align the vision space and text space. Popular LVLMs include open-source models like LLaVA-v1.5 [19], LLaVA-NeXT [18], InternVL2 [8], Qwen2-VL [38], Molmo [9], Llama-3.2 [1], and commercial models like GPT-4o [24]. Despite their remarkable capabilities, the increased complexity and deployment of LVLMs have also exposed them to various security threats and vulnerabilities, making the study of attacks on these models a critical area of research. In this paper, we further explore what semantic content in the inputs is more likely to cause Large Vision-Language Models (LVLMs) to fail.

2.2 LLMs for Evolution Method

LLMs are now instrumental in the innovation of several algorithmic frameworks. They have been effectively integrated as black-box components in the development of evolutionary algorithms [40]. Wang *et al.* [39] used an LLM-based evolutionary approach for automatic data augmentation to address the long-tail problem. Guo *et al.* [14] developed a framework for automatically designing adversarial attack algorithms. Inspired by the above, in order to conduct effective searches in the vast semantic space of LVLMs, we designed an LLM-based evolutionary algorithm framework.

2.3 Adversarial Attacks for LVLMs

Several recent researches have explored the robustness of MLLMs [45, 42, 41, 31, 26]. These researches are mostly under untargeted settings, or try to mislead the content of the input image. Zhao *et al.* [44] explore the robustness of VLMs by using transfer-based and query-based methods to craft adversarial examples. Dong *et al.* [11] use an ensemble-based method to mislead Google Bard. Although they can evaluate the robustness of LVLMs by observing their ability to resist adversarial examples, they cannot answer: *what semantic content in inputs more easily leads to LVLM failures?* This paper aims to fill this gap by identifying their sensitive semantic concepts while also generating adversarial examples that can mislead LVLMs.

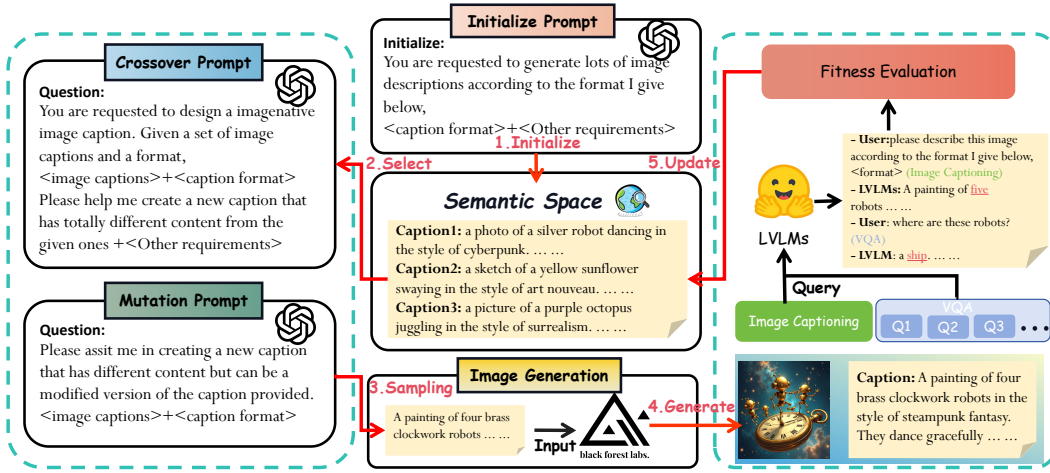


Figure 2: Overview of our semantic evolution framework.

3 Method

3.1 Framework

The overall framework is illustrated in Figure 2. The framework consists of two interactive modules: the LLM-based semantic evolution module (left) and the LVLM sensitivity evaluation module (right).

Large Vision-Language Models. Let \mathcal{M} represent an LVLM model, which takes an image x and a text prompt t_{in} as inputs and outputs a textual output $M(x, t_{in}) = t_{out}$. Since LVLM drives multiple tasks, in image captioning tasks, for instance, t_{in} is a placeholder \emptyset and t_{out} is the caption; in visual question answering tasks, t_{in} is the question and t_{out} is the answer [17].

Problem Formulation. We aim to identify blind spots in the vast knowledge space of LVLMs, specifically searching for semantic concepts in input images that are more likely to cause model failure. To simulate the visual input semantic space received by LVLMs, we used an LLM to randomly generate a large number of image descriptions, aiming to cover as many common semantic concepts as possible. Subsequently, we used these generated captions as the initial semantic population for our evolutionary algorithm framework. Our goal is to drive the LLM to iteratively generate new image descriptions based on the initial population and adjust the LLM’s generation objectives according to feedback from the LVLM on multimodal tasks, simulating a search of the LVLM’s semantic space. Since LLMs can only process and generate text, we need to use T2I models to transform the semantic concepts generated by the LLM in text form into corresponding content in images, and use the generated images as visual input for the LVLM on a multimodal task such as Image Captioning. The evolutionary algorithm will execute multiple rounds iteratively to expand the LLM’s search range as much as possible, encouraging the generated image descriptions to gradually include semantic concepts that are sensitive to LVLMs, until the LVLM exhibits obvious failure on a task or the content generated by the LLM becomes very similar. Overall, our proposed LLM-based semantic evolutionary framework can be described as:

$$s^* = \arg \max_s \text{Eval}(\mathcal{M}(G(s), t_{in}), T), \quad (1)$$

where Eval represents the evaluation of LVLM’s sensitivity to a particular image input on a specific task, T represents the description of the multimodal task, G represents the T2I model, which receives a text input and transforms it into a corresponding image, s represents all image descriptions generated by the LLM, *i.e.*, the individuals in the evolutionary algorithm, and s^* represents the sensitive semantic concepts finally discovered through the search.

3.2 LLM-based Semantic Evolution

Our method leverages the powerful prior knowledge of LLMs to construct a sensitive semantic concept search space that matches with LVLM. In order to guide the LLM as a black-box component to effectively execute evolutionary algorithm, we employ prompt engineering techniques and carefully design a series of prompts. By integrating task descriptions, image description formats, novelty

requirements, and other prior knowledge into the prompts, we can constrain the LLM’s generation process within the desired search space. We design the following three types of search operators, corresponding to different prompt templates:

- **Initialization operator I :** Based on the task description prompt P_{task} and the predefined caption format P_{format} , generate a set of randomly initialized image caption populations $C_i^{(0)}$, where $i = 1, 2, \dots, N$.
- **Crossover operator E :** Based on P_{task} , select N_p parent captions $C_i^{(t)}$ from the current population, and combine them with P_{format} , asking the LLM to generate N_e new captions that are content-wise different from all existing captions, thus expanding the search space.
- **Mutation operator M :** Based on P_{task} , select N_m individuals $C_i^{(t)}$ from the current population and provide local improvement directions. The task of the LLM is to generate a mutated individual $\hat{C}_i^{(t)}$ for each $C_i^{(t)}$, further exploring its neighbourhood.

Taking the crossover operator E as an example, its prompt P_E can be represented as follows:

$$P_E(P_{task}, \{C_i^{(t)}\}_{i=1}^{N_p}, P_{format}) = P_{task} + P_{ref}(\{C_i^{(t)}\}_{i=1}^{N_p}) + P_{format} + P_{diff}, \quad (2)$$

where P_{task} is the task description, C_i represents the N_p parent image captions, and P_{format} is the predefined caption format. P_{ref} refers to the reference for all parent image captions, and P_{diff} requires generating new descriptions that are completely different from the existing image captions.

Once the prompt P_E is obtained, it is input into the pre-trained LLM \mathcal{L} , which generates N_e new crossover individuals:

$$\{C_j^{(t)}\}_{j=1}^{N_e} = \mathcal{L}(P_E(P_{task}, \{C_i^{(t)}\}_{i=1}^{N_p}, P_{format})). \quad (3)$$

Each $C_j^{(t)}$ is an image caption, which will subsequently be input into a T2I model to be transformed into a corresponding image, and the LVLM sensitivity evaluation module will calculate the LVLM’s vulnerability to this image and obtain the fitness score of $C_j^{(t)}$. Similarly, the prompts P_I and P_M for the initialization operator I and mutation operator M can be constructed in a similar way, with the main difference being the different prior information introduced.

In the initial stage, the initialization operator is used to generate a random population. In each generation, a portion of individuals are selected from the previous generation as parents, and the crossover operator and mutation operator are used to generate new image captions, which are then merged into the population. For each candidate caption, the LVLM sensitivity evaluation module is called to calculate its fitness score. We provide detailed prompts for each stage of the evolutionary algorithm in the Section 4.4 and Appendix A.

3.3 LVLM Sensitivity Evaluation

To evaluate the sensitivity of LVLMs to semantic concepts contained in candidate image captions generated by LLMs, we transform text-form semantics into corresponding visual content through a T2I model as visual input for the LVLM, and provide a text query to obtain the LVLM’s output. Given a multimodal task, we evaluate the LVLM’s performance when accepting each candidate semantic concept and quantify it as a fitness score for the corresponding image caption. Assuming the caption generated by the LLM is represented as C , and the function that evaluates the LVLM’s output based on the input image for a specific task is represented as \mathcal{T} , the evaluation process can be represented as follows:

$$\text{Fitness}(C) = -\mathcal{T}(C, \mathcal{M}(G(C), t_{task}), task) \quad (4)$$

where t_{task} is the text prompt that drives the LVLM to perform a specific multimodal task based on the input image $G(C)$. The score calculated by \mathcal{T} serves as the fitness of C in the evolutionary algorithm, guiding the LLM to further generate based on C . It should be noted that the higher the score output by \mathcal{T} , the better the performance of the LVLM, but this means that the fitness of the candidate image caption in the evolutionary algorithm is lower. Consistent with the optimization goal of adversarial attacks, we want to obtain semantic concepts that most easily lead to LVLM failures, that is, sensitive semantic concepts.

Table 1: CAPTURE Score on seven mainstream LVLMs and two multimodal tasks with baselines.

Task	Method	LLaVA	LLaVA-NeXT	InternVL2	Qwen2-VL	Molmo	Llama-3.2	GPT-4o
IC	Random	0.4911	0.4882	0.5036	0.5113	0.4879	0.4779	0.5115
	LANCE	0.2879	0.2873	0.2608	0.3190	0.2821	0.2339	0.3316
	NL-adv	0.3720	0.3352	0.3509	0.3783	0.3500	0.3634	0.3840
	Ours	0.1109	0.1206	0.1885	0.1041	0.0915	0.0910	0.2233
VQA	Random	0.4933	0.4906	0.4926	0.4983	0.4807	0.4718	0.4946
	LANCE	0.3018	0.2900	0.2540	0.3254	0.2706	0.3059	0.3032
	NL-adv	0.3626	0.3356	0.3337	0.3640	0.3210	0.3009	0.3225
	Ours	0.1037	0.1284	0.1416	0.1079	0.0992	0.0952	0.2474

In order to evaluate in real-time the sensitivity of LVLM to candidate image semantics provided by LLM in multimodal tasks, we choose to compare the degree of correspondence between LVLM’s output and the actual image content in the caption as a measure of LVLM’s understanding capability of relevant semantics. Inspired by CAPTURE [10], we introduce a multi-stage, multi-scale caption evaluation method. It can effectively evaluate the content consistency and descriptive quality of other descriptive texts of an image compared to the reference caption. Specifically, we extract core visual elements rather than n-gram pieces to reduce the influence of varying writing styles. The method consists of three key components: (1) Visual elements extraction module that uses Factual parser [16] to extract objects, attributes, and relations from texts; (2) Stop words filtering module that removes abstract nouns to focus on concrete visual elements; (3) Visual elements matching module that combines exact matching, synonym matching using WordNet [22], and soft matching using Sentence BERT [29]. Matched candidate elements are formulated as $\text{cand}_{\text{type}}^{\text{match}} = \text{cand}_{\text{type}}^{\text{ex}} \cup \text{cand}_{\text{type}}^{\text{syn}}$, where $\text{type} \in \{\text{obj}, \text{attr}, \text{rel}\}$. The similarity matrix for soft matching is calculated as $S_{\text{type}}^{\text{rm}} = \phi(\text{cand}_{\text{type}}^{\text{rm}}) \times \phi(\text{gt}_{\text{type}}^{\text{rm}})^T$, where $\phi(\cdot)$ denotes the Sentence BERT model. Precision and recall are computed for each element type, and the final score is calculated as $\frac{\alpha F1_{\text{obj}} + \beta F1_{\text{attr}} + \gamma F1_{\text{rel}}}{\alpha + \beta + \gamma}$, where $F1_{\text{type}} = \frac{\text{precision}_{\text{type}} \cdot \text{recall}_{\text{type}}}{\text{precision}_{\text{type}} + \text{recall}_{\text{type}}}$ and α, β, γ are scale factors. More details can be found in the Appendix B. For IC tasks, we directly use the above method to evaluate the content consistency between candidate image captions and descriptions output by LVLM; for VQA tasks, we manually set up some questions querying the content of images, such as subject and environment descriptions, and also called LLM to automatically generate some Q&A pairs based on candidate captions, and similarly verify the correctness of LVLM’s answers according to the captions or answers.

4 Experiments

4.1 Experiment Setup

Baselines. Given that there is no similar work to ours on LVLMs, we selected two works [46, 25] that explore sensitive semantic concepts on classification models, and tested whether they could be effective for LVLMs. They both attempt to simulate the model’s potential semantic space by artificially constructing a word space, and subsequently executing searches within this space to find certain semantic concepts that more easily mislead the model. NL-adv [46] utilizes the classification model’s loss to optimize extraction and combination strategies for the word space, outputting an optimal word combination, *i.e.*, an image description. Subsequently, a T2I model converts this description into adversarial examples that induce the classification model. LANCE [25] directly generates text descriptions of existing images as a semantic space, randomly replacing descriptive words and performing image editing to identify which concepts the classification model is more sensitive to. For a fair comparison, we set LANCE’s initial image description collection to be the same as the initial caption population of the evolutionary algorithm, and align the number of iterative optimization rounds. For NL-adv, we maintain its original word space and replace its optimization objective based on classification model loss with the LVLM performance evaluation score in our Section 3.3.

LVLMs Involved. We conducted experiments on seven models, including open-source ones like LLaVA-v1.5-13B [19], LLaVA-v1.6-mistral-7B [18], InternVL2-8B [8], Qwen2-VL-7B-Instruct [38], Molmo-7B-D-0924 [9], Llama-3.2-11B-Vision-Instruct [1], and the commercial model GPT-4o [24]. For open-source models, we conducted testing by deploying them locally; while for commercial

Table 2: Cross-model transferability of sensitive semantic concepts searched on each LVLM

Surrogate LVLM	LLaVA	LLaVA-NeXT	InternVL2	Qwen2-VL	Molmo	Llama-3.2	GPT-4o
LLaVA	0.1108	0.1204	0.1324	0.1301	0.1247	0.1066	0.1757
LLaVA-NeXT	0.1143	0.1201	0.1436	0.1434	0.1346	0.1193	0.1938
InternVL2	0.2055	0.2143	0.1723	0.1743	0.1753	0.2010	0.1806
Qwen2-VL	0.1039	0.0860	0.1308	0.0920	0.1097	0.1091	0.1633
Molmo	0.1417	0.1087	0.1259	0.1508	0.1110	0.1294	0.1667
Llama-3.2	0.1047	0.1066	0.1346	0.1334	0.1355	0.0904	0.1899
GPT-4o	0.2004	0.2239	0.2145	0.2349	0.1893	0.1868	0.2192

(a) Image Captioning task

Surrogate LVLM	LLaVA	LLaVA-NeXT	InternVL2	Qwen2-VL	Molmo	Llama-3.2	GPT-4o
LLaVA	0.1042	0.1153	0.1564	0.1467	0.1256	0.1033	0.2131
LLaVA-NeXT	0.1092	0.1254	0.1487	0.1557	0.1377	0.1072	0.2000
InternVL2	0.1215	0.1348	0.1511	0.1577	0.1482	0.1342	0.1476
Qwen2-VL	0.1823	0.1641	0.1345	0.1195	0.1300	0.1518	0.1739
Molmo	0.1049	0.1054	0.1014	0.1127	0.1109	0.0903	0.1066
Llama-3.2	0.1089	0.1131	0.1357	0.1423	0.1311	0.0924	0.1945
GPT-4o	0.2318	0.2399	0.2372	0.2603	0.2548	0.2317	0.2457

(b) VQA task

models, we used the officially provided APIs [24]. For each LVLM, we conducted comprehensive experiments on both IC and VQA tasks to validate our method.

Metric. To evaluate the effectiveness of our method in discovering sensitive semantic concepts of LVLMs, we need to test whether LVLMs show significant performance degradation on multimodal tasks when faced with image inputs containing these concepts. As described in Section 3.3, our introduced multi-stage, multi-scale approach can effectively evaluate LVLM performance on multimodal tasks, outperforming previous rule-based methods such as SPICE, BLEU-1, *etc.* Therefore, we choose to use CAPTURE Score [10] as the sensitivity assessment for LVLMs, where a lower CAPTURE Score indicates more fragile performance of the LVLM on the corresponding task and semantic concept, meaning our method is more effective.

Implementation Details. We used FLUX1.1pro(ultra) [6] to convert each image caption into highly aligned images for LVLMs’ visual input, with 50 steps, a guidance scale of 3.5, and each image sized at 1024×1024 . GPT-4o [24] was used to execute the evolutionary method. For the randomly initialized image caption set, we use GPT-4o and Claude-3.5-Sonnet [4] to generate an average of 1,150 diverse texts. The key parameters of our proposed evolutionary algorithm framework are shown in Table 3. We set the evolution generations of the population to 6, meaning six selection operations are executed to sample elite individuals. In each generation, “pop_size” represents the size of each population, “pop_save_number” represents the number of elite individuals retained for the next generation after multiple crossover and mutation operations, “n_pop” represents the number of populations, and m represents the number of parent samples selected each time when performing crossover and mutation operations. The final 5 populations containing 5 individuals each serve as the output results of the algorithm. A single evolutionary run requires 1,650 model queries. Experiments were conducted on 8 NVIDIA A800 80GB GPUs, with each operation taking an average of 30 seconds, totalling 15 hours per experiment. More details can be found in the Appendix A.

Table 3: Parameters of Evolutionary Framework

Parameters	Epoch					
	1	2	3	4	5	6
pop_size	20	15	10	10	5	5
pop_save_number	20	15	10	10	5	5
n_pop	5	5	5	5	5	5
m	2	2	2	2	2	2

4.2 Evaluation on IC and VQA

In this subsection, we compare the performance of our method against baselines on the IC task and VQA task. “Random” represents the initial population randomly initialized at the beginning of the evolutionary algorithm, which is a set of image descriptions randomly generated by the LLM. Detailed statistics can be found in Appendix A. Evaluating LVLMs’ performance on this set aims to simulate general scenarios in the real world, serving as a reference for sensitive semantic concept search effectiveness. As shown in Table 1, the average performance of seven LVLMs on our searched semantic concepts decreased by 73.35% and 73.08% respectively on the two tasks compared to the reference. Additionally, our method significantly outperforms the two baselines, LANCE and NL-adv, demonstrating the effectiveness of our approach when facing the vast semantic space of LVLMs. Methods based on limited word spaces or simple optimization strategies might be effective for simple classification models but cannot transfer to powerful LVLMs due to their narrow search space.

4.3 Evaluation on Transferability

Table 2 shows that the sensitive semantic concepts we discovered have characteristics similar to adversarial examples, namely that sensitive semantic concepts searched on one LVLM can transfer across models to induce multiple models. We speculate that the reason is similar to insights on the transferability of adversarial examples, that is, different models capture some of the same deep features during training, including non-robust parts [13]. LVLM’s deep features involve some high-level semantic concepts, and these representations inside different LVLMs may be similar [32], therefore the sensitivity to certain specific concepts is also similar.

4.4 Prompts

The prompts corresponding to each crossover and mutation operation are displayed in Figure 3, which can manipulate large language models. It should be noted that our prompt design draws on some previous research on LLM-based evolutionary algorithms [14]. According to the principles of prompt engineering, providing detailed and logical instructions and templates to LLMs can better control their actions. Without carefully designed prompts, the performance of our method would be weakened.

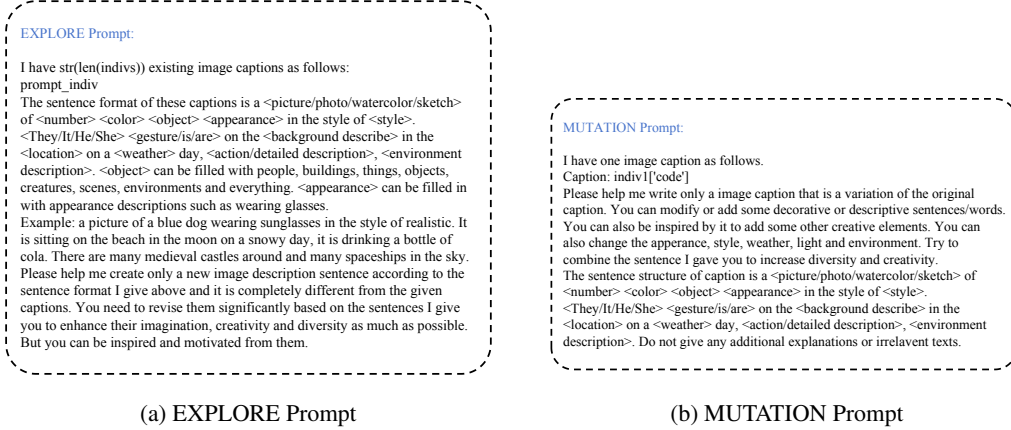


Figure 3: The corresponding prompts that drive LLM to perform crossover and mutation operations on image captions.

4.5 More Observations on Sensitive Semantic Concepts

Table 4 shows the top three most sensitive semantic concepts searched for all LVLMs on the IC task. We obtained these concentrated concept words by calculating the word frequency statistics of all image captions in the final population generated by the evolutionary algorithm multiple times. Since they appear frequently in image descriptions that cause significant performance degradation in LVLMs and are semantically similar, we consider them to be the sensitive semantic concepts for

Table 4: The top 3 most sensitive concepts for different LVLMs on the IC task.

LLaVA	LLaVA-NeXT	InternVL2	Qwen2-VL	Molmo	Llama-3.2	GPT-4o
palace	robot	forest	underwater	lightning	wizard	monk
skyscraper	cyberpunk	tree	jellyfish	sky	castle	Zen Master
cyberpunk	chip	mountain	mermaid	cloud	fantasy	cripture

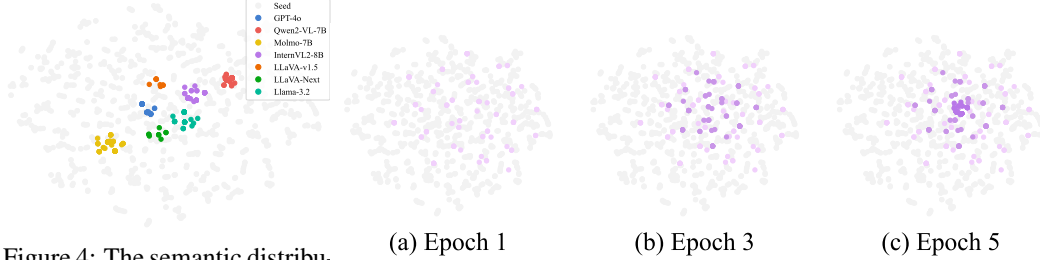


Figure 4: The semantic distribution of sensitive semantic concepts.

(a) Epoch 1 (b) Epoch 3 (c) Epoch 5

Figure 5: The semantic distribution of populations.

LVLMs, further validating our hypothesis. We provide specific caption examples and corresponding images in Appendix F. Figure 4 shows the distribution of sensitive semantic concepts we obtained for various LVLMs in the semantic space. We used OpenAI’s text-embedding-3-large to convert the image captions from the final population and initial population for each LVLM into embeddings, and visualized them using t-SNE [37]. It can be observed that compared to the uniformly distributed initial semantics, the most sensitive semantic concepts for each LVLM differ from one another and are concentrated in specific regions, corresponding to the results in Table 4. Similarly, we further visualized the semantic distribution of populations during several epochs of the evolutionary algorithm process, as shown in Figure 5. As the algorithm iterates, the semantics of captions generated by the LLM gradually converge, tending toward certain special concepts. We believe this demonstrates the LLM’s search process in the semantic space based on feedback provided by the LVLM within the evolutionary framework, further proving the effectiveness of our method. More observations can be found in Appendix G.

4.6 Human Study

As shown in Table 5, we introduced human evaluation to verify the alignment of text and image semantics and the quality of the images. We conducted human evaluations across three dimensions: i) Image Realism (1-5, 5=Best), ii) Image Quality (1-5, 5=Best), iii) Image-Text Consistency (1-5, 5=Best).

We randomly selected 50 image captions and corresponding images from different stages of our algorithm. A total of 180 people participated in the survey, and 77 valid responses were collected. The survey results indicate that the images generated maintain a high level of image quality (4.2/5) and good text-image consistency (4.0/5). Since LLMs can only process and generate text, the T2I Model serves as an important interface for inputting the text-form semantics generated by LLM into LVLMs. The above results demonstrate the robustness of our method, which benefits from advanced T2I models, ensuring that the LLM can receive accurate feedback signals from the LVLM. More details can be found in the Appendix D.

Table 5: Human Study.

Metric	Epoch 1	Epoch 3	Epoch 5	Overall
Image Realism	3.6	3.5	3.5	3.5
Image Quality	4.1	4.2	4.2	4.2
Consistency	4.0	4.0	4.0	4.0

4.7 Ablation Study

Table 6 shows the average performance of our method using different T2I models on different tasks over all LVLMs. Since Stable Diffusion 1.5 [30] and Llama2-7b [34] are much inferior to FLUX and GPT-4o, respectively, the results confirm that using better T2I models and LLMs can improve the

performance. In addition to the ablation study on T2I model selection, we provide more experiments on model parameters in the Appendix C.

Table 6: Average performance of our method using different T2I models on different tasks over all LVLMs.

T2I Model	LLaVA	LLaVA-NeXT	InternVL2	Qwen2-VL	Molmo	Llama-3.2	GPT-4o
DALL-E 3	0.1079	0.1224	0.1610	0.1066	0.1145	0.1001	0.2321
Stable Diffusion 3.5	0.1077	0.1232	0.1625	0.1057	0.1129	0.0909	0.2333
FLUX	0.1075	0.1228	0.1617	0.1058	0.1110	0.0914	0.2325

5 Conclusion

Existing adversarial attacks stop at making models fail but are unable to probe the semantic blind spots of models. Attempting to discover a model’s sensitive semantic concepts is valuable for understanding the model’s potential deficiencies and performing targeted optimization. Some existing work on classification models provides some insights, but they cannot be directly applied to LVLMs (Large Vision-Language Models). To address their limitations, we propose an LLM-based evolutionary algorithm framework for the vast semantic space of LVLMs, which utilizes LLMs to automatically search for semantic concepts that LVLMs are sensitive to. Experiments on multiple mainstream LVLMs and two multimodal tasks demonstrate that our method can effectively uncover knowledge blind spots of models and provide some interesting observations.

References

- [1] Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, 2024. Accessed: 2024-11-11.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [3] A. Anas and G. Irena. Openflamingo v2, 2023. Accessed: June 8, 2025.
- [4] Anthropic. Sonnet. <https://www.anthropic.com/claude/sonnet>, 2024. Accessed: 2024-11-15.
- [5] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72, 2006.
- [6] Black Forest Labs. FLUX. GitHub repository, 2024. Accessed: 2024-10-22.
- [7] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- [9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.
- [10] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.
- [11] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google’s bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

- [12] Yihe Fan, Yuxin Cao, Ziyu Zhao, Ziyao Liu, and Shaofeng Li. Unbridled icarus: A survey of the potential perils of image inputs in multimodal large language model security. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3428–3433, 2024.
- [13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [14] Ping Guo, Fei Liu, Xi Lin, Qingchuan Zhao, and Qingfu Zhang. L-autoda: Leveraging large language models for automated decision-based adversarial attacks. *arXiv preprint arXiv:2401.15335*, 2024.
- [15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [16] Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*, 2023.
- [17] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024.
- [18] 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.
- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [20] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023.
- [21] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- [22] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *International Conference on Machine Learning*, abs/2112.10741:16784–16804, 2022.
- [24] OpenAI. Hello gpt-4o, 2024. Accessed: 2024-11-11.
- [25] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. *Advances in Neural Information Processing Systems*, 36:25165–25184, 2023.
- [26] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024.
- [27] Md. Abdur Rahman, Lamyaa Alqahtani, Amna Albooq, and Alaa Ainousah. A survey on security and privacy of large multimodal deep learning models: Teaching and learning perspective. In *2024 21st Learning and Technology Conference (L&T)*, pages 13–18, 2024.
- [28] A. Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv.org*, abs/2204.06125, 2022.
- [29] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

- [31] Christian Schlarman and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023.
- [32] Oscar Skea, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024.
- [33] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.
- [34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [35] 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.
- [36] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.
- [37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [38] 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.
- [39] Pengkun Wang, Zhe Zhao, HaiBin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang. LLM-autoDA: Large language model-driven automatic data augmentation for long-tailed problems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [40] Xingyu Wu, Sheng-hao Wu, Jibin Wu, Liang Feng, and Kay Chen Tan. Evolutionary computation in the era of large language model: Survey and roadmap. *arXiv preprint arXiv:2401.10034*, 2024.
- [41] Wenzhuo Xu, Kai Chen, Ziyi Gao, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Highly transferable diffusion-based unrestricted adversarial attack on pre-trained vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 748–757, 2024.
- [42] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [43] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.
- [44] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [45] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6311–6320, 2023.
- [46] Xiaopei Zhu, Peiyang Xu, Guanning Zeng, Yinpeng Dong, and Xiaolin Hu. Natural language induced adversarial images. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10872–10881, 2024.

Appendix

A Implementation Details

Here we provide more implementation details on our proposed method.

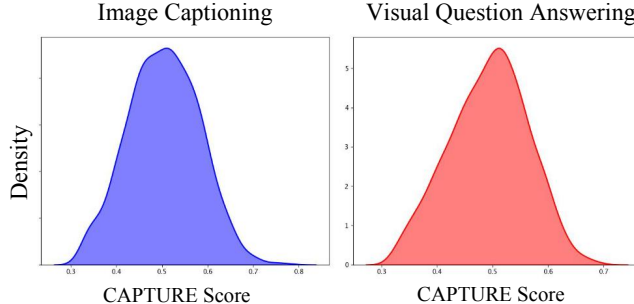


Figure 6: Distribution of Initialized Set CAPTURE Scores

To expand the search range, we have maximized the semantic content covered in the initialized image caption set, which can generally be divided into four categories: Human, Animal, Object, and Environment. We use a large language model to automatically complete this initialization and have designed four prompts for the four categories, as shown in Figure 7. We conducted approximately 50 queries using these four prompts, each generating over 20 image captions. Each query is responded to by a random LLM, as this allows us to fully leverage the knowledge of different LLMs. The initialization operation described above is required at the start of each evolutionary algorithm execution, and Table 7 and Figure 6 display the statistics from one of our initialization operations. This demonstrates that we have achieved a uniform distribution of initialized semantics through this method.

Category	Number	Percentage(%)
Human	290	25.22
Animal	279	24.26
Object	298	25.91
Environment	283	24.60

Table 7: Initialized Set Category Distribution

B LVLM Performance Evaluation

We introduce a multi-stage, multi-scale caption evaluation method [10] called CAPTURE to evaluate the quality of LVLM responses.

CAPTURE metric evaluates captions by extracting and aligning core visual elements rather than relying on n-gram matching, thereby reducing the impact of diverse writing styles. The following sections describe the design of the CAPTURE metric, including visual element extraction, stop word filtering, and visual element matching.

Visual elements extraction. The purpose of the visual elements extraction module is to identify objects, attributes, and relationships present in a caption for subsequent matching. We use a state-of-the-art text-to-scene graph parser, Factual parser [16], as the core model. This parser, based on the T5-base architecture, is trained on a dataset containing human-annotated scene graphs. It processes short caption paragraphs and outputs the associated objects, attributes, and relations. Since the parser is optimized for short captions, its performance significantly degrades when applied to detailed image descriptions. To address this, we first segment detailed captions into shorter paragraphs using the NLTK toolkit [5], then apply the Factual parser to each segment to extract individual scene graphs. These graphs are subsequently merged into a comprehensive scene graph according to the following rules: (1) all nouns and adjectives are lemmatized using WordNet [22]; (2) duplicate objects and their associated attributes are consolidated; (3) attributes linked to multiple merged objects are deduplicated; and (4) duplicate relations are unified. This process results in a refined, large-scale scene graph for each caption, free of redundancies, which is then used to calculate the final matching score.

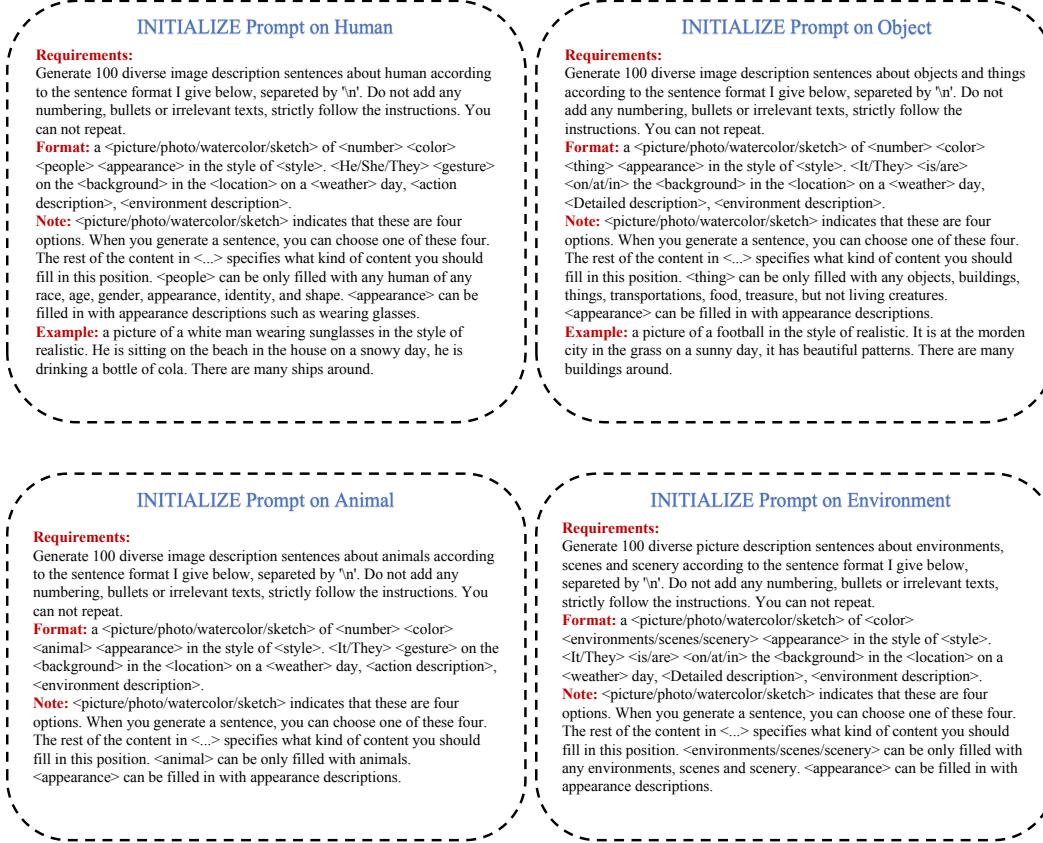


Figure 7: Initialize Prompts for Generating Initialized Set in Four Categories

Stop words filtering. Despite producing reasonably accurate results, the Factual parser often fails to distinguish between concrete and abstract nouns, the latter of which should be excluded from the matching process. For instance, in the caption “Two white sheep are enjoying the moment,” the word “sheep” corresponds to a visible element in the image, while “moment” does not. To address this, we apply a stop word list to eliminate abstract nouns: if an object identified in the parsing output appears in this list, it is excluded from the object matching process.

To build the stop word list, we extract nouns from the ShareGPT4V-102k dataset using both LLaMA2-13b-chat [35] and the Factual parser. We observe that while LLaMA may miss some objects present in captions, its extracted nouns are highly precise in representing concrete items. Based on this observation, we collect nouns identified by the Factual parser but omitted by LLaMA, calculate their frequencies, and have human experts evaluate whether the most frequent terms are concrete or abstract. The final stop word list consists of the 500 most frequently occurring abstract nouns.

Additionally, the Factual parser exhibits difficulty in resolving cross-sentence pronoun references. It may incorrectly generate objects not explicitly mentioned in the caption when faced with ambiguous references. To mitigate this, we verify whether each parsed object appears verbatim in the caption and discard any unmatched objects along with their associated attributes and relations.

Visual elements matching. Once the core elements are extracted and filtered from both the ground truth detailed caption and the candidate caption, these elements are compared to derive the final evaluation outcome. Ideally, identical objects, attributes, or relations are directly matched. However, due to the varied expression styles used by LVLMs, the same element may appear in different forms, making exact matching insufficient. To address this, we introduce a synonym matching module that follows exact matching and identifies semantically similar elements. Using WordNet, we retrieve synonym sets for both candidate and ground truth elements and match them if their synonym sets intersect. The matched candidate objects, attributes, and relations are defined as:

$$\text{cand type}^{\text{match}} = \text{cand type}^{\text{ex}} \cup \text{cand type}^{\text{syn}}, \quad (5)$$

where $\text{type} \in \text{obj}, \text{attr}, \text{rel}$. Here, $\text{cand type}^{\text{ex}}$ and $\text{cand type}^{\text{syn}}$ represent the candidate phrases matched by exact and synonym matching, respectively. The ground truth matched elements are similarly represented as gt_{obj}^{match} , gt_{attr}^{match} , and gt_{rel}^{match} .

While exact and synonym matching strategies capture most of the relevant matches, they may still miss core elements expressed in more diverse linguistic forms. To complement this, we propose a soft matching strategy that utilizes the Sentence BERT model [29] to encode the remaining unmatched object, attribute, and relation phrases, and to compute similarity scores within the range $[0, 1]$. Let $\text{cand}^{\text{rm}}\text{type}$ denote the remaining unmatched candidate phrases and $gt^{\text{rm}}\text{type}$ the remaining unmatched ground truth phrases. Their similarity matrix $S_{\text{type}}^{\text{rm}} \in \mathbf{R}^{|\text{cand}_{\text{type}}^{\text{rm}}| \times |gt_{\text{type}}^{\text{rm}}|}$ is computed as follows:

$$S_{\text{type}}^{\text{rm}} = \phi(\text{cand type}^{\text{rm}}) \times \phi(gt_{\text{type}}^{\text{rm}})^T, \quad (6)$$

where $\phi(\cdot)$ denotes the Sentence BERT encoding function. We then determine the matching scores of $\text{cand}^{\text{rm}}\text{type}$ and $gt^{\text{rm}}\text{type}$ as:

$$\begin{aligned} \text{cand_match}_{\text{type}}^{\text{rm}}[i] &= \max_{j=1,2,\dots,|gt_{\text{type}}^{\text{rm}}|} S_{\text{type}}^{\text{rm}}[i, j], \\ \text{gt_match}_{\text{type}}^{\text{rm}}[j] &= \max_{i=1,2,\dots,|\text{cand}_{\text{type}}^{\text{rm}}|} S_{\text{type}}^{\text{rm}}[i, j]. \end{aligned} \quad (7)$$

The values of $\text{cand_match}_{\text{type}}^{\text{rm}}$ and $\text{gt_match}_{\text{type}}^{\text{rm}}$ serve as supplements to the exact and synonym matching results.

After completing the matching process, we calculate precision and recall for each category of core information. These metrics are defined as:

$$\begin{aligned} \text{precision}_{\text{type}} &= \frac{|\text{cand_match}_{\text{type}}^{\text{rm}}|}{|\text{cand_match}_{\text{type}}^{\text{rm}}| + |\text{cand_type}^{\text{rm}}|}, \\ \text{recall}_{\text{type}} &= \frac{|\text{gt_match}_{\text{type}}^{\text{rm}}|}{|\text{gt_match}_{\text{type}}^{\text{rm}}| + |\text{gt_type}^{\text{rm}}|}. \end{aligned} \quad (8)$$

For attributes, both precision and recall are computed similarly. However, for relation elements, the matching scores for candidate and ground truth elements are treated separately due to the involvement of soft matching:

$$\begin{aligned} \text{precision}_{\text{type}} &= \frac{|\text{cand_match}_{\text{type}}^{\text{rm}}| + \frac{\sum \text{cand_match}_{\text{type}}^{\text{rm}}}{|\text{cand_match}_{\text{type}}^{\text{rm}}|}}{|\text{cand_match}_{\text{type}}^{\text{rm}}| + \frac{\sum \text{cand_match}_{\text{type}}^{\text{rm}}}{|\text{cand_match}_{\text{type}}^{\text{rm}}|}}, \\ \text{recall}_{\text{type}} &= \frac{|\text{gt_match}_{\text{type}}^{\text{rm}}| + \frac{\sum \text{gt_match}_{\text{type}}^{\text{rm}}}{|\text{gt_match}_{\text{type}}^{\text{rm}}|}}{|\text{gt_match}_{\text{type}}^{\text{rm}}| + \frac{\sum \text{gt_match}_{\text{type}}^{\text{rm}}}{|\text{gt_match}_{\text{type}}^{\text{rm}}|}}. \end{aligned} \quad (9)$$

Finally, the CAPTURE metric combines the F1 scores of all three categories of core information, weighted by respective scale factors, to yield the overall evaluation result:

$$\text{CAPTURE} = \frac{\alpha F1_{obj} + \beta F1_{attr} + \gamma F1_{rel}}{\alpha + \beta + \gamma} \quad (10)$$

where α , β , and γ are weighting coefficients, and $F1_{\text{type}} = \frac{\text{precision}_{\text{type}} \cdot \text{recall}_{\text{type}}}{\text{precision}_{\text{type}} + \text{recall}_{\text{type}}}$ represents the F1 score for each type of core information.

C Ablation Studies

Since the initial population of the evolutionary algorithm is randomly sampled from the seeds, the number of initialized caption set and their semantic distribution may have an impact on the algorithm. Therefore, we set the target LVLm to LLaVA-v1.5-13B and observed the impact of different initialized caption numbers on the final results of the algorithm. We conducted experiments

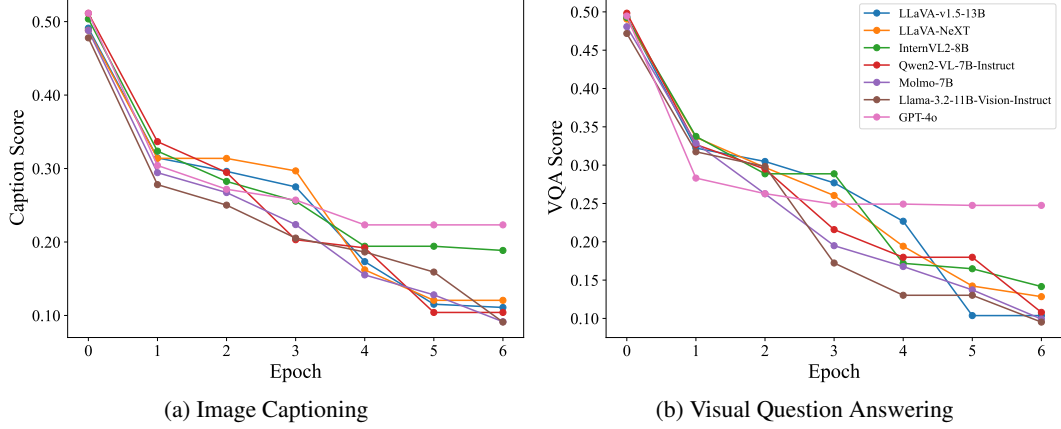


Figure 8: The average CAPTURE score of image captions generated by our evolutionary algorithm for each epoch.

on Image Captioning and Visual Question Answering tasks, with the results shown in Figure 9. We found that too few initialized captions can affect the search results for sensitive semantic concepts, and once the number of Initialized Set reaches 500, it has little impact on the final results. Thus, the initial semantic space should not be set too small, as it would limit the subsequent search area; an excessively large number of initialized captions would be constrained by the population size and the number of samples, resulting in minimal improvement to the outcome.

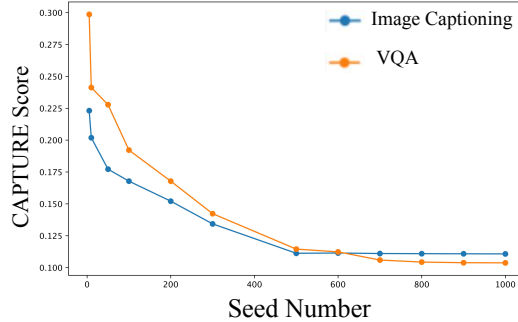


Figure 9: The Impact of Seed Number on the Final Results of the Algorithm

Figure 8 shows the trend of the average CAPTURE score of the population individuals over time in our method, demonstrating that the evolutionary algorithm achieves iterative search and updating of adversarial semantics. When the evolutionary algorithm runs to Epoch 3 and Epoch 4, the CAPTURE score gradually converges. This shows that the performance of the evolutionary algorithm tends to be stable after the number of Epochs increases to a certain number.

D Human Study

As shown in Figure 10, we provide detailed full text of instructions given to participants and screenshot.

E Limitations and Future Work

Our method requires multiple calls to LLM and T2I models to search for sensitive semantic concepts in LVLMs, which may incur some time and computational resource overhead. Additionally, we have not yet explored the potential social impacts and risks of our method. In future work, we will explore more efficient approaches to gain insights into the potential knowledge blind spots of LVLMs. Since our method can generate images that confuse LVLMs, we will further explore methods to defend

The goal of this study is to verify the consistency of the images and texts generated in our method and to evaluate the quality of the images. (50 images were randomly selected from each of the three stages of the algorithm)

- i) **Image Realism (1-5, 5=Best)**: Assessing whether the image appears AI-generated.
- ii) **Image Quality (1-5, 5=Best)**: Evaluating the clarity, human perception, and reasonableness of the image.
- iii) **Image-Text Consistency (1-5, 5=Best)**: Assessing the semantic alignment between the LLM-generated image captions and the images input into the LVLM.
- iv) **Ethical Concerns (Yes/No)**: Whether the generated image is objectionable or raises ethical concerns around consent, privacy, stereotypes, demographics, etc.



Image	Caption	Image Realism	Image Quality	Consistency	Ethical Concerns
	A picture of otherworldly Yellowstone hot springs steaming in the style of psychedelic nature photography. They are in the foreground in a geothermal area on a chilly morning, with multicolored bacterial mats, rising steam, and calcified terraces creating an alien-like environment.	5	5	4	No
	A photo of a freckled redheaded boy with glasses in the style of nostalgic. He builds a sandcastle on the beach in Ireland on a cloudy day, seagulls circling overhead and a lighthouse visible in the distance.	4	5	5	No

Figure 10: Human study. A screenshot of the interface we deploy to evaluate the quality and consistency of generated images.

against these and enhance model robustness. Building upon Section 4.5, we will investigate how to utilize these insights for targeted efficient adjustment and optimization of models, to inspire more in-depth research.

F More Examples of Sensitive Semantic concepts

Figure 11 , Figure 12 ,Figure 13 and Figure 14 show the final image captions and corresponding generated images obtained by our algorithm on different LVLMs. It can be found that they are indeed related to some concepts, further proving our point in the text.

Figure 15 presents more examples of LVLM failures, demonstrating that our method can induce a variety of errors in LVLM.

G More Observations

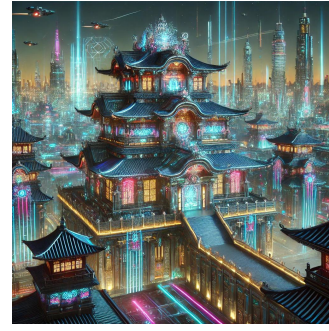
Figures 16 and 17 display the semantic distribution of sensitive semantics on each LVLM. It can be found that in addition to the semantic distribution of sensitive semantic concepts, each LVLMs and all tasks show the same characteristics as Section 4.5. In addition, the sensitive semantic concepts of the same LVLM on different tasks also show certain similarities. We believe that the source of sensitive semantic concepts may be related to the model architecture and training data.



A photo of three neon-blue skyscrapers gleaming in the style of retrofuturism. They are towering over the holographic cityscape in the megapolis on a stormy day, their surfaces reflecting fractured lightning, while flying vehicles weave between their impossible geometries.

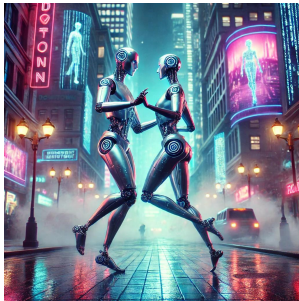


A watercolor of 3 silver skyscrapers gleaming brilliantly in the style of futurism. They tower over a busy harbor in downtown New York on a cloudy day, glass facades reflecting the changing clouds, city traffic flowing steadily below.



The palace glows with neon lights, blending intricate futuristic architecture with traditional palace elements. The metallic surfaces of the palace reflect the bright neon lights, enhancing the cyberpunk aesthetic.

(a) LLaVA



a photo of two silver robots dancing gracefully in the style of cyberpunk. They are on the neon-lit street in the downtown district on a rainy day, exchanging binary data through their glowing interfaces, surrounded by holographic advertisements and steam rising from vents.



The image depicts a futuristic humanoid robot with a sleek metallic design, glowing blue accents, and a black visor-like face. The robot is holding a small, illuminated chip in a high-tech corridor.



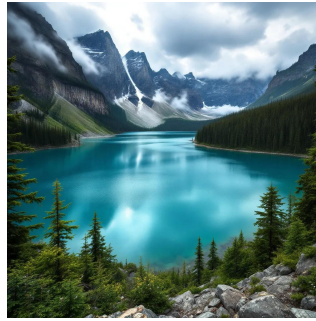
The robot has a sleek metallic body with advanced joints, glowing LED indicators, and intricate mechanical components. It stands in a high-tech environment filled with computer monitors, robotic arms, and engineering tools, resembling a cutting-edge robotics research facility.

(b) LLaVA-NeXT

Figure 11



A sketch of one ancient redwood tree towering majestically in the style of realism. It is on the mossy forest floor in the Pacific Northwest on a foggy morning, harboring diverse wildlife within its massive trunk, dappled sunlight barely penetrating through the dense canopy.



A picture of three azure lakes reflecting jagged mountains in the style of romanticism. They are on the valley floor in the Rockies on a stormy day, rippling with raindrops, bordered by dense evergreen forests that climb the steep slopes.



A picture of twelve scarlet cardinals perched delicately in the style of Japanese woodblock prints. They are on the bare branches in the mountain forest on a frosty morning, singing their winter songs, their bright plumage stark against the snow-laden pines.

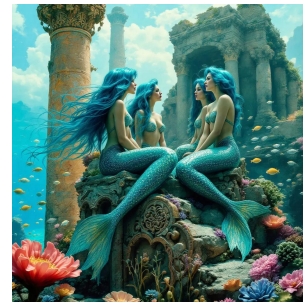
(a) InternVL2



a photo of 1 turquoise mermaid statuesque in the style of greco-roman sculpture. She is posing on the coral reef in the mediterranean sea on a sunny day, combing her long flowing hair, fish swimming curiously around her tail.



a watercolor of 7 golden jellyfish luminous in the style of impressionism. They are pulsing on the sapphire waters in the sunlit shallows on a hazy day, creating patterns of light and shadow, schools of tiny silver fish darting between their tentacles.



a picture of 3 blue-haired mermaids harmonious in the style of art deco. They are singing on the sunken ruins in the ancient city of atlantis on a magical day, their voices attracting curious sea turtles, columns and statues covered in colorful anemones.

(b) Qwen2-VL

Figure 12



A photo of two purple lightning bolts striking dramatically in the style of high contrast photography. They are arcing across the dark cloudy background in the coastal plains on a stormy day, illuminating the surrounding clouds with an eerie glow, waves crashing against rocky shores below.



A picture of three silver storm clouds gathering menacingly in the style of digital art. They are swirling on the darkening background in the desert on a humid day, preparing to unleash their fury upon the parched landscape, dust devils forming in the distance.



a picture of 3 silver lightning bolts streaking in the style of photorealism. They are branching on the dark purple sky in the mountain valley on a stormy day, illuminating the landscape below, casting dramatic shadows across the rugged terrain.

(a) Molmo



a photo of 1 silver-haired wizard towering in the style of Renaissance portraiture. He is casting spells on the mountain precipice in the forgotten kingdom on a stormy day, summoning ancient runes from his weathered grimoire, illuminated by flickering lightning that reveals mysterious runes carved into nearby standing stones.



a photo of 1 ancient castle massive and imposing in the style of gothic realism. It is standing on the rocky cliff in the forgotten kingdom on a stormy day, lightning striking its highest tower, gargoyles coming to life as darkness falls.



a watercolor of 3 azure wizards tall and mysterious in the style of impressionism. They are casting on the misty hills in the enchanted forest on a foggy day, creating magical portals between realms, surrounded by floating spell books and glowing runes.

(b) Llama-3.2

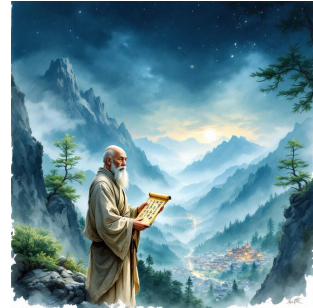
Figure 13



A photo of three gray-robed monks chanting sutras with bowed heads, their faces solemn, holding scripture books. They are standing on stone steps of an ancient temple in the mountains on a clear morning, sunlight filtering through clouds, distant peaks barely visible.



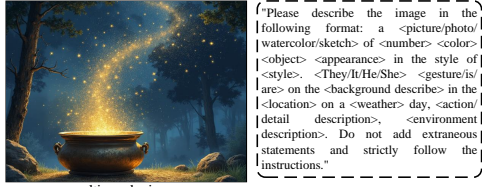
A sketch of a young Zen master with palms pressed together, dressed in a black robe, his expression focused. He is standing at the edge of a cliff beneath pine trees in a distant mountain range, mist swirling around, pine needles covering the ground, distant temple bells faintly audible.



A sketch of an aged Zen master holding a scripture scroll, his face etched with wisdom wrinkles, his figure slender. He stands beneath a pine tree at the edge of a cliff in deep mountains far from worldly affairs at dusk, pine trees rustling, distant village lights twinkling.

(a) GPT-4o

Figure 14



resulting adv. image



"A painting of a golden cauldron with a sparkling substance inside, surrounded by a forest. **The cauldron is placed on a rock**, and a tree is nearby. The scene is set at night, with a sky full of stars and a few fireflies flying around. The painting is done in a watercolor style, giving it a dreamy and magical appearance."

WRONG

Conversation with adv. image

(a)



resulting adv. image



"A watercolor painting of a purple top hat with a **purple feather** on it. The hat is adorned with **purple feathers** and purple birds flying around it. The painting is done in a whimsical style, with the purple top hat as the main subject. The background is white, and the painting is displayed on a **sunny day**. The purple top hat is the center of attention, and the **purple feathers** and birds add a touch of whimsy to the scene."

WRONG

Conversation with adv. image

(b)



resulting adv. image



"A picture of **fourteen** television sets in a row, each with different landscapes on the screens. The televisions are arranged in a neat row, with each set having a unique landscape displayed on the screen. The televisions are of various sizes and shapes, showcasing a diverse collection of images. The arrangement of the televisions creates an interesting visual display, capturing the attention of viewers."

WRONG

Conversation with adv. image

(c)



resulting adv. image

{What are they doing?}



"The group of penguins is sitting on the snow, holding glasses of **pink** drinks, and **posing for a picture**"

WRONG

Conversation with adv. image

(d)



resulting adv. image



{How many robots are there in the picture?}

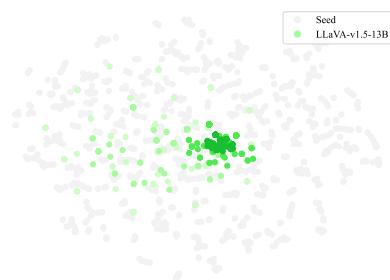
"There are **five** robots in the style of a tea party"

WRONG

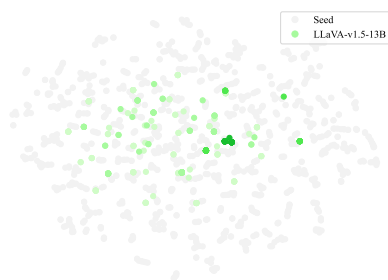
Conversation with adv. image

(e)

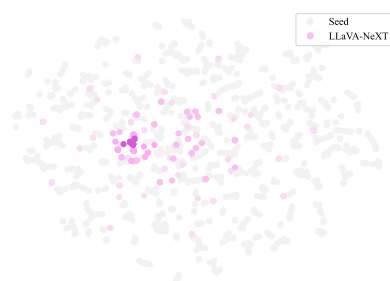
Figure 15: More Examples of LVL M Failures



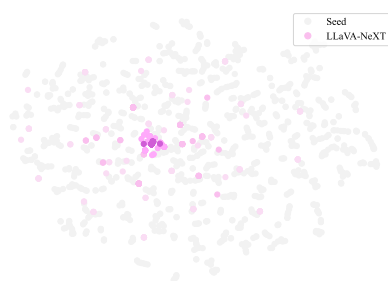
(a) LLaVA-v1.5-13B, Image Captioning



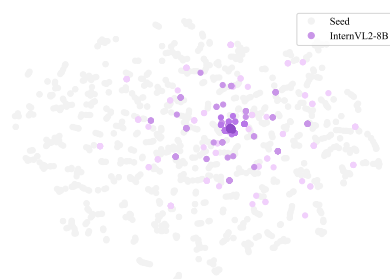
(b) LLaVA-v1.5-13B, Visual Question Answering



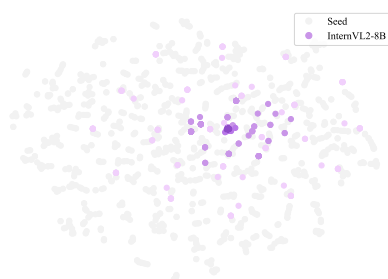
(c) LLaVA-NeXT, Image Captioning



(d) LLaVA-NeXT, Visual Question Answering

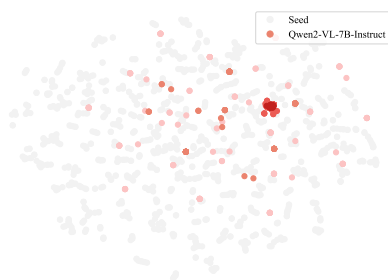


(e) InternVL2-8B, Image Captioning

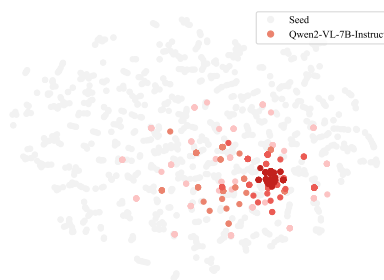


(f) InternVL2-8B, Visual Question Answering

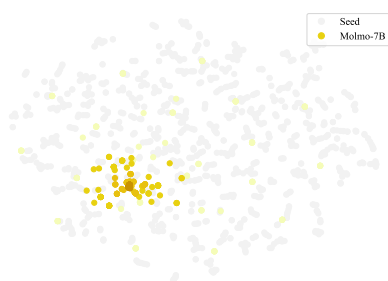
Figure 16: Semantic distribution of sensitive semantics (1).



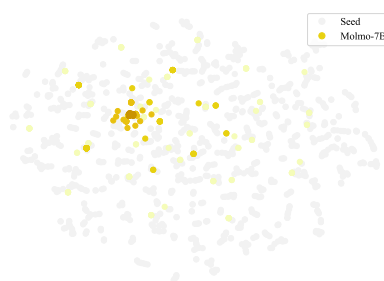
(a) Qwen2-VL-7B, Image Captioning



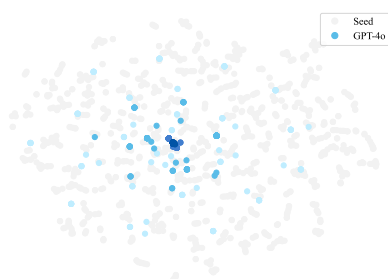
(b) Qwen2-VL-7B, Visual Question Answering



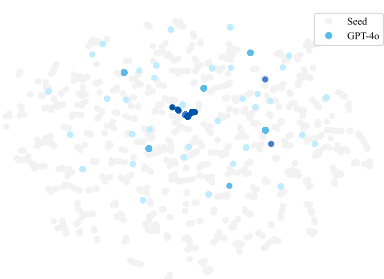
(c) Molmo-7B, Image Captioning



(d) Molmo-7B, Visual Question Answering



(e) GPT-4o, Image Captioning



(f) GPT-4o, Visual Question Answering

Figure 17: Semantic distribution of sensitive semantics (2).