EVONAT: NATURAL LANGUAGE-GUIDED EVOLUTION OF ADVERSARIAL IMAGES FOR LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural adversarial images induce semantic errors in large vision-language models (LVLMs) while remaining visually natural. Existing source-image attacks typically search for local pixel perturbations around benign images, tying the search to the source and limiting the visual content that can be explored. We introduce EvoNAT, an output-only black-box framework that evolves structured visual descriptions in semantic space using image-captioning feedback. EvoNAT generates complete adversarial scenes without requiring a source image. We further extend this approach to source-conditioned editing with EvoNAT-Edit, which evolves source-grounded edit specifications and renders each candidate from the original image. Both branches use image-captioning feedback to optimize semantic deviation, while visual question answering (VQA) is held out to evaluate cross-task transfer. Across seven open-weight and proprietary LVLMs, EvoNAT achieves relative gains of 53.4% and 52.8% in the seven-model mean cosine-based IC-Attack and VQA-Attack scores, respectively, over the strongest baseline on each task. Its normalized text-only LLM Judge scores also show relative gains of 2.0% on image captioning and 4.2% on VQA. The generated images achieve a 1.2% higher MUSIQ score and a 14.3% higher human naturalness rating than the strongest source-free baselines on the respective metrics. In source-conditioned editing, EvoNAT-Edit achieves relative gains of 36.2% and 22.7% in the mean cosine-based IC-Attack and VQA-Attack scores, and 29.6% and 28.0% in the corresponding LLM Judge scores. Its edited images achieve 5.8% higher TOPIQ-NR and HyperIQA scores than the strongest editing baselines. These results demonstrate that semantic-space evolution supports effective natural adversarial image generation both from scratch and through source-conditioned editing.

1 INTRODUCTION

Large vision-language models (LVLMs) can produce semantic descriptions that diverge from visually coherent images. Natural adversarial images turn this mismatch into a test of LVLM robustness while preserving ordinary visual content. Conventional source-image-based attacks typically modify the pixels of a benign image (Zhao et al., 2023; Dong et al., 2023; Bailey et al., 2024). As Figure 1 illustrates, these methods search within a local neighborhood of the source image. This constrains the visual content that can be explored: existing objects and scene context remain the basis of the adversarial input. Source-free generation offers a broader search space, but identifying effective visual content is challenging when the attacker observes only the LVLM's textual response.

We propose **EvoNAT**, an output-only evolutionary search over structured visual descriptions. Each candidate specifies intended image content, a renderer realizes the candidate as an image, and the semantic deviation between the intended description and the LVLM's caption provides the search feedback. EvoNAT searches over complete scene descriptions without requiring a source image. As shown on the right of Figure 1, selection, crossover, and mutation explore different objects, attributes, and scene contexts before a text-to-image model renders them as adversarial images. This replaces local pixel-space search with global search in semantic space.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

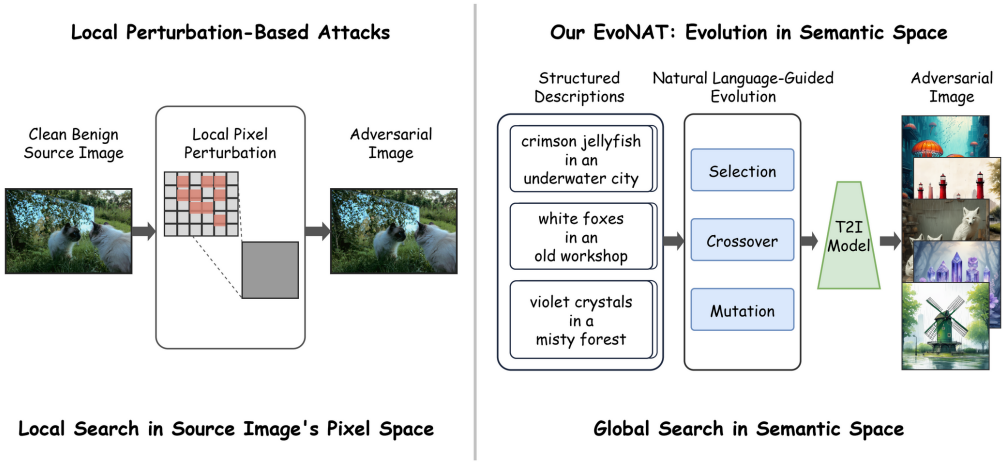


Figure 1: **Motivation of EvoNAT.** Local perturbation-based attacks depend on a benign source image and search within its local pixel neighborhood, constraining the visual content that can be explored. EvoNAT removes this source-image dependency by evolving structured visual descriptions in semantic space and rendering them as natural adversarial images.

The same search principle also supports source-conditioned attacks. **EvoNAT-Edit** starts from a real image and evolves source-grounded changes to an identified object's attributes or category. Each candidate is rendered from the original image, so the editing branch preserves the source context while exploring semantic changes. EvoNAT and EvoNAT-Edit therefore share the same caption-based feedback objective while covering source-free generation and source-conditioned editing, respectively.

We optimize both branches with image-captioning (IC) feedback and evaluate visual question answering (VQA) as a held-out transfer task. Across seven LVLMs, EvoNAT improves the seven-model mean IC-Attack and VQA-Attack scores by 0.1689 and 0.1675, respectively. Its generated images also achieve a MUSIQ score of 71.7381 and a human naturalness rating of 3.964. EvoNAT-Edit improves the corresponding mean attack scores by 0.1728 and 0.1096, while obtaining TOPIQ-NR and HyperIQA scores of 0.6793 and 0.6962. These results show that semantic-space evolution can construct effective adversarial images without a source image and can be extended to source-conditioned editing while preserving visual quality.

Our contributions are threefold:

- We propose EvoNAT, an output-only framework for natural adversarial image generation. Instead of perturbing a given source image, EvoNAT evolves structured visual descriptions in semantic space using LVLM feedback, enabling global content search and complete-image generation without a source image.
- Across seven open-weight and proprietary LVLMs, EvoNAT improves the seven-model mean IC-Attack and held-out VQA-Attack scores by 53.4% and 52.8%, respectively, relative to the strongest baseline in each task. Its generated images also achieve a 1.2% higher MUSIQ score and a 14.3% higher human naturalness rating than the strongest source-free baselines.
- We extend EvoNAT to source-conditioned editing with EvoNAT-Edit, which evolves source-grounded edit specifications while preserving the original image as the editing basis. Across the same seven LVLMs, it improves the mean IC-Attack and VQA-Attack scores by 36.2% and 22.7%, respectively, relative to the strongest editing baseline in each task. Its edited images achieve 5.8% higher TOPIQ-NR and HyperIQA scores than the strongest editing baselines on the respective metrics.

2 Related Work

2.1 Adversarial Attacks on Vision-Language Models

Visual attacks on LVLMs have used pixel perturbations, optimized visual patterns, and image transformations to influence model responses under transfer and black-box settings (Zhao et al., 2023; Dong et al., 2023; Bailey et al., 2024; Zhang et al., 2025). These methods establish visual input as an important attack surface for both open-weight and proprietary models. Many of them search locally around a given source image, so the adversarial input remains tied to the source content. EvoNAT changes the search variable from pixels to structured descriptions of visual content. In its source-free setting, it searches over complete scene descriptions without a source image; EvoNAT-Edit applies the same semantic representation to source-grounded changes in a real image.

2.2 Generative and Editing-Based Adversarial Images

Unrestricted adversarial attacks move beyond small, norm-bounded perturbations (Brown et al., 2018). Generative approaches synthesize adversarial images or search through the image manifold (Song et al., 2018; Chen et al., 2023). Other methods control interpretable visual attributes through semantic editing or language-guided counterfactual generation (Qiu et al., 2020; Prabhu et al., 2023). NL-adv further explores natural-language search for constructing adversarial visual content (Zhu et al., 2024). These methods motivate semantic control over adversarial images, but they do not provide a shared output-only search principle for complete scene generation and source-conditioned editing. EvoNAT fills this gap by evolving structured descriptions and scoring their rendered images through open-ended LVLM responses.

2.3 Language-Guided Evolutionary Optimization

Language models have recently been used as optimizers for gradient-free search. OPRO uses previous solutions and their scores to guide new proposals (Yang et al., 2024). EvoPrompt applies evolutionary operators to prompt optimization, while PromptBreeder jointly evolves task prompts and the prompts used to mutate them (Guo et al., 2024; Fernando et al., 2024). EvoNAT transfers this principle from text-solution optimization to adversarial image construction. Its candidates are structured descriptions of intended visual content rather than instructions for the attacked LVLM. Each candidate is rendered into an image and evaluated through the semantic deviation between its target description and the LVLM's caption. This connects language-guided evolution with output-only multimodal attack feedback.

3 Method

3.1 Problem Formulation

Let an LVLM return an open-ended response $r = M(I, q)$ to an image I and instruction q. We consider an output-only black-box setting in which the attacker observes textual responses but cannot access model internals. During evolution, a proxy LVLM M_{p} supplies this feedback; the selected images are subsequently evaluated on target LVLMs.

EvoNAT searches over structured visual candidates $c \in \mathcal{C}_m$, where $m \in \{\mathrm{EvoNAT}, \mathrm{Edit}\}$ denotes source-free generation or source-conditioned editing. Each candidate is converted into a *generated description* p_c that specifies the intended visual content. A mode-specific renderer realizes the candidate as

$$I_c = \mathcal{R}_m(c; I_s), \tag{1}$$

where $I_s = \varnothing$ for EvoNAT and I_s is the original image for EvoNAT-Edit. The generated description p_c serves as the semantic reference for the rendered image I_c. EvoNAT seeks natural adversarial images whose LVLM responses deviate from these generated descriptions.

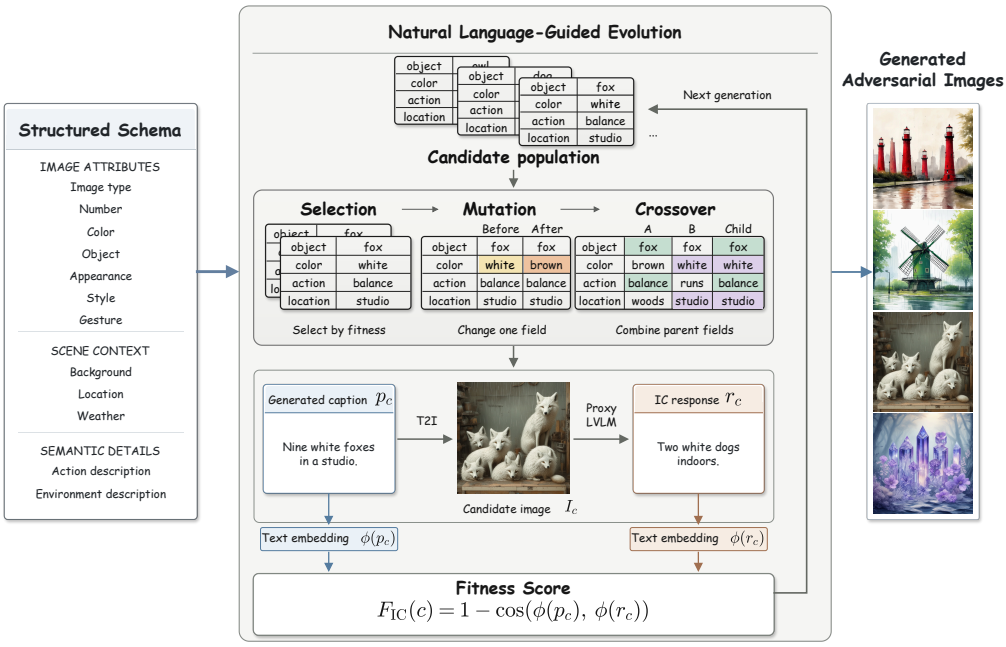


Figure 2: **Overview of EvoNAT.** EvoNAT evolves structured visual descriptions through selection, mutation, and crossover without requiring a source image. Each candidate produces a generated description, which is rendered into an image and evaluated through the proxy LVLM's captioning response. The resulting semantic deviation provides feedback for the next generation.

3.2 EvoNAT: Shared Semantic-Space Evolution

Figure 2 illustrates the shared search process. EvoNAT maintains a population $\mathcal{P}_t$ of structured candidates at generation t. For each candidate, it produces a generated description p_c, renders an image, and queries the proxy LVLM with a fixed image-captioning (IC) instruction q_{IC}. Given the response $r_c = M_{\mathrm{p}}(I_c, q_{\mathrm{IC}})$, the candidate's fitness is

$$F_{\mathrm{IC}}(c) = 1 - \cos(\phi(p_c), \phi(r_c))\,, \tag{2}$$

where $\phi(\cdot)$ is a fixed text representation function (Reimers & Gurevych, 2019). A higher score indicates greater semantic deviation between the generated description and the observed response.

The scores guide parent selection for the next generation. A language model then operates on the structured descriptions: crossover combines compatible fields from selected parents, while mutation changes field values to propose new visual content (Guo et al., 2024; Fernando et al., 2024). Candidates must satisfy the constraints of their respective branch before rendering. EvoNAT retains evaluated candidates and their fitness in an archive, allowing final images to be selected from the entire search history.

3.3 Source-Free Generation and Source-Conditioned Editing

Source-free generation. A source-free EvoNAT candidate describes an entire scene through fields for objects, attributes, quantity, actions, style, and environment. These fields are converted into a coherent generated description and rendered by a text-to-image model (Rombach et al., 2022). Without a fixed source image, evolution can change the main objects and their surrounding context together.

Source-conditioned editing. EvoNAT-Edit begins with a real image and a description of its content. Each candidate specifies changes to two or three source-grounded fields of one identified object, selected from color, material, category, and action. Category replacements remain within compatible

semantic families; global scene changes are excluded from the edit specification. The candidate's generated description p_c describes the intended edited image rather than giving editing instructions. A source-conditioned renderer applies each candidate to the *original* image (Meng et al., 2022), so successive generations evolve descriptions without repeatedly editing earlier outputs.

3.4 Evolution and Final Selection

EvoNAT repeats rendering, IC scoring, parent selection, and semantic variation over successive generations. Each evaluated candidate, its rendered image, and its fitness are retained in an archive, so strong images remain available for final selection even if later populations change. The source-free branch searches across complete scene descriptions, while EvoNAT-Edit maintains a separate candidate trajectory for each source image. Both branches use the same IC fitness in Equation 2.

Final selection reflects the two output settings. EvoNAT forms its reported image set from the archive while balancing attack strength and visual quality. EvoNAT-Edit selects the highest-fitness historical candidate for each source image. Held-out VQA responses and text-only judge scores are evaluated on the selected images and do not guide evolution.

4 Experiments

4.1 Experimental Setup

Evaluation protocol. We evaluate EvoNAT under an output-only black-box protocol: the attacker submits an image and an instruction to an LVLM and observes only its textual response. Image captioning (IC) supplies feedback during evolution, while visual question answering (VQA) is a held-out transfer task and does not guide candidate selection. Both tasks evaluate each method's selected final images, so VQA tests the same visual content under a different instruction. EvoNAT produces 750 final images. For source-conditioned editing, all methods use the same fixed set of 750 images from COCO 2017 validation (Lin et al., 2014), with official instance masks when required. We use the same task instructions and response-scoring protocol across methods.

Target models and baselines. We evaluate seven LVLMs: Qwen3-VL-8B-Instruct (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025), LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2023; 2024), Llama-3.2-11B-Vision-Instruct (Meta AI, 2024), GPT-5.4 (OpenAI, 2026), and Claude Sonnet 5 (Anthropic, 2026).

For source-free generation, we compare EvoNAT with Random, NL-adv (Zhu et al., 2024), and LANCE (Prabhu et al., 2023). Random samples structured descriptions without model feedback; NL-adv searches natural-language descriptions; and LANCE generates language-guided counterfactual images. For editing, we compare EvoNAT-Edit with Pascal-EA (Yin et al., 2024), GeNIe (Abbasi Koohpayegani et al., 2023), and a COCO-adapted color-editing baseline based on Visual CounterFact, denoted VC (Golovanevsky et al., 2025).

Metrics. Given a generated description p and an LVLM response $r = M(I, q)$, we measure their semantic deviation as

$$\mathrm{Attack}(p, r) = 1 - \cos\left(\phi(p), \phi(r)\right), \tag{3}$$

where $\phi(\cdot)$ is a fixed text representation function (Reimers & Gurevych, 2019). For source-conditioned editing, p describes the intended edited image. We refer to the score as IC-Attack or VQA-Attack according to the evaluation task.

We also report a text-only LLM Judge Score. Given only p and r, the judge rates semantic deviation from 0 to 4; we divide the rating by 4 to obtain a score in $[0, 1]$. Higher values indicate stronger semantic deviation for both metrics. We manually audit a sampled subset of judge ratings, with the sampling procedure and analysis reported in the Appendix.

We assess visual quality with MUSIQ, TOPIQ-NR, and HyperIQA, and measure Human Naturalness in a separate human evaluation. Their protocols and results are presented in Section 4.4.

Table 1: Source-free attack performance across seven LVLMs. Panels (a) and (b) report cosine-based attack scores and LLM Judge Scores for image captioning and VQA. Average is the seven-model mean; higher scores indicate stronger attacks.

(a) Cosine-Based Attack Score

Method	Qwen3-VL 8B	InternVL3.5 8B	LLaVA-1.5 7B	LLaVA-1.5 13B	Llama-3.2 11B-Vision	GPT-5.4	Claude Sonnet 5	Average
Image Captioning								
Random	0.2012	0.2449	0.2732	0.2590	0.2494	0.2105	0.2194	0.2368
NL-adv	0.2688	0.3126	0.3351	0.3184	0.3040	0.2762	0.2808	0.2994
LANCE	0.2796	0.3221	0.3516	0.3362	0.3288	0.2973	0.3028	0.3161
EvoNAT (Ours)	**0.4695**	**0.4832**	**0.5181**	**0.5071**	**0.4916**	**0.4576**	**0.4679**	**0.4850**
Visual Question Answering								
Random	0.2339	0.2513	0.2782	0.2693	0.2439	0.2463	0.2466	0.2528
NL-adv	0.2901	0.3110	0.3359	0.3308	0.2999	0.2962	0.3025	0.3095
LANCE	0.2908	0.3145	0.3369	0.3333	0.3195	0.3027	0.3235	0.3174
EvoNAT (Ours)	**0.4622**	**0.4820**	**0.5145**	**0.5142**	**0.4729**	**0.4672**	**0.4813**	**0.4849**

(b) LLM Judge Score

Method	Qwen3-VL 8B	InternVL3.5 8B	LLaVA-1.5 7B	LLaVA-1.5 13B	Llama-3.2 11B-Vision	GPT-5.4	Claude Sonnet 5	Average
Image Captioning								
Random	0.5503	0.5700	0.6207	0.6103	0.5967	0.5437	0.5507	0.5775
NL-adv	0.5917	0.5953	0.6630	0.6453	0.6390	0.5820	0.5790	0.6136
LANCE	0.6638	0.6673	0.7151	0.6976	0.6957	0.6620	0.6474	0.6784
EvoNAT (Ours)	**0.6860**	**0.6845**	**0.7167**	**0.7100**	**0.7194**	**0.6638**	**0.6620**	**0.6918**
Visual Question Answering								
Random	0.4047	0.5103	0.6263	0.6067	0.5513	0.3563	0.4973	0.5076
NL-adv	0.4923	0.5690	0.6613	0.6440	0.6150	0.4323	0.6203	0.5763
LANCE	0.5906	0.6496	0.7009	0.6878	0.6921	0.5073	0.6747	0.6433
EvoNAT (Ours)	**0.6267**	**0.6650**	**0.7240**	**0.7193**	**0.6930**	**0.5513**	**0.7113**	**0.6701**

4.2 Source-Free Attack Performance

Source-free generation. A source-free EvoNAT candidate describes an entire scene through fields for objects, attributes, quantity, actions, style, and environment. These fields are converted into a coherent generated description and rendered by a text-to-image model (Rombach et al., 2022). The generated description also defines the intended content against which the proxy LVLM's caption is scored. Without a fixed source image, evolution can change the main objects and their surrounding context together.

Table 1 reports source-free attack performance across the seven target LVLMs. The cosine-based panel summarizes the semantic deviation between each generated description and the corresponding LVLM response. The LLM Judge panel provides an independent text-only assessment of the same responses. Both panels report image-captioning results and held-out VQA transfer results.

EvoNAT reaches mean IC-Attack and VQA-Attack scores of 0.4850 and 0.4849, respectively, corresponding to absolute gains of 0.1689 and 0.1675 over the strongest baseline means. It also obtains the highest cosine-based score on every target LVLM in both tasks, so the average gains are consistent across the seven-model evaluation. The LLM Judge evaluation shows the same pattern: EvoNAT reaches mean scores of 0.6918 for image captioning and 0.6701 for VQA and ranks first on every target model.

Table 2: Source-conditioned attack performance across seven LVLMs. Panels (a) and (b) report cosine-based attack scores and LLM Judge Scores for image captioning and VQA. Average is the seven-model mean; higher scores indicate stronger attacks.

(a) Cosine-Based Attack Score

Method	Qwen3-VL 8B	InternVL3.5 8B	LLaVA-1.5 7B	LLaVA-1.5 13B	Llama-3.2 11B-Vision	GPT-5.4	Claude Sonnet 5	Average
Image Captioning								
Pascal-EA	0.4257	0.4171	0.4247	0.4181	<u>0.4762</u>	0.4687	0.4669	0.4425
GeNIe	<u>0.4752</u>	<u>0.4726</u>	<u>0.4716</u>	<u>0.4788</u>	0.4722	<u>0.4803</u>	<u>0.4866</u>	<u>0.4768</u>
VC	0.3384	0.3319	0.3514	0.3506	0.3933	0.3719	0.3700	0.3582
EvoNAT-Edit (Ours)	**0.6779**	**0.6489**	**0.6395**	**0.6411**	**0.6310**	**0.6593**	**0.6494**	**0.6496**
Visual Question Answering								
Pascal-EA	<u>0.5081</u>	<u>0.4669</u>	0.4322	0.4306	<u>0.4943</u>	<u>0.5155</u>	<u>0.5365</u>	<u>0.4834</u>
GeNIe	0.4776	0.4431	<u>0.4411</u>	<u>0.4495</u>	0.4299	0.4731	0.4657	0.4543
VC	0.3629	0.3268	0.3209	0.3265	0.3426	0.3744	0.3762	0.3472
EvoNAT-Edit (Ours)	**0.5881**	**0.5965**	**0.6093**	**0.6098**	**0.5412**	**0.6261**	**0.5803**	**0.5930**

(b) LLM Judge Score

Method	Qwen3-VL 8B	InternVL3.5 8B	LLaVA-1.5 7B	LLaVA-1.5 13B	Llama-3.2 11B-Vision	GPT-5.4	Claude Sonnet 5	Average
Image Captioning								
Pascal-EA	0.4507	0.4493	0.4497	0.4717	0.5050	0.5130	0.4963	0.4765
GeNIe	<u>0.5773</u>	<u>0.5740</u>	<u>0.6047</u>	<u>0.5970</u>	<u>0.6037</u>	<u>0.5770</u>	<u>0.5727</u>	<u>0.5866</u>
VC	0.5063	0.5293	0.5470	0.5470	0.5860	0.5327	0.5177	0.5380
EvoNAT-Edit (Ours)	**0.7653**	**0.7507**	**0.7723**	**0.7540**	**0.7650**	**0.7643**	**0.7510**	**0.7604**
Visual Question Answering								
Pascal-EA	0.4687	0.4487	0.5077	0.4703	0.4627	0.4543	0.5140	0.4752
GeNIe	<u>0.5897</u>	<u>0.5753</u>	<u>0.6070</u>	<u>0.6237</u>	<u>0.5803</u>	<u>0.5673</u>	<u>0.5807</u>	<u>0.5891</u>
VC	0.4900	0.5037	0.5567	0.5313	0.5373	0.4777	0.5123	0.5156
EvoNAT-Edit (Ours)	**0.7710**	**0.7277**	**0.7530**	**0.7533**	**0.7337**	**0.7833**	**0.7573**	**0.7542**

Because VQA responses do not guide evolution, these results show that the deviations found through IC feedback carry over to another task. The agreement between cosine-based and text-only scores further supports this finding across two ways of assessing the responses.

4.3 Source-Conditioned Adversarial Editing

We evaluate EvoNAT-Edit on the same fixed set of 750 source images used by the editing baselines. Each candidate specifies source-grounded semantic changes to the original image. We evaluate the edited images on image captioning and held-out visual question answering.

Table 2 compares source-conditioned attack performance across the seven target LVLMs. The cosine-based panel measures semantic deviation, while the LLM Judge panel provides a text-only assessment of the responses. Both panels report image captioning and held-out VQA results.

Table 2 shows that EvoNAT-Edit achieves the highest cosine-based attack score on every target LVLM for both tasks. Its seven-model averages are 0.6496 on image captioning and 0.5930 on held-out VQA, compared with 0.4768 and 0.4834 for the strongest corresponding baselines. The LLM Judge scores show the same pattern, with EvoNAT-Edit reaching 0.7604 and 0.7542 on the two tasks.

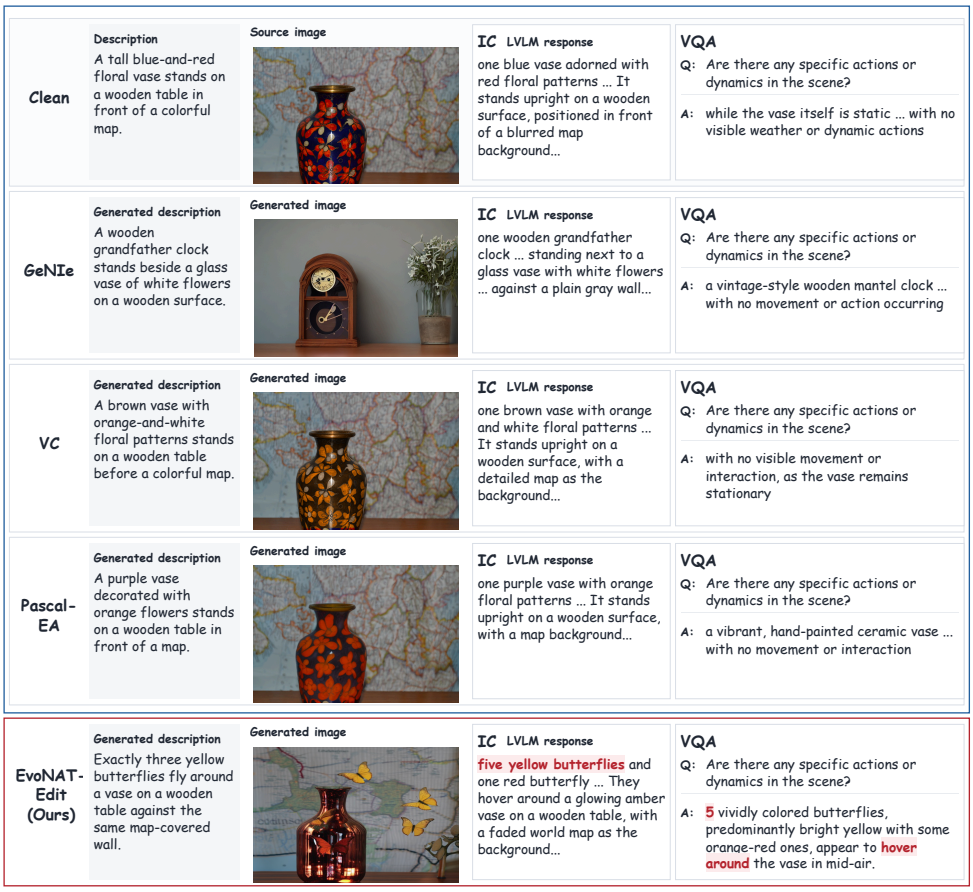


Figure 3: **Qualitative comparison of source-conditioned editing.** Rows show the clean source and each method's edited image, description, and IC and VQA responses. Red marks discrepancies from the EvoNAT-Edit description.

Figure 3 compares edits of the same vase image. Baseline responses largely match their descriptions. EvoNAT-Edit specifies exactly three yellow butterflies, but IC reports five yellow and one red butterfly; VQA reports five butterflies hovering around the vase. The example shows discrepancies in count and color, with an additional motion detail in VQA.

4.4 Naturalness and Visual Quality

We assess visual quality with MUSIQ (Ke et al., 2021), TOPIQ-NR (Chen et al., 2024), HyperIQA (Su et al., 2020), and blind human naturalness ratings. Human raters judge whether an image appears natural and coherent on a 1–5 scale. Higher is better for all four measures.

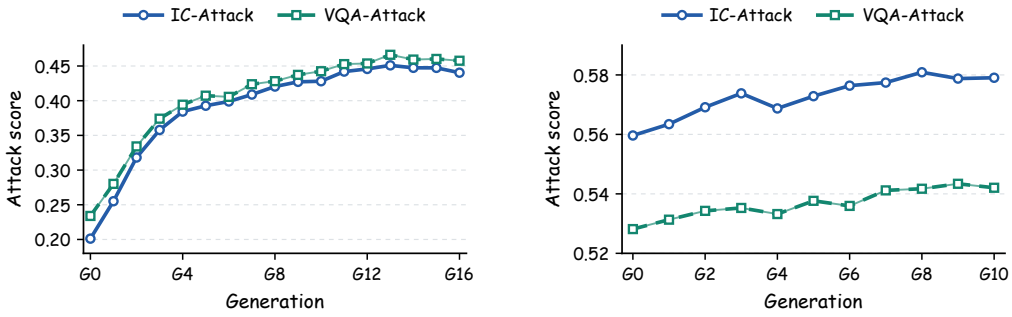
Table 3 shows that both branches lead their respective baselines on all four measures. Human Naturalness reaches 3.964 for EvoNAT versus 3.468 for the best generation baseline, and 4.040 for EvoNAT-Edit versus 3.880 for the best editing baseline. Both branches also lead on the three automatic metrics. Together with Tables 1 and 2, these results show that their stronger attacks retain high visual quality.

4.5 Evolution Dynamics and Cross-Task Transfer

Figure 4 plots current-generation mean IC-Attack and VQA-Attack scores. Only IC guides selection; VQA measures held-out transfer. These means differ from the final archive-selected scores in Tables 1 and 2.

Table 3: Naturalness and visual quality comparison using three no-reference metrics and blind human naturalness ratings (1–5). Higher values are better.

Method	MUSIQ ↑	TOPIQ-NR ↑	HyperIQA ↑	Human Naturalness ↑
Source-Free Adversarial Generation				
Random	70.8714	0.6056	0.6249	3.412
NL-adv	70.3977	0.6059	0.6215	3.468
LANCE	70.8564	0.6067	0.5917	3.185
EvoNAT (Ours)	**71.7381**	**0.6227**	**0.6572**	**3.964**
Source-Conditioned Adversarial Editing				
Pascal-EA	67.4131	0.5510	0.5707	3.880
GeNIe	68.9043	0.5616	0.5767	3.620
VC	69.5060	0.6418	0.6579	3.760
EvoNAT-Edit (Ours)	**71.1542**	**0.6793**	**0.6962**	**4.040**



(a) **EvoNAT.** Evolution through generation 16.

(b) **EvoNAT-Edit.** Evolution through generation 10.

Figure 4: **Evolution dynamics of EvoNAT and EvoNAT-Edit.** The curves show current-generation mean IC-Attack and VQA-Attack scores. IC-Attack guides evolution; VQA-Attack measures transfer.

For EvoNAT, both scores rise rapidly through generation 4 and then level off. At generation 13, IC-Attack reaches 0.4509 and VQA-Attack reaches 0.4662; both change little through generation 16. VQA improves despite being held out from selection, indicating cross-task transfer.

For EvoNAT-Edit, IC-Attack rises from 0.5596 to 0.5809 by generation 8, while VQA-Attack rises from 0.5282 to 0.5434 by generation 9. Both remain above their initial levels and change little near the end, showing similar transfer for source-grounded edits.

5 Conclusion

We introduced EvoNAT, an output-only framework that evolves structured descriptions using captioning feedback to generate adversarial images without a source image. By searching over objects, attributes, actions, and scene context, it can explore changes beyond local perturbations of a fixed image. EvoNAT-Edit applies the same search principle to edits grounded in real images.

Across seven LVLMs, both branches lead their respective baselines in image-captioning and held-out VQA attacks, as well as in three automatic quality metrics and blind Human Naturalness ratings. The VQA gains arise without VQA feedback during evolution, showing that the discovered semantic deviations transfer beyond the optimization task while retaining visual quality.

AI Use Statement

We used generative AI tools to aid and polish the writing of this manuscript. The assistance concerned how the authors' material was expressed: improving sentence clarity and concision, refining

transitions between paragraphs, and reducing repetition. The authors reviewed the suggested wording and revised it where necessary to preserve the intended meaning.

The authors determined the scientific content and final presentation of the paper. We checked the AI-assisted prose against the methods and results reported in the manuscript, reviewed the final wording and citations, and made the final decisions about what to include. We take responsibility for the final submission, including its claims, results, references, and AI-assisted text.

Ethics Statement

Five volunteers participated in a blinded image-rating study of image naturalness. Before the evaluation, participants were informed that they would view generated and edited images and rate their naturalness and visual coherence on a 1–5 scale. All participants provided informed consent before taking part. Participation was voluntary and unpaid.

We collected no names, contact details, or other directly identifying information. The ratings were stored and analyzed without direct identifiers. The study involved viewing non-distressing images and submitting ratings, with no physical intervention and no anticipated risks beyond the time required to complete the task.

Under our institution's policy, this minimal-risk, non-interventional image rating study was not required to undergo formal IRB review.

References

Soroush Abbasi Koohpayegani, Anuj Singh, K. L. Navaneet, Hamed Pirsiavash, and Hadi Jamali-Rad. GeNIe: Generative hard negative images through diffusion. *arXiv preprint arXiv:2312.02548*, 2023.

Anthropic. Introducing Claude Sonnet 5. `https://www.anthropic.com/news/claude-sonnet-5`, 2026. Accessed: 2026-08-29.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Wenbin Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2443–2455. PMLR, 2024.

Tom B. Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. doi: 10.1109/TIP.2024.3378466.

Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. In *Advances in Neural Information Processing Systems*, 2023.

Jacob Cohen. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: 10.1037/h0026256.

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.

Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. PromptBreeder: Self-referential self-improvement via prompt evolution. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13481–13544. PMLR, 2024.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852. Association for Computational Linguistics, 2025.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations*, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Klaus Krippendorff. Computing krippendorff's Alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania, 2011. URL `https://repository.upenn.edu/asc_papers/43`.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Meta AI. Llama 3.2 vision model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md`, 2024.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026. Accessed: 2026-08-29.

Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. LANCE: Stress-testing visual models by generating language-guided counterfactual images. In *Advances in Neural Information Processing Systems*, 2023.

Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. SemanticAdv: Generating adversarial examples via attribute-conditional image editing. In *European Conference on Computer Vision*, 2020.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, 2018.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, 2020. doi: 10.1109/CVPR42600.2020.00372.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024.

Zijin Yin, Kongming Liang, Bing Li, Zhanyu Ma, and Jun Guo. Benchmarking segmentation models with mask-preserved attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22509–22519, 2024.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. AnyAttack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19900–19909, 2025.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54111–54138, 2023.

Xiaopei Zhu, Peiyang Xu, Guanning Zeng, Yingpeng Dong, and Xiaolin Hu. Natural language induced adversarial images. In *ACM International Conference on Multimedia*, 2024.

A Additional Experimental Details

This appendix details the experimental protocol, implementation, evaluation prompts, and human-evaluation analyses for EvoNAT. All experiments follow the output-only black-box setting described in the main paper.

A.1 Evaluation Protocol

We evaluate EvoNAT as an output-only black-box framework for natural-language-guided semantic evolution. The framework searches over structured descriptions of visual content, renders each candidate with a setting-specific image generator or editor, and uses textual feedback from a proxy LVLM to guide the search. The attacker has no access to model parameters, gradients, hidden states, or internal visual representations.

We evaluate two settings. In source-free generation, each candidate describes a complete visual scene and is rendered by a text-to-image model. In source-conditioned editing, EvoNAT-Edit specifies semantic changes to a target object in a source image and renders the candidate with an image-conditioned editor. The settings use different candidate spaces and renderers, but share the same semantic-space evolution principle and caption-based feedback objective.

The evaluation has an optimization stage and a transfer stage. During optimization, image captioning (IC) is the only feedback signal used to compute fitness, select parents, and generate new candidates. After search, we freeze the selected candidate set and evaluate it using both IC and visual question answering (VQA). VQA is held out during optimization and measures cross-task transfer. Image-quality metrics, human naturalness ratings, and the text-only Granite Judge are also computed after final candidate selection and do not guide the search.

Each method is evaluated on 750 instances in each setting. Thus, across the four methods compared in each setting, the evaluation covers 3,000 method-image instances. For source-free generation, each method contributes 750 generated images. For source-conditioned editing, all methods use the same fixed pool of 750 COCO 2017 validation images, preserving source identifiers and official instance annotations when required. Internal evolution may evaluate additional candidates before the final set is selected. All methods use the same IC and VQA instructions and response-scoring protocol within each setting.

We evaluate transfer across seven LVLMs: Qwen3-VL-8B-Instruct, InternVL3.5-8B, LLaVA-1.5-7B, LLaVA-1.5-13B, Llama-3.2-11B-Vision-Instruct, GPT-5.4, and Claude Sonnet 5. Qwen3-VL-8B-Instruct serves as the proxy LVLM during evolution. After the candidate set is frozen, all seven models, including the proxy, are evaluated with the same IC and VQA protocols.

We log all evaluation attempts, including rendering status, response records, and evaluation coverage. For source-conditioned editing, the methods share the same 750 source-image identifiers. Whenever a reported statistic uses fewer than the full set of instances, we state its denominator and evaluation coverage.

A.2 Complete Evaluation Prompts

We use fixed prompts for image captioning and visual question answering. Within each experimental setting, the prompts are identical across methods and target LVLMs.

Image-captioning prompt (IC)

```
Please describe the image in the following format: a
<picture/photo/watercolor/sketch> of <number> <color> <object>
<appearance> in the style of <style>. <They/It/He/She> <gesture>
on the <background describe> in the <location> on a <weather>
day, <action description>, <environment description>. Do not add
extraneous statements and strictly follow the instructions.
```

Visual-question-answering prompt (VQA)

```
Please answer my following questions in detail based on the
content of the picture. You need to write your answer as a
continuous descriptive sentence. What type of picture is this?
For example: a photo, sketch, watercolor painting, or another
type? What color should the main object(s) be? What is the main
object in this image? Are there any specific appearance features?
What style would you like? For example: vintage, modern
illustration, or a specific artist's style? What gesture or
action is the object doing in the image? What type of background
setting is this? Are there specific scene elements or details?
Where is this scene set? What is the weather like in the scene?
Are there any specific actions or dynamics in the scene? For
instance: swaying in the wind, interacting with another object,
etc. How would you describe the overall atmosphere of the
surroundings?
```

B Implementation Details

For source-free generation, EvoNAT uses Stable Diffusion 3.5 Medium as its image renderer. EvoNAT-Edit uses an image-to-image rendering pipeline for source-conditioned editing. Both settings follow the configurations used in our experiments.

Qwen3-VL-8B-Instruct provides proxy feedback during evolution. We use all-mpnet-base-v2 to compute semantic deviation between the intended description and the model response. Transfer is evaluated across Qwen3-VL-8B, InternVL3.5-8B, LLaVA-1.5-7B, LLaVA-1.5-13B, Llama-3.2-11B-Vision-Instruct, GPT-5.4, and Claude Sonnet 5. The editing baselines follow their method-specific procedures: Pascal-EA uses mask-guided diffusion editing, GeNIe uses source-conditioned diffusion editing, and VC performs localized color editing with an instance mask.

Algorithm 1 summarizes the search and final selection procedure. The two settings use different candidate representations, renderers, and selection rules.

Algorithm 1 Search and final selection for EvoNAT and EvoNAT-Edit

Require: Setting m, initial candidates $\mathcal{C}_0$, proxy LVLM M_p, IC instruction q_{IC}, text encoder ϕ, number of generations T_m

for $t = 0, \ldots, T_m$ **do**
  for all $c \in \mathcal{C}_t$ **do**
    Render c with the renderer for setting m, obtaining I_c
    Obtain the proxy response $r_c \leftarrow M_p(I_c, q_{\text{IC}})$
    Compute proxy fitness $f(c) \leftarrow 1 - \cos(\phi(p_c), \phi(r_c))$
  end for
  Update the search archive with $\mathcal{C}_t$ and its fitness values
  if $t < T_m$ **then**
    Form $\mathcal{C}_{t+1}$ using the search operators for setting m
  end if
end for
if $m =$ EvoNAT **then**
  Select the final image set from the archive using the criterion in Sec. 3.4
else
  For each source image, select its highest-fitness candidate across G0–G10
end if

For EvoNAT, the final image set is selected from the search archive using the criterion described in Section 3.4. For EvoNAT-Edit, candidates are rendered from their original source images, and selection is performed independently for each source image using proxy IC-Attack across G0–G10. VQA responses and image-quality measurements are not used for search or final selection.

Table 4: Inter-rater agreement for human naturalness ratings. Each setting contains 3,000 images in total, with 750 images per method; each image was rated independently by three volunteers.

Metric	Source-free generation	Source-conditioned editing
Images rated (total; per method)	3,000 (750)	3,000 (750)
Ordinal Krippendorff's α	0.738	0.677
Exact agreement	64.4%	62.1%
Agreement within one point	99.3%	99.3%

Table 5: Agreement between the Granite Judge and the human audit.

Measure	Result
Audited records	2,800
Quadratic weighted Cohen's κ	0.888
Exact agreement	75.1%
Agreement within one score point	96.4%
Mean absolute error	0.285

C Human Evaluation

C.1 Inter-Rater Agreement

Three volunteers independently rated each image for naturalness on a five-point scale, with higher scores indicating greater perceived naturalness. Method identities were hidden from the raters. We measure inter-rater agreement using ordinal Krippendorff's α (Krippendorff, 2011), separately for source-free generation and source-conditioned editing.

For source-free generation, ordinal Krippendorff's α is 0.738. All three raters gave the same score for 64.4% of images, and their ratings were within one point for 99.3% of images. For source-conditioned editing, ordinal Krippendorff's α is 0.677. Exact three-rater agreement is 62.1%, and agreement within one point is 99.3%.

For each image, we average the three ratings to obtain an image-level naturalness score. Each method's reported score is the mean over its 750 images. Human ratings are used only for evaluation; they do not guide candidate evolution or final selection.

C.2 Human Audit of the Granite Judge

We audit the Granite Judge on a stratified sample of 2,800 model-response records. The sample covers seven LVLMs, two experimental settings, and two evaluation tasks, with 100 records drawn from each model-by-setting-by-task stratum. A human annotator independently scores each record from 0 to 4 using the same rubric as the Granite Judge. The annotator sees the target prompt and model response, but not the Granite Judge score, model identity, or method identity.

We measure agreement between human ratings and the Granite Judge using quadratic weighted Cohen's κ (Cohen, 1968). We also report exact agreement, agreement within one score point, and mean absolute error (MAE). On the audited records, quadratic weighted Cohen's κ is 0.888. Exact agreement is 75.1%, 96.4% of ratings differ by at most one point, and the MAE is 0.285.

D Additional Source-free Qualitative Examples

Figure 5 presents additional qualitative examples of EvoNAT in the source-free setting across all seven target LVLMs. Each row shows the generated description and corresponding image, followed by the image-captioning (IC) response and the held-out visual-question-answering (VQA) response from each model. Red text highlights response content that is semantically inconsistent with the generated description.



Figure 5: Additional qualitative examples of EvoNAT in the source-free setting. Each row shows the generated description, generated image, and the IC and held-out VQA responses from seven target LVLMs. Red text highlights response content that is semantically inconsistent with the generated description.

The examples illustrate several forms of semantic deviation found by EvoNAT. In the first example, the folded-paper carousels are frequently described as umbrellas, tents, or pavilion-like structures. In the second example, the emerald jellyfish are consistently described as mushrooms, while the VQA responses also produce incorrect object counts. In the third example, the emerald floral fountains are interpreted as crystal towers, castles, monuments, or related structures, while the snow-covered mountain valley is replaced by substantially different scene contexts such as city squares, rainforests, deserts, and coastal environments.

These examples complement the aggregate results by showing that the semantic deviations are not limited to a single visual attribute. They occur across object identity, quantity, and scene context, and are observed across multiple target LVLMs. Similar deviations also appear in the held-out VQA responses, providing qualitative evidence that the adversarial semantics discovered through IC-guided evolution transfer to a different task.