

Advdreamer Unveils: Are Vision-Language Models Truly Ready for Real-World 3D Variations?

Supplementary Material

Overview

This supplementary material provides some essential details that complement our main paper. Sec. A presents the mathematical derivation of the AdvDREAMER distribution update formulation, which builds upon the canonical form of CMA-ES [7, 8]. Sec. C provides detailed experimental settings and results. Specifically, Sec. C.1 details the implementation of AdvDREAMER and the configuration of critical hyperparameters. Sec. C.2 presents a comprehensive ablation study quantifying the impact of NRM feedback on Adv-3DT’s naturalness. Sec. C.3 offers additional results and analysis of physical experiments. Sec. C.5 provides the computational cost of AdvDREAMER. Sec. C.6 provides a comparison of Adv-3DT samples from AdvDREAMER with those from previous studies. Sec. D provides all prompt templates provided to GPT-4 for automatic annotation of naturalness datasets and GPT-Score / GPT-Acc metrics. Sec. E showcases additional visualization of Adv-3DT examples and its performance across various tasks. Finally, Sec. F provides an in-depth description of MM3DTBench and presents detailed quantitative benchmarking results. Sec. B reports the ImageNet categories utilized in experiments and examples of clean images.

Open Source Declaration

We will release the source code of AdvDREAMER, along with the MM3DTBench dataset and the evaluation scripts. We will also release our naturalness reward model’s weights and training dataset. We believe these open-source contributions will facilitate future research in VLMs’ 3D variation robustness evaluation.

A. Details of Optimization Algorithm

A.1. Derivation of Eq.(10) & (11)

AdvDREAMER optimizes adversarial 3D transformation distributions through Covariance Matrix Adaptation Evolution Strategy (CMA-ES). Among various evolutionary optimization algorithms, CMA-ES stands out as one of the most prominent and effective approaches, demonstrating superior performance particularly on medium-scale optimization problems (typically involving 3-300 variables) [7]. Its gradient-free nature eliminates the dependency on gradient information, making it an ideal choice for optimizing the adversarial 3D transformation distributions in AdvDREAMER.

Following the canonical formulation of CMA-ES, Adv-

DREAMER generates a population of 3D transformations in each optimization iteration t by sampling from the current adversarial distribution with a specified step size σ^t , formally expressed as:

$$\mathbf{z}_i = \boldsymbol{\mu}^t + \sigma^t \cdot \boldsymbol{\gamma}_i \text{ where } \boldsymbol{\gamma}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}^t), \quad (\text{A.1})$$

where each $\boldsymbol{\gamma}_i$ represents a search direction in the transformation space. The sampling process for $\boldsymbol{\gamma}_i$ can be performed through either Eigendecomposition of the current adversarial distribution’s covariance matrix $\boldsymbol{\Sigma}^t = B D^2 B^T$, or Cholesky decomposition, followed by transformation of samples drawn from a standard normal distribution:

$$\boldsymbol{\gamma}_i = B D \boldsymbol{\delta}_i, \text{ where } \boldsymbol{\delta}_i \sim \mathcal{N}(0, \mathbf{I}). \quad (\text{A.2})$$

The sampled population $\{\mathbf{z}_i\}_{i=1}^K$ undergoes reparameterization to obtain a constrained set of 3D transformation parameters $\{\boldsymbol{\Theta}_i\}_{i=1}^K$, as formulated in Eq.(8). By applying $\{\boldsymbol{\Theta}_i\}_{i=1}^K$ to the original images under generative process, we generate a batch of Adv-3DT samples $\{X'_i\}_{i=1}^K$ and compute their \mathcal{L}_{Nat} and \mathcal{L}_{IPS} losses. The samples are then ranked according to their \mathcal{L}_{IPS} values, which quantify the adversarial effectiveness against the target model:

$$\mathcal{L}_{\text{IPS}}(X'_{1:K}) \leq \mathcal{L}_{\text{IPS}}(X'_{2:K}) \leq \dots \leq \mathcal{L}_{\text{IPS}}(X'_{K:K}). \quad (\text{A.3})$$

We employ truncation selection to identify the samples contributing to distribution parameter updates. Specifically, we select the k most adversarial samples based on \mathcal{L}_{IPS} scores, then retain the top $k/2$ samples with the highest \mathcal{L}_{Nat} values. This two-stage selection mechanism effectively serves as a naturalness regularization for the optimization process. For the sake of simplicity, we denote this selected subset of population as $\{\mathbf{z}_{(i:k)}\}_{i=1}^k$ and utilize it to update both the distribution parameters $\boldsymbol{\mu}^{t+1}$, $\boldsymbol{\Sigma}^{t+1}$ and step size σ^{t+1} . The updated mean of the distribution is computed as the weighted maximum likelihood estimate of the selected population, which can be expressed as:

$$\boldsymbol{\mu}^{t+1} = \sum_{i=1}^k w_i \cdot \mathbf{z}_{(i:k)}^{t+1}, \text{ where } \sum_{i=1}^k w_i = 1. \quad (\text{A.4})$$

This formulation indicates that the updated mean shifts along the average search direction. In our implementation, we assign uniform weights to all selected population, where $w_i = 1/k$. The covariance matrix $\boldsymbol{\Sigma}$ update comprises both rank-1 and rank- μ terms. The rank-1 term incorporates historical search information through an evolution path, which

Table S.1. Category-wise accuracy (%) of physical Adv-3DT samples under different tasks.

Task	airliner	ambulance	fire truck	forklift	water jug	laptop	police van	shoe	street sign	tank	traffic light	trash bin
Zero-shot Cls.	49.4	68.4	1.0	85.0	0.0	99.5	25.3	100.0	48.2	90.9	47.2	0.4
VQA-Choice	3.1	35.9	67.2	9.4	0.0	99.5	34.7	100.0	57.2	5.3	0.6	26.3
VQA-Ture or False	0.3	14.1	0.5	37.2	99.5	99.5	2.6	100.0	96.4	85.2	0.3	58.8

is constructed as:

$$p_{t+1} = (1 - c)p_t + \sqrt{c(2 - c)}\sqrt{\mu_w} \frac{\mu^{t+1} - \mu^t}{\sigma^t}, \quad (\text{A.5})$$

where the factor $\mu_w = \frac{1}{\sum_{i=1}^k w_i^2}$, c represents the adaptation rate/learning rate, is designed based on $c^{-1} \propto n$. Notably, c is inversely proportional to the degrees of freedom (number of parameters) of the adjustable variables. The evolution path characterizes the movement of the distribution mean by maintaining an exponentially weighted average of the update directions $\frac{\mu^{t+1} - \mu^t}{\sigma^t}$ across iterations. This path effectively encodes one of the most promising search directions in the current optimization landscape. The covariance matrix update is based on:

$$\text{argmax}_p(p_{t+1} \mid \mu, \Sigma), \text{argmax}_{\Sigma} \prod_{i=1}^k p\left(\frac{\mathbf{z}_{(i:k)} - \mu^t}{\sigma^t} \mid \mu, \Sigma\right). \quad (\text{A.6})$$

Building upon this principle, the update rule can be formulated as:

$$\begin{aligned} \Sigma^{t+1} &= (1 - \eta_1 - \eta_\mu) \cdot \Sigma^t + \eta_1 \cdot p_{\Sigma}^{t+1} (p_{\Sigma}^{t+1})^T + \\ &\eta_\mu \cdot \sum_{i=1}^h w_i \cdot \left(\frac{\mathbf{z}_{(i:k)}^{t+1} - \mu^t}{\sigma^t}\right) \left(\frac{\mathbf{z}_{(i:k)}^{t+1} - \mu^t}{\sigma^t}\right)^T. \end{aligned} \quad (\text{A.7})$$

This update mechanism effectively increases the sampling probability along successful search directions by expanding the variance in these directions. The rank-1 term in Eq. (A.7), as previously mentioned, directly leverages the evolution path as a successful search direction. The second term (rank- μ), represents the weighted maximum likelihood estimate of the selected population, which can be interpreted as natural gradient optimization of Σ in the information geometry framework. The learning rates η_1, η_μ for both terms follow the same design principle as $c_1 \approx \frac{2}{n^2}$, $c_\mu \approx \frac{\mu_w}{n^2}$, where they are inversely proportional to the degrees of freedom (number of parameters) of the adjustable variables.

For step size σ adaptation, we adopt the Cumulative Step-size Adaptation (CSA) mechanism, which is widely recognized as the most successful and prevalent approach in the CMA-ES framework [5, 6]. The above derivations provide a comprehensive elaboration of the algorithmic details presented in Sec. 3.5 of the main paper.

A.2. Pseudocode of AdvDreamer Framework

We provide the pseudocode for AdvDreamer in Algorithm 1

Algorithm 1: Optimization Algorithm of AdvDreamer

```

Data: Natural images or text descriptions  $I \in \{X, T\}$ .
Result: Optimal distribution parameters  $\mu^*, \Sigma^*$ .
/* Foreground-Background Pairs Preparation */
1  $\{X_f, X_b\} \leftarrow \mathcal{F}(I)$ ;
/* Initialization of distribution parameters */
2  $\mu^0 \leftarrow \mathbf{0}, \Sigma^0 \leftarrow \mathbf{I}$ ;
3 while  $t < t_{\max}$  do
  /* Generate a batch of adversarial samples */
  4 for  $i = 1 \rightarrow K$  do
    5  $\mathbf{z}_i^{t+1} \sim \mathcal{N}(\mu^t, \Sigma^t)$ ;
    6  $\Theta_i^{t+1} \leftarrow \mathbf{A} \cdot \tanh(\mathbf{z}_i^{t+1}) + \mathbf{B}$ ;
    7  $(X')_i^{t+1} \leftarrow C_{w_1}(\mathcal{R}_{w_0}(X_f, \Theta_i^{t+1}), X_b)$ ;
    /* Calculate loss/fitness value */
    8  $\mathcal{L}_i^{t+1} = \mathcal{L}_{\text{IPS}}((X')_i^{t+1}, \mathcal{Y}) + \lambda \cdot \mathcal{L}_{\text{Nat}}((X')_i^{t+1})$ ;
    9 end
    10 Sorting  $\mathcal{L}_i^{t+1}$  in ascending order;
    11 Obtain  $\mu^{t+1}, \Sigma^{t+1}$  by Eq.(A.4) and (A.7);
  12 end
  13  $\mu^* \leftarrow \mu^{t_{\max}}, \Sigma^* \leftarrow \Sigma^{t_{\max}}$ .

```

Table S.2. The computational cost of AdvDreamer across different stages and total optimization process.

	Step.0	Step.1@iter	Step.2@iter	Step.3@iter	Toatal
Time Cost	$\leq 1s$	$\sim 0.6s \cdot k$	$\sim 6s \cdot k$	$\leq 1s$	$\sim 16.5min$

Table S.3. Computational Cost and Accuracy of Naturalness Assessment Across Human, GPT, and NRM Evaluations.

	Human	GPT-4o	NRM
Time cost	$\sim 43min$	$\sim 4min$	$\sim 5s$
Acc.	-	64.94%	67.82%

B. Selected ImageNet Categories

We conduct experiments of zero-shot classification task (Tab. 1) on 30 ImageNet Categories, generating Adv-3DT samples from both ImageNet test set images and synthetic images created using Stable-Diffusionv2. The selected categories are enumerated in Tab. S.4.

C. Additional Experimental Results

C.1. Implementation Details

Optimization Algorithm. we employ CMA-ES through the cmaes [12] library. The distribution mean μ is initialized as $[90, 180, 90, 0, 0, 1.0]$, with optimization bounds constrained to $\Theta_{\min} = [0, 160, 80, -100, -100, 0.5]$, $\Theta_{\max} = [360, 200, 100, 100, 100, 1.5]$. We initialize covariance matrix μ as a identity matrix, with a step size σ of

Table S.4. The selected 30 categories in ImageNet dataset.

0	backpack	5	carton	10	rocking chair	15	beaker	20	coffee mug	25	monitor
1	barber chair	6	chest	11	lamp	16	convertible	21	ballpoint	26	dustcart
2	minivan	7	moped	12	cocktail shaker	17	park bench	22	waggon	27	cleaver
3	folding chair	8	microwave	13	ambulance	18	jug	23	moving van	28	mailbox
4	hatchet	9	coffeepot	14	fire truck	19	crate	24	chiffonier	29	doormat

Table S.5. VQA accuracy(%) of representative VLMs under MM3DTBench.

VLMs	Choice	Free Answer	Avg.	Rank
GPT-4o [13]	69.30	47.44	58.37	1
GPT-4o-mini [13]	58.60	46.51	52.56	2
CogVLM [16]	60.00	40.00	50.00	3
InternVL [4]	46.05	49.77	47.91	4
Qwen-VL [2]	58.60	36.74	47.67	5
LLaVa-1.6 [10]	49.77	34.42	42.10	6
ShareGPT-4V [3]	50.23	33.02	41.63	7
MiniGPT-4 [18]	48.84	30.70	39.77	8
Mplug-Owl2 [17]	40.47	32.56	36.52	9
InternLM-2 [15]	38.60	27.90	33.25	10
LLaVa-1.5 [11]	41.40	24.60	33.00	11
Claude-3 [1]	35.81	24.19	30.00	12
LRV-Instruction [9]	23.70	13.02	18.36	13

0.5, which decays exponentially during optimization with a decay rate of 0.9. Other hyperparameters follow the default configuration in official implementation.

Training Details of NRM. To train the Natural Reward Model (NRM), we first establish a large-scale image naturalness assessment dataset, which comprises 100k images obtained by applying random 3D transformations to samples from ImageNet and generated images. Following the methodology detailed in Sec. 3.4 of main paper, we leverage GPT-4o for automated naturalness scoring, complemented by human verification. Representative examples of the annotation process are illustrated in Fig. S.1. The dataset is split into training and validation sets with a 9:1 ratio.

We formulate NRM training as a multi-classification task, where the model predicts specific scores for realism and physical plausibility. The backbone of NRM adopts DI-NOv2 (ViT-L/14 distilled), where we apply average pooling over patch tokens to obtain 1024-dimensional feature representations. These features are then fed into two separate prediction heads for parallel classification. Each head consists of two fully connected layers with non-shared weights, configured as (1024, 512) and (512, 5), respectively. The network is optimized using cross-entropy loss as the training objective. The NRM is trained for 100 epochs, achieving a prediction accuracy of 75.03% on the test set.

C.2. Naturalness Evaluation Effect

Alignment between NRM and Human Evaluation: A critical aspect of NRM is its ability to approximate human judgment for replacing manual Adv-3DT image fil-

tering during optimization. To validate this alignment, we conducted a human evaluation study on 175 test images with volunteer annotators. Tab. S.3 presents the agreement rates (difference within 1 point) between human scores and predictions from both NRM and GPT-4o. NRM achieves 67.82% accuracy, outperforming GPT-4o’s 64.94%. This superior performance stems from additional manual refinement of NRM’s training data to mitigate the inherent output instability of GPTs. Furthermore, Tab. S.3 demonstrates that NRM’s inference time is negligible compared to both human and GPT-based evaluations, significantly enhancing optimization efficiency.

C.3. Detailed Physical Experiment Results

Tab. S.1 presents a detailed category-wise accuracy of our physical-world experiments (Tab. 4 of main paper). Additionally, we visualize the zero-shot classification results using OpenCLIP ViT-B/16 on sampled frames (every 5 frames) from our physical reproduction video sequences, as shown in Fig. S.2. Green and red boxes indicate successful and failed classifications, respectively.

C.4. Comparison of Samples from AdvDreamer with Those from Previous Studies.

We further compare previous methods in overlapping ImageNet classes. Results in Appendix C.5 demonstrate that AdvDreamer exhibits superior performance in both adversarial effectiveness and sample naturalness. It outperforms GMVFool [14] by 20% in ASR while maintaining greater fidelity to human-captured real Adv-3DT samples in naturalness.

C.5. Computational Cost

The computational overhead for different stages within a single iteration of AdvDreamer is detailed in Tab. S.2. For each clean sample, optimizing the adversarial distribution requires approximately 16.5 GPU minutes, with image reconstruction (step-2) being the primary bottleneck. Our experiments on the ImageNet test set (300 samples) consumed 96 GPU hours on an NVIDIA RTX 3090.

C.6. Comparison with Prior Studies

We evaluate samples generated by different methods in overlapping ImageNet classes. The Tab. ?? shows that AdvDreamer outperforms in both attack success rate (ASR) and naturalness ($Score_R$ and $Score_P$, defined in Eq. (6)).

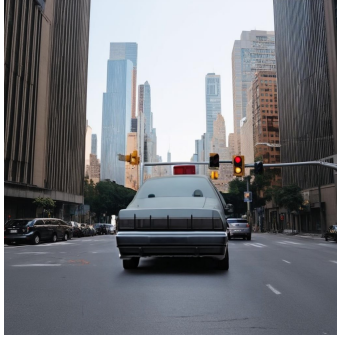


Step 1: Beginning with the overall visual realism of the image, the lighting appears soft and consistent with the presence of the lamp. The colors are natural, and the textures on the bed and walls seem realistic. However, there is an odd object (the pillow propped up) that detracts from the overall coherence of the scene, making it look slightly unnatural. Overall, the image is quite clear with minimal signs of artificial artifacts.

Step 2: Evaluating the physical plausibility of the objects in the scene, the main issue is the pillow appearing to be propped up in a manner that suggests it is floating slightly above the bed. This lack of grounding is not consistent with how one would expect the objects to interact. Additionally, the shadows produced by the lamp are in line with the lighting source, but the floating appearance of the pillow disrupts the physical logic of the scene. Score for Physical **Final Scoring:**

Realistic Score: **3**

Physical Score: **2**



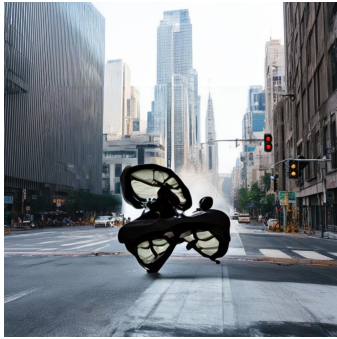
Step 1: Beginning with the overall visual realism of the image, the scene appears relatively clear with some good color consistency. However, there is a noticeable lack of detail in the police car, which makes it look less realistic compared to the background. Additionally, some areas exhibit slight blurriness, particularly around the car, which detracts from the overall clarity.

Step 2: Evaluating the physical plausibility of the objects in the scene, the police car appears grounded, but its details are minimal and lack realism. The shadows and lighting on the car do not perfectly align with the rest of the scene, making it less convincing. The traffic lights are also visible but their integration into the scene feels somewhat disconnected due to the car's lack of detail.

Final Scoring:

Realistic Score: **2**

Physical Score: **2**



Step 1: Beginning with the overall visual realism of the image, there appears to be some blurriness and color inconsistency, particularly with the main object that seems to be poorly integrated into the background. The edges of the object are not well-defined, leading to a somewhat artificial appearance. Overall, the image contains elements that suggest it may have been generated by an image model, particularly in the texture and detail levels.

Step 2: Evaluating the physical plausibility of the objects in the scene, the floating object does not appear grounded or supported in a way that is consistent with gravity. Additionally, there are issues with shadows that do not correspond to the light sources in the scene, contributing to a lack of realism regarding physics in the image.

Final Scoring:

Realistic Score: **1**

Physical Score: **1**

Figure S.1. Examples of image naturalness annotation results using GPT-4o.

Table S.6. comparison of Adv-3DT samples from AdvDreamer with those from previous studies.

Methods		OpenCLIP	BLIP-2	Qwen-VL	4o-mini	GPT-4o	$Score_R$	$Score_P$
Human-guided	[65]	17.3	13.8	4.2	2.1	2.0	3.3	3.3
	[8]	23.8	17.4	10.2	0.0	0.0	3.1	3.0
Simulation-based	[39]	21.6	16.9	6.9	4.6	3.4	2.5	2.9
Multiview-based	[18]	45.4	37.1	10.7	8.1	4.5	2.0	2.1
	[47]	61.9	52.2	16.0	13.2	7.6	2.0	1.0
Generative-based	Ours	82.0	81.3	41.6	39.1	38.4	<u>2.7</u>	<u>2.9</u>

D. Prompt Templates

Fig. S.3 illustrates the prompt template for automatic annotation of NRM training data. The prompt templates for computing GPT-Score and GPT-Acc metrics, which are employed to evaluate image captioning and VQA performance

in our main experiments (Tab. 2), are presented in Fig. S.4.

E. Visualization Examples

We provide additional visualization examples of physically captured Adv-3DT samples in Fig. S.5.

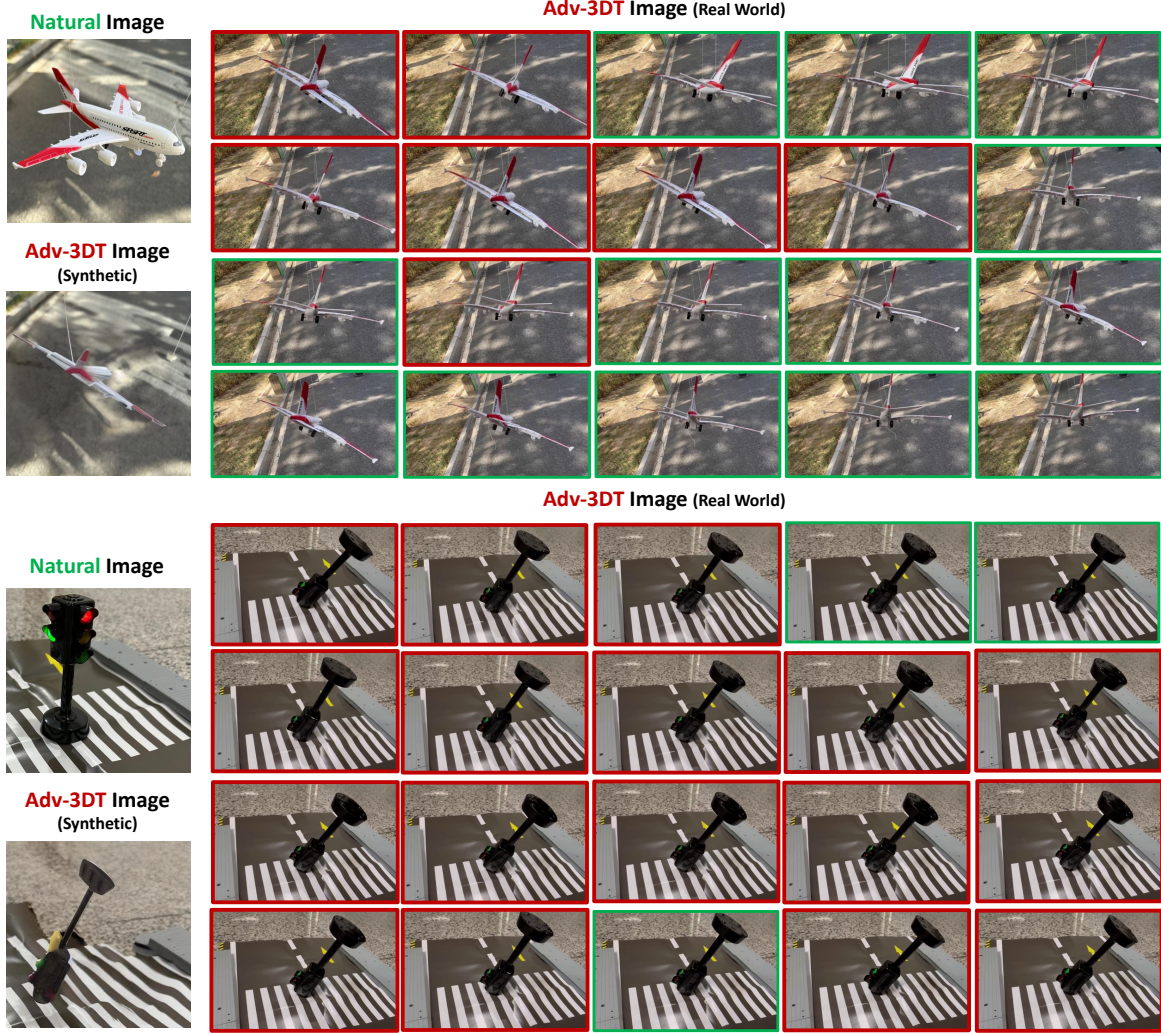


Figure S.2. Physical reproduction results of Adv-3DT Samples: zero-shot classification performance on video frame sequences.

F. MM3DTBench

Tab. S.5 presents comprehensive evaluation results of representative VLMs on our MM3DTBench benchmark. To provide deeper insights into the benchmark’s composition, we visualize exemplar Adv-3DT samples and their corresponding question templates and annotation format in Fig. S.6. The annotation framework consists of four semantic options per sample, including the ground truth label, which is carefully crafted to maintain semantic ambiguity. We also provide the spatial description of the target objects in the annotation file.

"system"

"You are an image analysis expert."

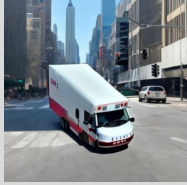
"user"

You are an image analysis expert tasked with evaluating the naturalness of images generated by a model. Your evaluation will be based on two key dimensions:

1. **Visual Realism:** Assess the overall visual realism of the image. Consider factors such as image clarity, color consistency, and whether there are any signs of use Image Generation model like stable diffusion and other artificial artifacts like blurring, ghosting, or meaningless patches. Assign a score based on the level of realism, where 0 is the lowest and 4 is the highest.
2. **Physical Plausibility:** Evaluate the physical plausibility of the objects in the image. Analyze whether all the objects conform to the laws of physics, considering factors like whether objects are appropriately grounded or if they appear to be floating unnaturally, , and whether the light and shadows are natural.. Assign a score based on the physical realism, where 0 is the lowest and 4 is the highest.

Follow the steps below to perform a thorough analysis and conclude with scores for both dimensions.

Example Context:



Step 1: Start by examining the overall visual realism of the image. The image has some blurring and the colors seem slightly off, with some areas showing patches that don't fit well with the rest of the scene.

Step 2: Next, analyze the physical plausibility of the main objects in the scene. The objects in the center of this image appears to be floating without any physical support, and the shadows are inconsistent with the lighting.

Final Scoring: Realistic Score: 1 Physical Score: 2

Now, apply the same chain of thought to evaluate the given image. Please remember to follow the steps carefully and conclude with 'Realistic Score: [score]' and 'Physical Score: [score]'. Please analysis the above image:

Figure S.3. The prompt template for image naturalness evaluation

"system"

"You are an caption quality analysis expert."

"user"

I will provide a predicted caption along with a list of 10 standard captions for an image. Please evaluate the quality of the predicted caption according to the following criteria:

1. **Semantic Accuracy:** Does the caption describe the same object(s) and attributes as in the true captions? The score ranges from 0 to 10, where a higher score means greater semantic similarity and relevance.
2. **Tone Confidence:** Does the caption sound confident? A lower score should be given if the caption contains uncertainty or words like "maybe" or "might". The score ranges from 0 to 10.
3. **Consistency:** Is the predicted caption consistent with the reference captions in an overall sense? The score ranges from 0 to 10.

Please provide the score for each metric, including an explanation for each score. Conclude with the total score in the format "Toatal Score: [score]".

Predicted Caption: "{predicted_caption}"

Reference Captions: "" + "\n".join([f"- {cap}" for cap in reference_captions])

"user"

I will provide a predicted answer to a visual question answering (VQA) task, along with 10 human-provided reference answers with varying confidence levels (e.g., "yes", "maybe", "no"). Please evaluate if the predicted answer is correct based on the reference answers and their confidence levels. If the predicted answer aligns with most high-confidence reference answers ("yes" or multiple "maybe"), mark it as "1" (correct). If it deviates significantly from these high-confidence answers, mark it as "0" (incorrect).

Please provide the score including an explanation for each score. Return a single integer in the format "Score: (1 or 0)" based on your evaluation.

Predicted Answer: "{predicted_answer}"

Reference Answers: {reference_answers_text}

Figure S.4. The prompt template for GPT-Score in image captioning tasks and for GPT-Acc in VQA tasks.



Figure S.5. Additional visualization of Adv-3DT samples.

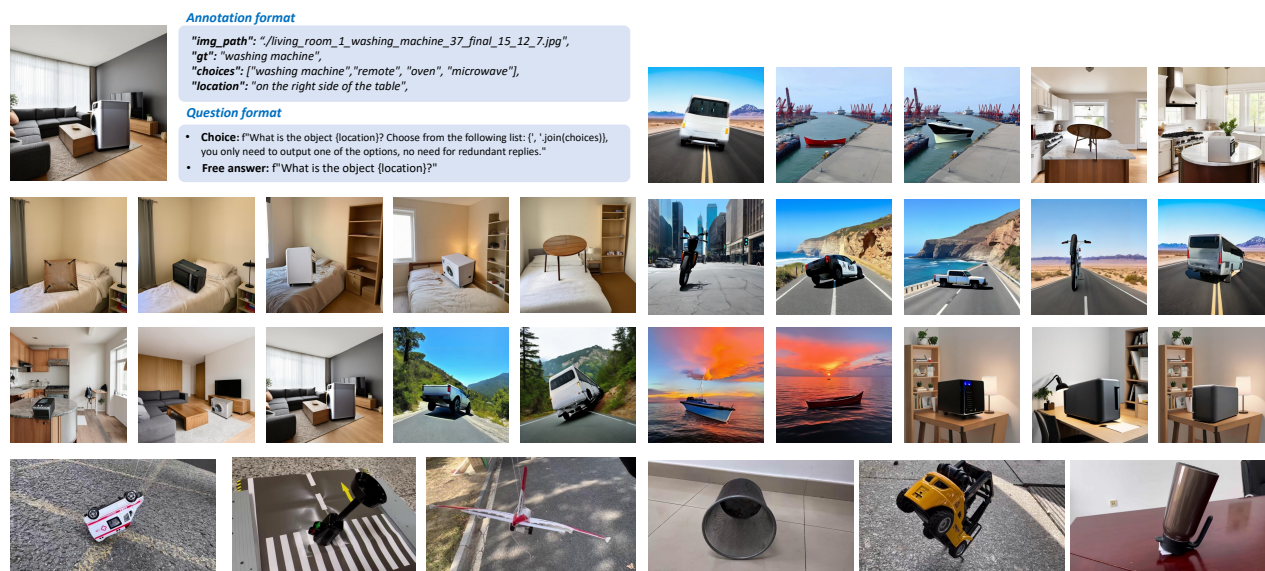


Figure S.6. Anotation format, question format and some examples in MM3DTBench.

References

- [1] Anthropic. claude-3. <https://www.anthropic.com/claude>, 2024. 3
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3
- [3] 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. 3
- [4] 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. 3
- [5] Alexandre Chotard, Anne Auger, and Nikolaus Hansen. Cumulative step-size adaptation on linear functions. In *Parallel Problem Solving from Nature-PPSN XII: 12th International Conference, Taormina, Italy, September 1-5, 2012, Proceedings, Part I 12*, pages 72–81. Springer, 2012. 2
- [6] Alexandre Chotard, Anne Auger, and Nikolaus Hansen. Markov chain analysis of cumulative step-size adaptation on a linear constrained problem. *Evolutionary Computation*, 23(4):611–640, 2015. 2
- [7] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and David Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1487–1495, 2017. 1
- [8] Nikolaus Hansen. The cma evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016. 1
- [9] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 3
- [10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3
- [11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3
- [12] Masahiro Nomura and Masashi Shibata. cmaes : A simple yet practical python library for cma-es. *arXiv preprint arXiv:2402.01373*, 2024. 2
- [13] openai. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. 3
- [14] Shouwei Ruan, Yinpeng Dong, Hang Su, Ning Chen, and Xingxing Wei. Towards viewpoint-invariant visual recognition via adversarial training. In *ICCV*, pages 1–10, 2023. 3
- [15] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023. 3
- [16] Weihai Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3
- [17] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 3
- [18] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3