Introducing GenCeption for Multimodal LLM Benchmarking: You May Bypass Annotations

Lele Cao[1,✉] **Valentin Buchner**[1] **Zineb Senane**[1,2,3,4] **Fangkai Yang**[2]
[1]Motherbrain, EQT Group, Stockholm, Sweden
[2]KTH Royal Institute of Technology, Stockholm, Sweden
[3]Télécom Paris, Palaiseau, France [4]Eurecom, Biot, France
{lele.cao,valentin.buchner,zineb.senane}@eqtpartners.com fangkai@kth.se
https://github.com/EQTPartners/GenCeption

Abstract

Multimodal Large Language Models (MLLMs) are commonly evaluated using costly annotated multimodal benchmarks. However, these benchmarks often struggle to keep pace with the rapidly advancing requirements of MLLM evaluation. We propose GenCeption, a novel and annotation-free MLLM evaluation framework that merely requires unimodal data to assess inter-modality semantic coherence and inversely reflects the models' inclination to hallucinate. Analogous to the popular DrawCeption game, GenCeption initiates with a non-textual sample and undergoes a series of iterative description and generation steps. Semantic drift across iterations is quantified using the GC@T metric. Our empirical findings validate GenCeption's efficacy, showing strong correlations with popular MLLM benchmarking results. GenCeption may be extended to mitigate training data contamination by utilizing ubiquitous, previously unseen unimodal data.

1 Introduction

Large Language Models (LLMs) have shown remarkable capability in natural language understanding, reasoning, and problem solving. Multimodal LLMs (MLLMs) extend these capabilities to multiple modalities, with the visual modality being predominant (Achiam et al., 2023; Liu et al., 2023b; Jiang et al., 2023; Ye et al., 2023). MLLMs harness the power of LLMs as a foundation to incorporate non-textual modality, promising richer interactions and broader applications in real-world scenarios. However, comprehensive evaluation methods that enable comparing different MLLM architectures and training methods are lacking (Fu et al., 2023).

In response, the community has swiftly developed several MLLM benchmarks, such as those detailed by Xu et al. (2022); Dai et al. (2023); Wang et al. (2023); Ye et al. (2023); Li et al. (2023); Zhao et al. (2023). Yet, these benchmarks encounter common challenges: (1) They predominantly rely on multimodal datasets that demand high-quality annotations, which is costly and restrictive in capturing the evolving capabilities of MLLMs (Fu et al., 2023). This has been shown to result in increasing speed in benchmark saturation (Kiela et al., 2021). (2) The evaluation scores may not reflect true performance on real-world tasks due to potential contamination of MLLM training data by benchmark datasets, as reported for LLM pretraining corpora (Dodge et al., 2021; Yang et al., 2023).

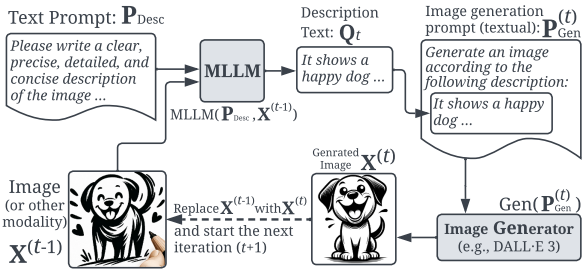

Figure 1: An illustration of the t-th iteration in the GenCeption evaluation procedure for MLLMs. Using the image modality as an example, the process begins with an existing image $\mathbf{X}^{(0)}$ sourced from a unimodal image dataset for the first iteration (t=1). The MLLM provides a detailed description of the image, which is then used by an image generator to produce $\mathbf{X}^{(t)}$.

To address these highlighted challenges, we propose GenCeption, a novel and simple approach for evaluating MLLMs. By iteratively generating and describing non-textual samples, GenCeption gauges MLLMs' ability to consistently maintain semantic coherence across modalities. This approach simultaneously measures the model's tendency to hallucinate, as this inversely correlates with semantic coherence. Further, an MLLM's ability to provide detailed descriptions of non-textual samples measures a diverse range of specialised abilities like object/posture/emotion recognition, numeracy, color perception, OCR, and even the knowledge of artistic styles. Leveraging easily accessible unimodal datasets, GenCeption reduces

Proceedings of the 4th Workshop on Trustworthy NLP (TrustNLP 2024), pages 196–201

Algorithm 1: Calculate GC@T via GenCeption

Input: MLLM to be evaluated, a unimodal dataset $\mathcal{D}$: $\mathbf{X}_1^{(0)}, \ldots, \mathbf{X}_n^{(0)}, \ldots, \mathbf{X}_N^{(0)}$, fixed textual prompt $\mathbf{P}_{\text{Desc}}$, a sample generator Gen($\cdot$), and a sample encoder Enc($\cdot$)
Output: Average GC@T metric over $\mathcal{D}$
Parameter: The number of iterations T

GC@T = 0
for $(n = 1; n \leq N; n++)$ **do**
  $\mathbf{z}^{(0)}$:= Enc($\mathbf{X}_n^{(0)}$);
  for $(t = 1; t \leq T; t++)$ **do**
    Generate description $\mathbf{Q}_t$ for $\mathbf{X}_n^{(t-1)}$ using (1);
    Create sample generation prompt $\mathbf{P}_{\text{Gen}}^{(t)}$;
    Generate a new sample $\mathbf{X}_n^{(t)}$ according to (2);
    $s^{(t)}$:= CosineSimilarity($\mathbf{z}^{(0)}$, Enc($\mathbf{X}_n^{(t)}$));
  end
  Calculate GC@T += $\sum_{t=1}^{T}(t \cdot s^{(t)}) / \sum_{t=1}^{T} t$; (3)
end
return GC@T / N;

the cost and complexity of dataset procurement, facilitating scalability. Moreover, this facilitates the use of previously unseen datasets for MLLM evaluation, minimizing the risk of training data contamination with evaluation data (Dodge et al., 2021). We will detail the GenCeption procedure and our initial experimental findings in the upcoming sections.

2 GenCeption

Our approach, GenCeption, is inspired by a multiplayer game DrawCeption[1] (a.k.a., Scrawl or Whispernary). In this game, the first player in a queue is presented with an image, which they describe verbally to the next player. This subsequent player then draws based on the description, and the cycle continues, often leading to amusing deviations from the original image as the game progresses. The challenge and objective of the game lie in preserving the initial information across iterative switches between two modalities: verbal description and drawing. Similarly, a proficient MLLM, which inherently models multiple modalities like text and images, should excel at playing such game, minimizing the semantic drift from the original input. Recognizing that MLLMs can encompass modalities beyond just visual cues, such as audio and graphs, we name our approach GenCeption, covering a broader scope than the visually-centric DrawCeption.

2.1 Procedure

Unlike existing MLLM benchmarks that rely on multimodal samples, GenCeption is designed to operate on unimodal datasets, significantly streamlining dataset acquisition efforts. For illustrative purposes, we employ the image modality as a representative non-textual modality throughout this exposition. Let's consider an image dataset $\mathcal{D}$ comprising images $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, akin to well-established datasets like ImageNet (Deng et al., 2009), CIFAR (Krizhevsky et al., 2009), and STL (Coates et al., 2011). Without loss of generality, any image from $\mathcal{D}$ is denoted as $\mathbf{X}$.

GenCeption operates iteratively, spanning from t=1 to a pre-defined maximum iteration t=T. Each iteration, as depicted in Figure 1, begins with an image $\mathbf{X}^{(t-1)}$, and yields a new image $\mathbf{X}^{(t)}$. The first iteration (t=1) commences with the original image $\mathbf{X}^{(0)}$ from $\mathcal{D}$. During any given iteration t, the MLLM receives a textual prompt $\mathbf{P}_{\text{Desc}}$ (Table 1), instructing the MLLM to articulate a comprehensive description $\mathbf{Q}_t$ for the input image $\mathbf{X}^{(t-1)}$:

$$\mathbf{Q}_t := \text{MLLM}(\mathbf{P}_{\text{Desc}}, \mathbf{X}^{(t-1)}). \quad (1)$$

Following this, an image generation prompt $\mathbf{P}_{\text{Gen}}^{(t)}$ is constructed as "*Generate an image that fully and precisely reflects this description*: <$\mathbf{Q}_t$>". This prompt guides a pretrained image generation model, such as DALL·E (Ramesh et al., 2021), to create a new image, $\mathbf{X}^{(t)}$:

$$\mathbf{X}^{(t)} := \text{Gen}(\mathbf{P}_{\text{Gen}}^{(t)}), \quad (2)$$

where Gen($\cdot$) signifies the chosen image generator. Each subsequent iteration t+1 commences by using the image $\mathbf{X}^{(t)}$ generated in the previous iteration. Upon completion of all iterations, we obtain a series of T+1 images: $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(T)}$, with the initial image being the original, and the rest sequentially produced across the iterations.

Please write a clear, precise, detailed, and concise description of all elements in the image. Focus on accurately depicting various aspects, including but not limited to the colors, shapes, positions, styles, texts and the relationships between different objects and subjects in the image. Your description should be thorough enough to guide a professional in recreating this image solely based on your textual representation. Remember, only include descriptive texts that directly pertain to the contents of the image. You must complete the description using less than 500 words.

Table 1: The fixed textual prompt $\mathbf{P}_{\text{Desc}}$ instructs the MLLM to produce a description of the input $\mathbf{X}^{(t-1)}$.

2.2 Metric: GC@T

Our primary objective is to measure the semantic divergence of each generated image $\mathbf{X}^{(t)}$ (for t=1, ..., T) from the original image $\mathbf{X}^{(0)}$. To

[1] https://wikipedia.org/wiki/drawception

(a) Correlations between GC@T, OpenCompass (OC), MME, and HallusionBench (HB) scores.

(b1) Seed image from the "color" category and its generated images for 3 VLLMs (GPT-4V, mPLUG-Owl2, LLaVA-7B&13B) over 5 GenCeption iterations.

(b2) Seed image from "OCR" category and its generated images for 3 VLLMs over 3 iterations.

Figure 2: Correlation analysis (a) and demonstration of GenCeption evaluation procedure on a visual-intensive image (b1) and a textual-intensive image (b2). The similarity $s^{(t)}$ and GC@T scores are printed on the top and bottom of each image, respectively.

achieve this, we utilize a pretrained image encoder, such as ViT (Dosovitskiy et al., 2021), to transform all images, resulting in T+1 image embeddings denoted as $\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)}$, where $\mathbf{z}^{(t)} := \text{Enc}(\mathbf{X}^{(t)})$. Afterwards, we compute the cosine similarity between $\mathbf{z}^{(0)}$ and each $\mathbf{z}^{(t)}$ (for t=1, . . . , T), yielding T similarity scores: $s^{(1)}, s^{(2)}, \ldots, s^{(T)}$. Here, $s^{(t)} \in [-1.0, 1.0]$ approximates the level of semantic drift observed in the t-th iteration of the aforementioned GenCeption procedure. To quantify the overall speed and magnitude of semantic drift, we propose to calculate the GenCeption score over T iterations, denoted as GC@$T \in [-1.0, 1.0]$, computed as follows:

$$\text{GC@}T := \textstyle\sum_{t=1}^{T}(t \cdot s^{(t)}) / \sum_{t=1}^{T} t. \qquad (3)$$

This is a normalized and continuous[2] metric that progressively weights later iterations more heavily for two reasons: (1) analogous to the DrawCeption game, it is the deviation from the initial image at the end that is most telling; (2) we aim to capture performance and dynamics across the entire iterative sequence. A high GC@T value signifies an exceptional and consistent ability to maintain inter-modal (text-image) semantic congruence, effectively curbing the propensity for rapid or extensive deviation from the semantics encapsulated in the original image. It is worth noting that GC@1 is equivalent to $s^{(1)}$. For the pseudo code detailing GenCeption procedure and the calculation of the average GC@T metric over the entire dataset $\mathcal{D}$, please see Algorithm 1.

[2]The GC@T metric progressively enhances with MLLM performance, counteracting the limitations of discontinuous metrics like accuracy prevalent in MLLM benchmarks that may falsely suggest emergent abilities (Schaeffer et al., 2023). This continuous metric facilitates more predictable projections of performance improvements resulting from model scaling, either through increased parameters or expanded training data.

3 Experiments

In this section, we embark on an empirical investigation of the GenCeption framework, focusing on its potential and implications for evaluating MLLMs, with a special focus on Vision LLM (VLLM), the predominant category in this area. Although GenCeption's innovative design merely requires unimodal image datasets, we choose to employ the most recent multimodal MLLM benchmark dataset – MME (Fu et al., 2023). This decision stems from two key considerations: (1) to allow for a direct comparison with metrics that incorporate additional textual QA (question-answering) annotations; and (2) to achieve a fine-grained assessment of MLLM performance across MME's 14 carefully crafted sample categories. We select four VLLMs – GPT-4V (Achiam et al., 2023), LLaVA-7B/13B (Liu et al., 2023b) and mPLUG-Owl2 (Ye et al., 2023) – based on their superior performance on the OpenCompass multimodal leaderboard (OpenCompass, 2023), which incorporates a comprehensive set of benchmarks like MME (Fu et al., 2023) and HallusionBench (Liu et al., 2023a). We will demonstrate GenCeption's efficacy through both quantitative and qualitative assessments, highlighting its validity and the correlations between unimodal and multimodal metrics.

3.1 Quantitative results

We partition the 14 MME categories into two groups based on content type: visual-intensive (10 categories) and textual-intensive (4 categories). GC scores and MME Accuracy are reported for each category in Table 2. Additionally, rankings for visual and textual intensive samples are compared against the OpenCompass multimodal leaderboard scores (OpenCompass, 2023) and HallusionBench (Liu et al., 2023a). Notably, GPT-4V leads

	Sample Category	GPT-4V				mPLUG-Owl2				LLaVA-13B				LLaVA-7B		
		ACC	GC@1	GC@3	GC@5	ACC	GC@1	GC@3	GC@5	ACC	GC@1	GC@3	GC@5	GC@1	GC@3	GC@5
visual-intensive samples	Existence	**96.67**	**0.505**	**0.422**	**0.358**	95.00	0.427	0.323	0.285	95.00	0.416	0.305	0.276	0.418	0.308	0.248
	Count	**86.67**	**0.498**	**0.404**	**0.360**	85.00	0.378	0.299	0.244	85.00	0.408	0.294	0.241	0.341	0.253	0.222
	Position	65.00	**0.501**	**0.408**	**0.347**	61.67	0.346	0.306	0.260	**76.67**	0.359	0.255	0.218	0.350	0.285	0.248
	Color	80.00	**0.506**	**0.403**	**0.325**	88.33	0.345	0.290	0.254	**90.00**	0.420	0.300	0.252	0.318	0.284	0.247
	Poster	**96.94**	**0.444**	**0.324**	**0.265**	86.73	0.338	0.243	0.210	86.39	0.303	0.215	0.176	0.305	0.214	0.182
	Celebrity	0.00	**0.433**	**0.332**	**0.284**	**87.94**	0.319	0.232	0.197	83.53	0.284	0.206	0.176	0.263	0.188	0.154
	Scene	83.50	**0.497**	**0.393**	**0.337**	83.25	0.385	0.299	0.252	**86.75**	0.355	0.277	0.230	0.350	0.266	0.223
	Landmark	79.25	**0.458**	**0.353**	**0.302**	85.74	0.363	0.275	0.223	**90.00**	0.376	0.242	0.191	0.334	0.252	0.215
	Artwork	**82.00**	**0.504**	**0.421**	**0.363**	77.25	0.333	0.252	0.211	70.75	0.308	0.212	0.166	0.294	0.210	0.176
	Comm.	**79.29**	**0.563**	**0.471**	**0.405**	71.43	0.425	0.353	0.290	73.57	0.429	0.334	0.273	0.417	0.294	0.235
	Vis mean	74.93	**0.491**	**0.393**	**0.335**	82.23	0.366	0.287	0.243	**83.77**	0.366	0.264	0.220	0.339	0.255	0.215
	Vis rank	3	**1**	**1**	**1**	2	2	2	2	**1**	2	3	3	4	4	4
text-intensive	Code.	**90.00**	**0.333**	**0.193**	-	45.00	0.281	0.176	-	42.50	0.260	0.144	-	0.186	0.107	-
	Num.	**75.00**	0.325	**0.240**	-	35.00	0.322	0.192	-	37.50	**0.336**	0.195	-	0.259	0.155	-
	Text trans.	55.00	**0.359**	**0.157**	-	**67.50**	0.173	0.081	-	57.50	0.200	0.116	-	0.212	0.111	-
	OCR	**95.00**	**0.482**	**0.393**	-	45.00	0.358	0.276	-	75.00	0.368	0.239	-	0.351	0.222	-
	Txt Mean	**78.75**	**0.375**	**0.246**	GC rank*	48.13	0.284	0.181	GC rank*	53.13	0.291	0.174	GC rank*	0.252	0.149	GC rank*
	Txt rank	**1**	**1**	**1**	**1.00**	3	3	2	2.14	2	2	3	2.62	4	4	4.00
	HallusionBench†	score: 46.5, rank: **1**				score: 25.7, rank: 4				score: 29.4, rank: 2				score: 27.4, rank: 3		
	OpenCompass†	score: 64.2, rank: **1**				score: 47.8, rank: 3				score: 49.7, rank: 2				score: 46.8, rank: 4		

* "GC rank" for each VLLM is a weighted (by the number of categries) average of blue-colored "Vis rank" and "Txt rank", i.e., $\frac{10}{14} \times \overline{\text{vis\_ranks}} + \frac{4}{14} \times \overline{\text{txt\_ranks}}$.
† Results are taken from https://rank.opencompass.org.cn/leaderboard-multimodal as of Feb. 2024.

Table 2: Evaluation results on visual(Vis)-intensive (*existence, count, position, color, poster, celebrity, scene, landmark, artwork, and commonsense reasoning*) and textual(Txt)-intensive (*code reasoning, numerical calculation, text translation, and OCR*) sample categories. Best results per metric and category are **bolded**.

our rankings, followed by mPLUG-Owl2, LLaVA-13B/7B, diverging from MME scores but aligning with HallusionBench and OpenCompass rankings.

Figure 2(a) presents a correlation matrix among GC@T, MME, OpenCompass, and HallusionBench scores, where the "GC@T" is averaged over the GC@T scores of all MME categories. It reveals a strong correlation between GC@T and HallusionBench, indicating effective hallucination measurement without human annotation or multimodal data. Further, the moderately strong correlation with OpenCompass suggests GenCeption's comprehensive evaluation capability. The negative correlation with MME scores suggests that GenCeption measures distinct aspects not covered by MME, using the same set of samples.

3.2 Qualitative results

We conduct a qualitative inspection by visualizing artifacts (descriptions and images) alongside cosine similarity and GC@T scores for two seed images across different categories, as shown in Figure 2(b). This visualization reveals a correlation between these scores and the images' visual characteristics in relation to the seed image. A notable observation is the addition of nonexistent elements or styles to the generated images, a trend that intensifies with subsequent iterations. For a broader spectrum of examples across all MME image categories and accompanying descriptions from each evaluated VLLM, we direct readers to Appendix A. It is apparent that later iterations exhibit an increased propensity for producing unreal imagery.

4 Conclusion and Future Work

To enable scalable and continuous evaluation of rapidly evolving MLLMs without relying on expensive annotated multimodal benchmark datasets, we propose GenCeption, an intuitive, simple and effective approach. Our preliminary tests on VLLMs demonstrate that the GC@T metric proficiently assesses semantic coherence and consistency across modalities, aligning closely with results from existing comprehensive MLLM benchmarks. Looking ahead, future work includes: (1) Broadening its application across all VLLM benchmark datasets to comprehensively understand its capabilities. (2) Adapting GenCeption for various modalities, such as audio and graphs, by selecting modality-specific generation and embedding models. (3) Enhancing understanding through comparisons with human performance on GenCeption tasks. (4) Tailoring MLLM prompts to different sample categories for nuanced analysis. (5) Improving similarity metrics by incorporating object recognition models to better quantify sample distances. (6) Directly leveraging sample descriptions in similarity score calculations for a more inclusive evaluation.

References

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

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint:2305.06500*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of closed-source large language model. *arXiv preprint arXiv:2305.12870*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. Technical report, Massachusetts Institute of Technology and New York University.

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

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

OpenCompass. 2023. OpenCompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint:2305.11175*.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint:2212.10773*.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint:2311.04257*.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. *arXiv preprint:2305.16934*.

A GenCeption Demonstration

To provide a comprehensive, intuitive and qualitative understanding of the GenCeption procedure and GC@T metric, we illustrate the input, output, intermediate artifacts, similarity scores, and GC@T values throughout the GenCeption process. An example from one of the 14 MME image categories is showcased in Figures 1 to 12 of our supplementary material that needs to be downloaded separately.

B Limitations and Societal Impact

The limitations, outlined in Sections 3 and 4, primarily pertain to our initial experimental focus on image-based experiments, excluding other modalities. A critical assumption is the minimal influence of stochastic variability in image generation and MLLM text generation processes. While we have not delved into ethical risks, our framework's purpose – to assess inter-modality semantic drift and susceptibility to hallucination in MLLMs—is clearly articulated. Societally, the exclusive use of the English language in GenCeption experiments may inadvertently marginalize non-English-speaking user groups.

C Dataset and Reproducibility

In Sections 1, 2.1, 2.2 and 3 of the main paper, we cite the creators of all artifacts used. Detailed citations can be found in references. The MME dataset is not directly downloadable, and is released for research purposes only upon a request from authors to gain access to it. We followed the guidelines provided by the authors and respected the intended terms of use. The specific licenses and terms for the use and distribution of publicly available artifacts can be found in the corresponding original papers or GitHub repositories, as cited. As per this research work and aligning with the MME copyrights, we are not releasing this asset. Regarding the created artifacts, we introduce a new metric called GC@T, and detail its creation and intended use in Section 2.2 of the main paper. Our study exclusively utilizes images from the MME dataset, omitting textual QA annotations, and generates textual data in the form of English descriptions as part of our methodology. Given the nature of our research centered on quantifying the inter-modality coherence and consistency, we do not use or report any statistics related to the data splits. The metrics reported in Table 2 are from a single run.

In our study, we adopt several state-of-the-art models to facilitate our experiments, including GPT-4V, LLaVa-13B, LLaVa-7B, and mPLUG-Ow12 for text description generation, ViT for image embedding, and DALL·E 3 for image generation, adhering to default parameter settings as outlined in their original specifications. We set the temperature parameter (whenever relevant) to 0 in both the MLLM and DALL-E 3 models to minimize the stochasticity inherent in these models' outputs. The text descriptions generated by GPT-4V are obtained through API calls, while experiments involving the other models are conducted on A100 GPUs, totaling approximately 96 GPU hours. Image generation was also performed via a call to OpenAI's DALL-E 3 API. To compute the GC@T metric, we employ the cosine similarity metric from the Scikit-learn library (Version 1.4.0).