TokenSwap: Benchmarking and Reducing the Modality Gap in Multimodal LLMs

Andong Hua[1] **Colton Bishop**[2] **Igor Mordatch**[2] **Arian Hosseini**[2]
Jindong Gu[2] **Aleksandra Faust**[2] **Rebecca Roelofs**[2] **Yao Qin**[1,2]

[1]University of California, Santa Barbara [2]Google DeepMind

Abstract

Multimodal large language models (MLLMs) should exhibit cross-modal consistency, generating consistent responses given semantically equivalent inputs across modalities. However, we observe a systematic discrepancy in model predictions under such cross-modal variations. Specifically, we define the modality gap as the difference in model performance under semantically equivalent textual and multimodal inputs. We introduce TokenSwap, a method that constructs such inputs by replacing textual concepts with semantically aligned images, resulting in sequences where visual tokens are interleaved with text tokens. Based on TokenSwap, we transform existing text-based benchmarks (e.g., MMLU [17]) into image-interleaved counterparts, resulting in TOKENSWAP-BENCH. Across 42 MLLMs, we observe a pervasive modality gap, with performance decreasing by 4.2% to 47.4% when moving from text-only to image-interleaved inputs, averaging 19.6% $\pm$ 3.3% across models. Notably, reasoning models exhibit consistently smaller gaps, achieving an average gap of 10.1% compared to 25.5% for non-reasoning models, suggesting greater robustness in multimodal reasoning across equivalent textual and visual representations. In contrast, neither prompting strategies nor scaling training compute alone reliably reduces the modality gap. Finally, we demonstrate that incorporating TokenSwap during training effectively mitigates this gap, improving cross-modal alignment while preserving strong text-only and vision-language model performance. The project page is available at `https://dongxzz.github.io/TokenSwap/`.

1 Introduction

Multimodal Large Language Models (MLLMs) are expected to be semantically invariant, producing consistent predictions across semantically equivalent inputs regardless of whether they are presented as pure text or with interleaved images [42, 37, 7]. Yet, this consistency across modalities often breaks in practice. Consider the example in the Final Benchmark panel in Figure 2: the image-interleaved version is semantically aligned with the original, but key concepts (e.g., "cards", "aces", "gas tank") have been replaced by corresponding images. Of 42 evaluated MLLMs, 11 answer the text-only version correctly but fail on the image-interleaved input, revealing a systematic violation of semantic invariance.

We formalize this phenomenon as the *modality gap*, defined as the difference in model performance when semantically equivalent content is presented in textual versus multimodal form. To systematically study this gap, we introduce TokenSwap, a general method for transforming text-only datasets into image-interleaved counterparts while preserving semantic alignment across modalities. Rather than converting entire inputs into images (e.g., rendering text as images [48, 49]), TokenSwap operates at the concept level, replacing individual textual concepts with semantically aligned natural images while preserving the surrounding context and structure (Figure 2).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

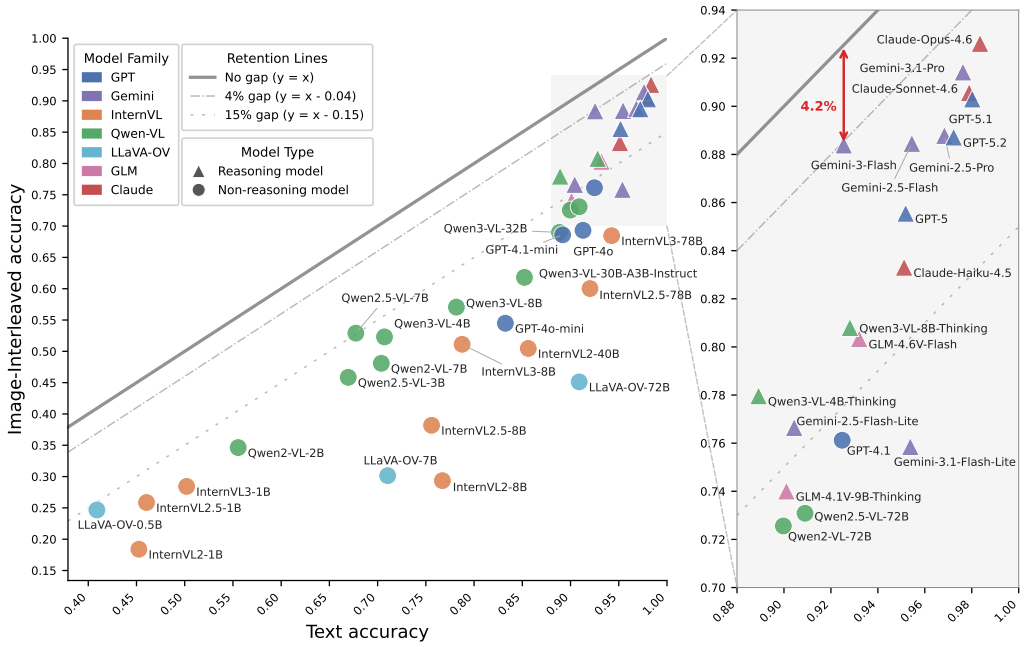


Figure 1: **A pervasive modality gap exists: all models perform worse on image-interleaved inputs than on text-only inputs, with gaps ranging from 4.2% to 47.4%.** Each point represents a model, colored by family, with triangles indicating reasoning models. The diagonal $y = x$ denotes equal performance, and dashed lines indicate constant gap levels (4.0%, and 15.0%). All models lie below the diagonal, indicating consistently lower performance under image-interleaved inputs. The right panel zooms into high-performing models. Gemini-3-Flash achieves the smallest gap (4.2%). Reasoning models typically exhibit smaller gaps (4.0%–15.0%), while non-reasoning models tend to have substantially larger gaps.

We apply TokenSwap to MMLU [17] to construct TOKENSWAP-BENCH, a benchmark containing 1,516 samples with 6,946 total images, specifically designed to quantify the modality gap. Across 42 state-of-the-art MLLMs, we observe a clear modality gap: performance drops from text-only to image-interleaved inputs range from 4.2% to 47.4%, with an average decrease of 19.6% ± 3.3% (95% confidence interval), as shown in Figure 1. Taking a closer look, we discover that reasoning-oriented models exhibit substantially smaller gaps (10.1% on average) compared to non-reasoning models (25.5%). We further find that common prompting strategies, such as chain-of-thought and few-shot prompting cannot reliably reduce the modality gap. These findings suggest that the smaller gaps of reasoning-oriented models are more likely linked to special training-time mechanisms. In addition, as shown in Figure 4, scaling training compute yields only marginal gains, with a $10\times$ increase in FLOPs reducing the gap by approximately 2.8%.

Finally, unlike prior work that is primarily diagnostic [48, 49, 37], we show that augmenting text data with image-interleaved counterparts via TokenSwap effectively mitigates the modality gap in both pre-training and post-training settings. To our knowledge, this is the first training-based approach to mitigate the modality gap for multimodal LLMs. Importantly, these improvements do not come at the cost of either text-only or vision-language performance, and in some cases even lead to slight gains. Our contributions are summarized as follows:

1. We propose TokenSwap, a data-centric method for constructing image-interleaved inputs by replacing textual concepts with semantically aligned images, and introduce TokenSwap-Bench, a benchmark for systematically quantifying the modality gap in MLLMs.

2. We evaluate 42 MLLMs and find the modality gap to be pervasive. Reasoning models consistently exhibit smaller gaps, while prompting strategies and scaling alone provide limited improvement.

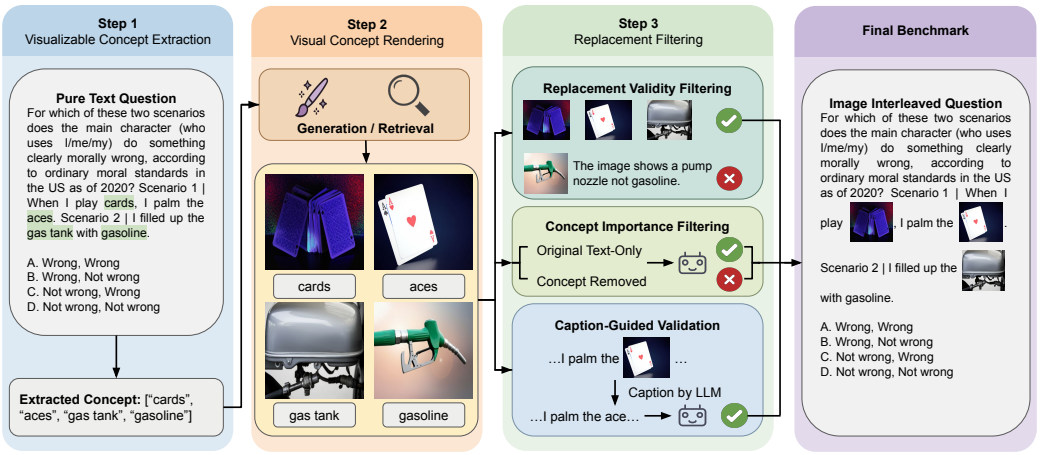


Figure 2: **TokenSwap constructs semantically aligned image-interleaved questions, where Step 1 panel shows the original MMLU question and the Final Benchmark panel shows its image-interleaved counterpart.** We first extract visualizable concepts and generate or retrieve images representing them, followed by rigorous filtering to ensure validity, task relevance, and semantic consistency, resulting in semantically aligned image-interleaved questions.

3. We demonstrate that TokenSwap training effectively reduces the modality gap at both pre-training and post-training stages, while preserving strong text-only and vision-language performance.

2 Related Works

Modality Gap and Cross-Modal Consistency. The notion of *modality gap* has been extensively studied in contrastive vision–language models such as CLIP [31, 20, 9, 36, 47, 39], where image and text representations may remain separated despite being embedded in a shared space, leading to representation-level discrepancies [22, 34, 12, 44, 13, 32]. Related work on multimodal large language models (MLLMs) instead examines *cross-modal consistency*, evaluating whether models produce consistent predictions when the same semantic content is presented in different modalities [7, 46, 1, 48, 49, 40, 37, 42]. These studies reveal that MLLMs often exhibit inconsistent behavior across modalities, influenced by factors such as rendering choices, visual attributes, and domain structure. Additional details on related work are provided in Appendix D.

Compared with these works, our setting is distinct in two key aspects. First, prior studies typically focus on whole-input modality conversion, rendered-text settings, or highly structured domains with standardized symbolic notations, whereas we consider an open-domain setting where natural images are interleaved into textual inputs. Notably, TokenSwap enables transforming arbitrary existing textual benchmarks into their image-interleaved counterparts. As discussed in Section 4.7, our benchmark provides a complementary perspective on modality gap compared to prior benchmarks. Second, most prior work is primarily diagnostic, with limited exploration of mitigation strategies. In contrast, we not only conduct a comprehensive evaluation across 42 MLLMs, including recent reasoning models, but also demonstrate that the modality gap can be effectively reduced through TokenSwap-based training.

3 Quantifying Modality Gap: TokenSwap-Bench

To systematically study the modality gap in Multimodal Large Language Models (MLLMs), we introduce TokenSwap-Bench, a benchmark designed to measure the performance discrepancy when the same semantic information is presented in different modalities. Unlike conventional vision-language benchmarks [24, 19, 35, 14, 45] that evaluate models on multimodal understanding with complementary information, our benchmark instead focuses on measuring the modality gap under semantic equivalence.

3.1 Formalizing Modality Gap via TokenSwap

We study modality gap under semantic equivalence: an ideal MLLM should make consistent predictions when the same content is presented in textual or visual form. Given a text input X_{text}, TokenSwap replaces selected visualizable concepts (e.g., words or phrases) with semantically aligned images, producing an image-interleaved input $X_{\text{interleaved}}$ while preserving the surrounding textual context, as illustrated in Figure 2. We provide a more detailed token-level formalization of TokenSwap in Appendix A.1. We quantify the modality gap as

$$\Delta_{\text{Gap}} = \text{Eval}(X_{\text{text}}) - \text{Eval}(X_{\text{interleaved}}), \tag{1}$$

where $\text{Eval}(\cdot)$ denotes the model's performance metric (e.g., accuracy). $\Delta_{\text{Gap}} > 0$ indicates the presence of a modality gap, while larger values correspond to a more substantial modality gap.

3.2 Benchmark Construction

We construct TOKENSWAP-BENCH by applying TokenSwap to MMLU [17]. As illustrated in Figure 2, the construction follows three stages. The final benchmark contains 1,516 samples with 6,946 image replacements, averaging 4.58 replacements per sample. Representative examples are shown in Figure 15.

3.2.1 Visualizable Concept Extraction

We use Gemini-2.0-Flash [38] to identify concepts in each question and answer option that can be visually represented, typically concrete entities and visually grounded phrases. The full prompt is provided in Appendix A.4.

3.2.2 Visual Concept Rendering

We render each visualizable concept into an image using a text-to-image model. Specifically, given a concept c, we synthesize a corresponding image I_c with Gemini using the prompt: *"generate an image of a* $\{c\}$*"*.

3.2.3 Replacement Filtering

To ensure that measured gaps are not driven by poor or irrelevant substitutions, we apply a rigorous sequence of filtering procedures. Across all filters, we consistently use Gemini-2.0-Flash [38] as the underlying LLM for filtering and as a proxy model. We provide additional implementation details in Appendix A.2.

Validity filtering: We use an LLM to verify that the generated image faithfully represents the intended concept in context. Only replacements that preserve the original meaning are retained.

Importance filtering: We ensure that the replaced concepts are task-relevant. Let $\mathcal{C}$ denote the set of valid textual concepts in a sample, and let $X_{\text{text}}^{-\mathcal{C}}$ be the input obtained by removing them from the original text. We retain only samples satisfying:

$$\text{Eval}(X_{\text{text}}) = 1 \quad \text{and} \quad \text{Eval}(X_{\text{text}}^{-\mathcal{C}}) = 0. \tag{2}$$

This ensures that the selected concepts collectively affect the model's prediction.

Caption-guided validation: We verify semantic recoverability by captioning each substituted image and replacing the visual tokens in $X_{\text{interleaved}}$ with the generated captions, yielding $\tilde{X}_{\text{text}}$. We retain only samples satisfying:

$$\text{Eval}(X_{\text{text}}) = 1 \quad \text{and} \quad \text{Eval}(\tilde{X}_{\text{text}}) = 1. \tag{3}$$

This round-trip process reduces the chance that the measured gap is caused by unrecognizable or semantically mismatched images.

We use Gemini-2.0-Flash as the underlying model for all filtering stages. To avoid directly evaluating the same model used as the proxy, we exclude Gemini-2.0-Flash from our model comparison. To test

whether our results are sensitive to the choice of filtering proxy, we further rebuild TOKENSWAP-BENCH with GPT-5-mini as the proxy, rerunning the two stages that depend most on the proxy: importance filtering and caption-guided validation. Across nine evaluated models, the ranking is unchanged (Spearman $\rho = 1.00$, Kendall $\tau = 1.00$, Pearson $r = 0.99$), and Gemini-3-Flash still achieves the smallest modality gap. Full results are provided in Appendix A.3.

3.3 Semantic Equivalence and Human Validation

TokenSwap relies on semantic equivalence between textual concepts and their visual counterparts. In practice, perfect equivalence is not achievable because textual concepts are abstract but images are concrete instances. We therefore use *semantic recoverability* as a practical proxy: the substituted image should convey information that can be recovered as text and preserve the task prediction. In particular, validity filtering and caption-guided validation explicitly ensure this recoverability.

We further validate substitution quality through a human study. We randomly sample one question from each of the 57 subjects in TOKENSWAP-BENCH, covering 213 substituted concepts. Two annotators achieve 92.0% (196/213) and 89.2% (190/213) accuracy in identifying the intended concept from the image, both substantially above the 25% random baseline, with 93.4% inter-annotator agreement. These results suggest that the substituted images reliably convey the intended concepts. Details are provided in Appendix A.5.

4 Analyzing Modality Gap: Results and Insights

4.1 Experimental Setup

Models. We evaluate **42 models**, covering a wide range of scales from sub-billion ($\sim$0.5B) to tens-of-billions (e.g., 78B) parameters, as well as larger-scale proprietary models beyond this range. These include state-of-the-art open-source MLLMs such as Qwen-VL (2/2.5/3) [41, 6, 5], InternVL (2/2.5/3) [8, 50], LLaVA-OV [21], and GLM [18], as well as proprietary models including GPT 4o/4.1/5/5.1/5.2 [26–30], Gemini 2.5/3/3.1 [11, 16], and Claude 4.5/4.6 [2, 4, 3]. Our selection spans diverse training paradigms, including instruction-tuned, reasoning, and mixture-of-experts (MoE) [33] models. This diverse pool enables a systematic evaluation of modality gap across architectures, scales, and capabilities.

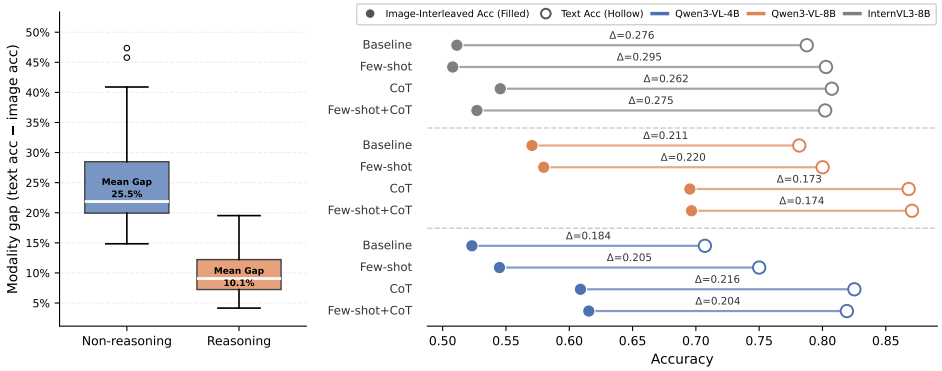
Evaluation Protocol. For each sample in TOKENSWAP-BENCH, we construct two versions: (1) a text-only input X_{text}, and (2) an image-interleaved input $X_{\text{interleaved}}$ obtained via TokenSwap. We evaluate model accuracy on both versions, referred to as text accuracy and image-interleaved accuracy, respectively. We quantify the modality gap as defined in Eq. 1, i.e., the difference between text accuracy and image-interleaved accuracy. Appendix B provides additional methodological details, including prompt templates and answer extraction methods, as well as additional results, including per-class analysis, results by the number of image replacements, and full numerical results for all models.

Beyond TOKENSWAP-BENCH, we also apply TokenSwap to GSM8K [10], a free-form reasoning benchmark, and the details are provided in Appendix B.2.

4.2 Modality Gap Persists Across Models and Scales

All models exhibit a non-trivial modality gap. As shown in Figure 1, all evaluated models lie strictly below the $y = x$ line, indicating that replacing text with semantically aligned images consistently degrades performance. Notably, all models fall below the 4.0% retention line, demonstrating a non-trivial modality gap. Even the best-performing model, Gemini-3-Flash, exhibits a gap of **4.2%**, while weaker models such as InternVL2-8B suffer substantially larger drops (up to **47.4%**).

Modality gap varies across model families. We further examine the modality gap across model families and observe consistent trends. Proprietary models, such as Claude, exhibit the smallest gaps (mean gaps 8.3%), while open-source models generally show larger gaps, with InternVL2 and LLaVA-OV often exceeding 35.0%. Although Qwen3-VL shows relative improvements (mean gaps $\sim$17.6%), it still lags behind proprietary systems.



(a) Modality gap: reasoning vs. non-reasoning

(b) Performance under different prompting strategies

Figure 3: **Reasoning models exhibit a smaller modality gap, while prompting strategies fail to reduce it.** (a) Distribution of modality gap for reasoning and non-reasoning models. (b) We compare different prompting strategies. Each pair shows text accuracy (hollow) and image-interleaved accuracy (filled), with the modality gap Δ indicated.

Modality gap increases with the number of image replacements. We find that the modality gap generally increases as more textual concepts are replaced by images, rising from 13.5 percentage points with one image to around 20 percentage points with more replacements (Figure 13). Detailed analysis are provided in Appendix B.5.

Modality gap is not explained by visual-input confounds. Replacing text with images introduces two potential confounds. First, images contain substantially more visual tokens than the text they replace, so longer inputs may cause the performance drop. To rule out this factor, we restore the original text and add pure-white images of the same size. Accuracy remains within ± 1.0 percentage point of text-only performance (Table 4), indicating that increased input length alone cannot explain the modality gap.

Second, the drop may come from processing information through the visual pathway itself. We then replace each textual concept with a rendered-text image containing the same text. The resulting gap is substantially smaller than that with natural images (Table 5), indicating that most of the degradation arises from recognizing natural visual concepts rather than from routing the input through the visual modality itself.

Additional experimental details and full numerical results for both controls are provided in Appendix B.6.

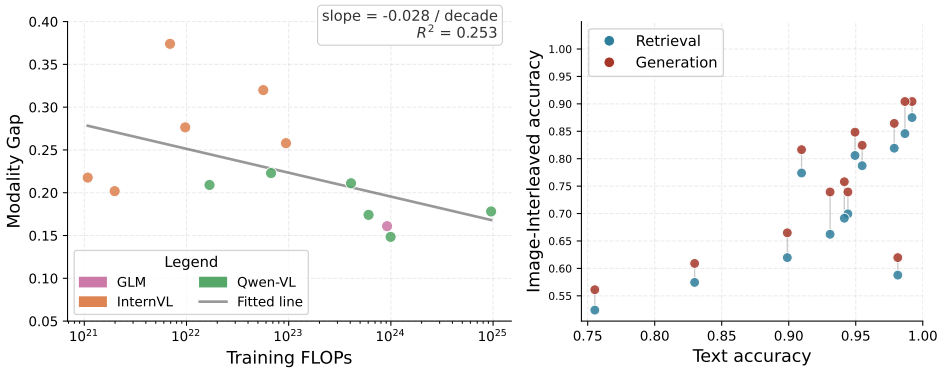
4.3 Reasoning Models Exhibit Smaller Modality Gap

We observe a clear distinction between reasoning and non-reasoning models. Reasoning models not only achieve stronger performance on both text and image-interleaved inputs, but also exhibit consistently smaller modality gaps. As shown in Figure 3(a), most reasoning-oriented models cluster within a 4.0%–15.0% gap range, whereas non-reasoning models typically exhibit larger gaps, often exceeding 15.0%. On average, reasoning models exhibit substantially smaller modality gaps than non-reasoning models (10.1% ± 2.2% vs. 25.5% ± 3.7%), a difference that is statistically significant (Welch's t-test: $t = -7.48$, $p < 0.001$).

This trend is consistent when considering the **Relative Modality Gap**, which normalizes the absolute gap by text-only performance:

$$\Delta_{\text{Rel Gap}} = \frac{\text{Eval}(X_{\text{text}}) - \text{Eval}(X_{\text{interleaved}})}{\text{Eval}(X_{\text{text}})}. \tag{4}$$

As shown in Figure 12, reasoning models still exhibit smaller relative modality gaps. This suggests that the observed advantage is not solely attributable to stronger text capabilities, and may be partly associated with reasoning-oriented training.



(a) Absolute modality gap

(b) Retrieval vs. Generation

Figure 4: **Scaling provides limited improvement in modality gap, and generation-based benchmarks exhibit smaller gaps.** (a) plots absolute modality gap versus training FLOPs. FLOPs are collected for all open-source non-reasoning models with publicly available training compute estimates, based on original papers and Epoch AI[1]. (b) Each blue/red pair corresponds to the same model evaluated with retrieved versus generated images, sharing identical text accuracy since only the image source differs. Red points consistently lie above blue ones, indicating that generated images yield higher image-interleaved accuracy and smaller modality gaps.

4.4 Prompting Strategies Do Not Consistently Reduce Modality Gap

A natural question is whether improved prompting strategies can reduce the modality gap. To investigate this, we evaluate three representative non-reasoning MLLMs on TOKENSWAP-BENCH under several prompting strategies, including Chain-of-Thought (CoT), few-shot prompting, and their combination. The corresponding prompt templates are provided in Appendix B.

As illustrated in Figure 3(b), while these prompting strategies improve performance in both text-only and image-interleaved settings, they do not reliably reduce the modality gap. The effect is inconsistent across models: in some cases the gap even increases, while in others it decreases marginally. For example, under CoT prompting, the gap widens by approximately 3.2% for Qwen3-VL-4B-Instruct, but shrinks by a similar margin for the 8B variant.

Moreover, even with advanced prompting, these models still lag significantly behind their reasoning counterparts (e.g., Qwen3-VL-4B/8B-Thinking), which exhibit clearly smaller modality gaps of 10.9% and 12.4%, respectively. This suggests that inference-time prompting alone is insufficient to bridge the modality gap, and that reducing the modality gap likely requires changes at the training level rather than prompting.

4.5 Scaling Alone Provides Limited Improvement in Modality Gap

We further analyze the relationship between training compute and modality gap by plotting the training FLOPs of all open-source models with publicly available compute estimates, as shown in Figure 4(a). The correlation is weak, with a low $R^2 = 0.253$ for the fitted regression line, indicating that scaling explains only a limited portion of the variance in modality gap.

Interestingly, a similar trend with a better fit holds when considering the relative modality gap, as described in Appendix B.3. In both cases, the negative slope suggests that increasing training compute reduces the modality gap. However, the magnitude of this reduction remains limited: a $10\times$ increase in training FLOPs reduces the absolute modality gap by only 2.8%.

Overall, these results suggest that scaling alone provides limited improvement in mitigating the modality gap. While the models included in this analysis are not the latest models due to limited public availability of training compute information, we believe this trend likely extends to newer frontier models as well.

[1] https://epoch.ai/data/ai-models/

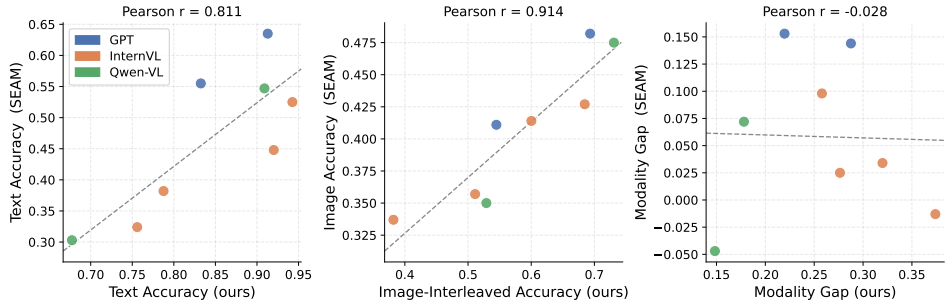


Figure 5: **TOKENSWAP-BENCH is complementary to the existing benchmark.** We plot 8 overlapping models, with SEAM [37] results taken from the original paper; numerical values are provided in the Appendix B.7. We observe high Pearson correlation with SEAM in both text accuracy ($r = 0.811$) and image-interleaved accuracy ($r = 0.914$), but near-zero correlation in modality gap ($r = -0.028$).

4.6 Retrieval-Based Benchmarks Exhibit Larger Modality Gap than Generation-Based Ones

In addition to the benchmark constructed using generated images, we construct a retrieval-based benchmark variant to study how image sourcing affects the measured modality gap. Specifically, for each concept c, we retrieve images from the DataComp-Small [15] dataset using CLIP ViT-L/14 [31] embeddings with the prompt: *"An image describing* $\{c\}$*"*. We retrieve the top-5 nearest images and retain those with cosine similarity greater than 0.3.

To ensure a fair comparison, we keep all other factors identical across the two benchmark variants and vary only the image source. In practice, we construct a matched subset of 376 samples for which both retrieved and generated images are available. Figure 4(b) shows model performance under both settings.

Across all models, we observe a consistent modality gap, where image-interleaved accuracy remains lower than text-only accuracy regardless of how the images are obtained. Furthermore, while absolute scores vary, the ranking of different models stays largely invariant across both benchmark settings. This suggests that the gap is intrinsic and persists across different input constructions.

However, the magnitude of the gap varies with the image source. Generation-based benchmarks consistently achieve 4.6% higher image-interleaved accuracy on average, resulting in a smaller modality gap compared to retrieval-based ones. We attribute this to the fact that generated images are typically more focused and better aligned with the replaced concept, whereas retrieved images often contain additional irrelevant or distracting details despite passing our validity filtering. Figure 9 provides representative examples illustrating these differences between generated and retrieved images. These results indicate that the observed modality gap is not solely a property of the model, but is also influenced by benchmark construction and image quality.

4.7 Complementary to Existing Benchmarks

We compare our benchmark with SEAM [37], a recent benchmark that evaluates cross-modal reasoning consistency using semantically equivalent inputs across modalities. Instead of using natural images for semantic substitution in TOKENSWAP-BENCH, SEAM focuses on textual-symbolic domains such as chess, chemistry, music, and graph theory.

As illustrated in Figure 5, we observe a strong positive correlation with SEAM for both text-only and image accuracies, indicating consistent performance trends across benchmarks. In contrast, the modality gap exhibits near-zero correlation. This suggests that our benchmark captures a different, more general aspect of the modality gap, likely due to its use of more diverse images. In contrast, SEAM focuses on structured, domain-specific representations, making the two benchmarks complementary to each other.

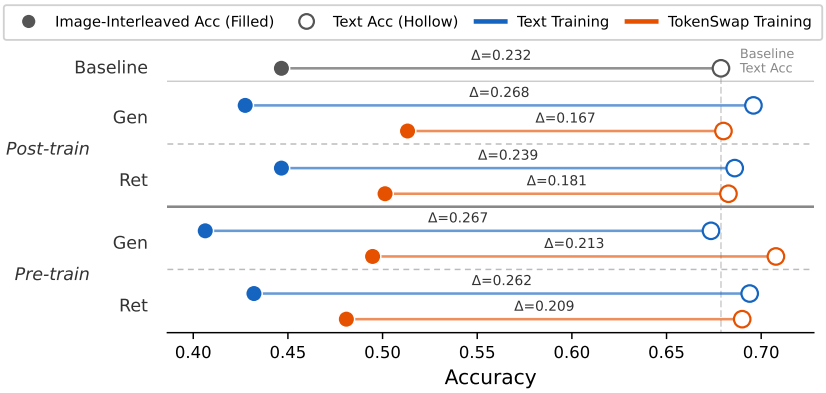


Figure 6: **TokenSwap reduces the modality gap while preserving text performance.** We compare Text training (blue) and TokenSwap training (red) across both pre-training and post-training, using either generated (Gen) or retrieved (Ret) images. Text training includes Gen and Ret variants, as it is paired with the TokenSwap variants constructed from the same samples. Each pair shows text accuracy (hollow) and image-interleaved accuracy (filled), with the modality gap Δ indicated.

5 Reducing Modality Gap: TokenSwap Training

To mitigate the modality gap, we apply TokenSwap during training by augmenting text-only data with image-interleaved counterparts constructed through visual concept substitution.

5.1 Experimental Setup

We construct TokenSwap training data from Magpie-Pro [43] by replacing visualizable concepts in user queries with images while keeping assistant responses unchanged. This yields 116,722 samples with an average of 1.75 images per sample. To isolate the effect of visual substitution, we construct paired text-only and TokenSwap variants using the same underlying samples.

We use Qwen2-VL-7B as the base model and evaluate three settings. **Baseline** denotes training only on the original LLaVA post-training data. **Gen** and **Ret** denote additional training with data constructed using generated and retrieved images, respectively. For each Gen/Ret setting, we compare TokenSwap training with a paired text-only variant that uses the same samples but without replacing concepts with images. We evaluate this comparison under both post-training and continuous pre-training. Additional data construction details are provided in Appendix C.1.

5.2 Main Results

Figure 6 illustrates the performance of models trained with TokenSwap training across various training stages within the TOKENSWAP-BENCH. Our observations are as follows:

TokenSwap reduces the modality gap. TokenSwap significantly mitigates the modality gap compared to both the baseline and text training. For example, in post-training with generated images, the gap decreases from 0.268 to 0.167. Similar reductions are observed in pre-training and with retrieved images, indicating that TokenSwap training effectively reduces the modality gap.

TokenSwap preserves text-only and vision-language performance. TokenSwap training maintains comparable text accuracy and in some settings slightly improves it. Standard vision-language benchmarks also show no noticeable degradation, as reported in Appendix C.5. These results indicate that the reduction in modality gap does not come at the expense of either textual or general multimodal capabilities.

Text-only training exacerbates the modality gap. Interestingly, we observe that training on text-only data consistently increases the modality gap. For instance, in post-training, the gap increases from 0.232 for the baseline to 0.268 after text-only training. This suggests that text-only training further biases the model toward textual modality, thereby widening the modality gap.

Finally, to examine whether the effectiveness of TokenSwap training generalizes beyond Qwen2-VL-7B, we replicate the post-training experiment on InternVL3-8B-Pretrained using a smaller-scale subset of the same training data. TokenSwap training reduces the gap from 0.303 to 0.230, and the improvement comes largely from the image-interleaved side (0.438 $\rightarrow$ 0.518), while text-only accuracy is almost unchanged (0.741 $\rightarrow$ 0.748). These results indicate that the effectiveness of TokenSwap training is not specific to Qwen2-VL-7B and generalizes to a second model family. Full results are provided in Appendix C.2.

5.3 Generalization to Text Recognition Task

As an additional case study beyond natural-image substitution, we apply TokenSwap to text recognition using the IIIT 5K-Word dataset [25]. In this setting, the same word can appear either as text or as a rendered image, making text recognition a special case of the modality gap under semantic equivalence. We construct TokenSwap training samples by replacing target words in text prompts with their corresponding word images. Detailed data construction and experimental setup are provided in Appendix C.3.

Table 1: **TokenSwap training improves OCR performance.** We report accuracy on the IIIT 5K-Word dataset.

Method	Accuracy
Baseline	0.9003
Text Training	0.8947
TokenSwap Training	**0.9270**

As shown in Table 1, TokenSwap improves OCR accuracy from 90.03% to 92.70%, outperforming both the baseline and text-only augmentation. This suggests that TokenSwap training can benefit related cross-modal recognition settings beyond our benchmark.

6 Conclusion

In this paper, we introduce TOKENSWAP-BENCH for quantitative evaluation of modality gap and find that all models exhibit a non-trivial gap, with at least a 4% drop under image-interleaved inputs. To mitigate this, we propose TokenSwap as a training strategy that constructs image-interleaved data, effectively reducing the modality gap while preserving text performance.

Our approach has several limitations. The measured gap depends on image quality, where relatively lower-quality (e.g., retrieval-based) images can introduce a bigger modality gap, although model rankings remain consistent. In addition, TokenSwap training requires in-distribution data: domain mismatch between training and evaluation (e.g., natural images vs. OCR data) may limit transfer, as discussed in Appendix C.6. We believe applying TokenSwap at a larger training scale is a promising direction for improving multimodal performance while preserving strong text-only capabilities and reducing the modality gap. As a data-centric approach, TokenSwap can be seamlessly integrated into existing training pipelines and extended to diverse datasets at scale, which we leave for future work.

Acknowledgements and Disclosures

Funding. This work was supported in part by research funding provided to Yao Qin at the University of California, Santa Barbara, and by Google DeepMind.

Competing Interests. Several authors are affiliated with Google DeepMind. The authors declare no other competing interests.

References

[1] Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez De Lacalle. Vision-language models struggle to align entities across modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18846–18862, 2025.

[2] Anthropic. Introducing Claude Haiku 4.5. `https://www.anthropic.com/news/claude-haiku-4-5`, 2025. Accessed: 2026-05-06.

[3] Anthropic. Introducing Claude Opus 4.6. `https://www.anthropic.com/news/claude-opus-4-6`, 2026. Accessed: 2026-05-06.

[4] Anthropic. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, 2026. Accessed: 2026-05-06.

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

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv e-prints*, art. arXiv:2502.13923, February 2025. doi: 10.48550/arXiv.2502.13923.

[7] Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, et al. Omnixr: Evaluating omni-modality language models on reasoning across modalities. *arXiv preprint arXiv:2410.12219*, 2024.

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

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[12] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024.

[13] Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

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

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

[16] Google. Gemini API Models. https://ai.google.dev/gemini-api/docs/models, 2026. Accessed: 2026-05-06.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[18] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[22] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[25] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012.

[26] OpenAI. GPT-4o System Card. `https://openai.com/index/gpt-4o-system-card/`, 2024. Accessed: 2026-05-06.

[27] OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, 2025. Accessed: 2026-05-06.

[28] OpenAI. Introducing GPT-5. `https://openai.com/index/introducing-gpt-5/`, 2025. Accessed: 2026-05-06.

[29] OpenAI. Introducing GPT-5.1 for developers. `https://openai.com/index/gpt-5-1-for-developers/`, 2025. Accessed: 2026-05-06.

[30] OpenAI. Introducing GPT-5.2. `https://openai.com/index/introducing-gpt-5-2/`, 2025. Accessed: 2026-05-06.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[32] Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. *arXiv preprint arXiv:2404.07983*, 2024.

[33] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[34] PY Shi, M Welle, M Bjørkman, and D Kragic. Understanding the modality gap in clip. *ICLR, Stockholm, Sweden*, pages 8–10, 2023.

[35] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[36] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[37] Zhenwei Tang, Difan Jiao, Blair Yang, and Ashton Anderson. Seam: Semantically equivalent across modalities benchmark for vision-language models. *arXiv preprint arXiv:2508.18179*, 2025.

[38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. arxiv 2023. *arXiv preprint arXiv:2312.11805*, 2024.

[39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[40] Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M Asano. Same content, different answers: Cross-modal inconsistency in mllms. *arXiv preprint arXiv:2512.08923*, 2025.

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

[42] Xingrui Wang, Jiang Liu, Chao Huang, Xiaodong Yu, Ze Wang, Ximeng Sun, Jialian Wu, Alan Yuille, Emad Barsoum, and Zicheng Liu. Xmodbench: Benchmarking cross-modal capabilities and consistency in omni-language models. *arXiv preprint arXiv:2510.15148*, 2025.

[43] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024.

[44] Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*, 2024.

[45] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

[46] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: a more robust multi-discipline multimodal understanding benchmark (2025). *URL https://arxiv.org/abs/2409.02813*, 2025.

[47] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[48] Xiang Zhang, Senyu Li, Zijun Wu, and Ning Shi. Lost in translation: When gpt-4v (ision) can't see eye to eye with text. a vision-language-consistency analysis of vllms and beyond. *arXiv preprint arXiv:2310.12520*, 2023.

[49] Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2411.09273*, 2024.

[50] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Details for TOKENSWAP-BENCH Construction

A.1 Detailed Token-Level Formalization of TokenSwap

In Section 3.1, we define the modality gap under semantic equivalence and use TokenSwap to construct image-interleaved inputs. Here, we provide a more explicit token-level formalization of the TokenSwap operation.

Formally, let a task be represented by a sequence of textual tokens $X_{\text{text}} = \{t_1, t_2, \ldots, t_n\}$. We define a concept c as a semantically coherent unit (e.g., a word or a phrase) that corresponds to a contiguous span of tokens $c = \{t_i, \ldots, t_j\}$. We then introduce the **TokenSwap** operation $\mathcal{S}(c) \to I_c$, which replaces the textual concept c with an image I_c that conveys the equivalent semantic meaning.

Since images are encoded as a sequence of visual tokens, replacing a single textual concept with its corresponding image yields the interleaved input

$$X_{\text{interleaved}} = \{t_1, \ldots, t_{i-1}, v_1, \ldots, v_m, t_{j+1}, \ldots, t_n\},$$

where $\{v_1, \ldots, v_m\}$ are the visual tokens corresponding to I_c. While we illustrate the formulation with a single replacement for simplicity, TokenSwap can replace multiple textual concepts within the same input, resulting in multiple visual token sequences interleaved with text tokens. An example with three concept replacements is shown in Figure 2, where the Step 1 panel presents the original pure-text question X_{text}, and the Final Benchmark panel shows the image-interleaved version $X_{\text{interleaved}}$ after TokenSwap.

A.2 Detailed Replacement Filtering

This section provides additional details for the replacement filtering procedures.

Validity Filtering We utilize LLM to judge whether the generated image I_c faithfully represents the intended concept c within the context of the question. Specifically, we adopt a one-at-a-time restoration strategy: for each sample, we restore all other replaced words to the text and highlight only the target concept. The model then evaluates whether the candidate image can replace the highlighted text while preserving the original sentence's meaning and clarity. Only replacements that preserve the original meaning without introducing ambiguity are retained. The full prompt for validity filtering is provided in Appendix A.4.

Importance Filtering To ensure that the selected concepts are truly important for solving the task, rather than details that do not affect the final answer, we perform importance filtering at the sample level. Specifically, for each sample, we construct a reduced input $X_{\text{text}}^{-\mathcal{C}}$ by removing all valid concepts from X_{text}, and re-evaluate the model. We retain only samples satisfying

$$\text{Eval}(X_{\text{text}}) = 1 \quad \text{and} \quad \text{Eval}(X_{\text{text}}^{-\mathcal{C}}) = 0, \tag{5}$$

This ensures that the selected concepts collectively play a critical role in the model's prediction.

Caption-Guided Validation To further ensure the quality of visual substitutions I_c, we adopt a round-trip validation strategy. Given the interleaved input $X_{\text{interleaved}} = \mathcal{S}(X_{\text{text}})$, we first generate a textual description $\hat{c}$ for each image I_c using a model. We then construct a reconstructed text-only input $\tilde{X}_{\text{text}}$ by replacing each visual token span $\{v_1, \ldots, v_m\}$ (corresponding to I_c) in $X_{\text{interleaved}}$ with the tokenized caption $\hat{c}$. The prompts for caption generation and reconstruction are provided in Appendix A.4.

We retain only samples for which

$$\text{Eval}(X_{\text{text}}) = 1 \quad \text{and} \quad \text{Eval}(\tilde{X}_{\text{text}}) = 1, \tag{6}$$

This round-trip process enforces that the semantic information in I_c is recoverable from the image, ensuring that the visual substitution does not introduce ambiguity or information loss. As a result, it significantly mitigates the risk that the measured modality gap is driven by unrecognizable or low-quality visual inputs.

Table 2: Detailed results using GPT-5-mini and Gemini-2.0-Flash as filtering proxies.

Model	Text	Image	GPT Proxy Gap	Gemini Proxy Gap
Gemini-3-Flash	0.983	0.906	0.077	0.042
Claude-Opus-4.6	0.982	0.901	0.081	0.057
GPT-5.1	0.982	0.897	0.085	0.077
Qwen3-VL-4B-Thinking	0.859	0.729	0.130	0.109
GPT-4.1	0.886	0.726	0.160	0.164
Qwen3-VL-4B	0.704	0.509	0.195	0.184
InternVL3-1B	0.491	0.260	0.231	0.218
InternVL3-78B	0.917	0.612	0.305	0.258
LLaVA-OV-7B	0.704	0.258	0.446	0.409

A.3 Robustness to Filtering Proxy

All three filtering stages use Gemini-2.0-Flash as the underlying model. To test whether the choice of proxy biases the cross-model comparison, we rebuild the benchmark with GPT-5-mini as the proxy. We rerun the two stages that depend most on the proxy: importance filtering (Eq. 2, text-correct $\wedge$ concept-removed-wrong) and caption-guided validation (Eq. 3, text-correct $\wedge$ image-caption-correct). This results in a benchmark of 1,390 samples, on which we reevaluate nine models.

Across these nine models, the ranking is unchanged (Spearman $\rho = 1.00$, Kendall $\tau = 1.00$, Pearson $r = 0.99$), showing that the cross-model comparison is largely insensitive to the choice of proxy. Gemini-3-Flash still achieves the smallest modality gap.

A.4 Prompt Templates

Prompt Template for Visualizable Concept Extraction

```
You are an advanced text analysis assistant. I will provide you with a piece of text, and your
task is to identify words or phrases that can be replaced with images without changing the meaning
 of the text.

Requirements:

- Output only a Python list (list[str]), without any extra explanations.
- Select concrete nouns (e.g., "apple," "car") and visually representable phrases (e.g., "cup of
coffee," "blue sky").
- Do not include **verbs**, **adjectives**, **numbers** or abstract ideas unless they are part of
a clearly visual scene.
- Prefer shorter, concise phrases over long, complex expressions.

Input:{text}
```

Prompt Template for Validity Filtering

```
Given a piece of text where one word or phrase is enclosed in double asterisks (**like this**),
and a set of candidate images, determine for each image whether the word/phrase can be replaced
with that image while keeping the text's meaning the same.
For each image, consider:
- Whether the image clearly represents the same concept as the word/phrase.
- Whether replacing the word/phrase with the image would preserve the original sentence's meaning.
- Whether the image introduces any ambiguity or confusion in context.

Output a JSON dictionary where each key is the image ID (e.g., "image1", "image2", etc.). The
value should be a dictionary with two keys:
- `"can_replace"`: a boolean indicating whether the image can replace the word/phrase.
- `"reason"`: a brief explanation (1-2 sentences) justifying the decision based on the criteria
above.

Example output:
{
  "image1": {
    "can_replace": true,
    "reason": "The image clearly depicts a cat, which is exactly what the phrase refers to and
fits the sentence without ambiguity."
  },
  "image2": {
    "can_replace": false,
    "reason": "The image shows a dog, which does not match the intended meaning and would confuse
the reader."
  }
}


Here is the text, the {{image_num}} candidate images, and the target word or phrase enclosed in
double asterisks: **{{word}}**
```

Prompt Template for Caption-Guided Validation

```
Replace each image tag in the question and options below with a short, clear word or phrase.
Then, return only the full question and options with the image tags replaced.

Question: {question}
A. {choice_1}
B. {choice_2}
C. {choice_3}
D. {choice_4}
```

A.5 Human Study Details

Setup. We randomly sample one question from each of 57 subjects in TOKENSWAP-BENCH. For each sampled question, we consider all concepts that are replaced by images, resulting in a total of 213 substituted concepts. For each substituted concept, we construct a multiple-choice question asking *"What concept does this image represent?"* given the substituted image.

Each question contains four options, including one correct answer and three distractors. The distractors are sampled from a subject-specific concept pool constructed from all replaced concepts within the same subject. An example is shown in Figure 7.

Results. The evaluation is conducted by two human annotators (one author and one independent annotator), achieving 92.0% (196/213) and 89.2% (190/213) accuracy, respectively, both substantially above the random baseline of 25%. The inter-annotator agreement is 93.4% (199/213). Errors tend to cluster in domain-specific subjects requiring specialized knowledge, particularly virology and chemistry, as well as economics and management. We note that while some distractors are designed to be plausible alternatives, they are generally not highly confusable (e.g., "doubly linked list" vs. "computers"), which makes the task relatively easier. Therefore, the reported accuracy can be viewed as a conservative measure of semantic alignment. Overall, the results indicate that the substituted images reliably convey the intended concepts.

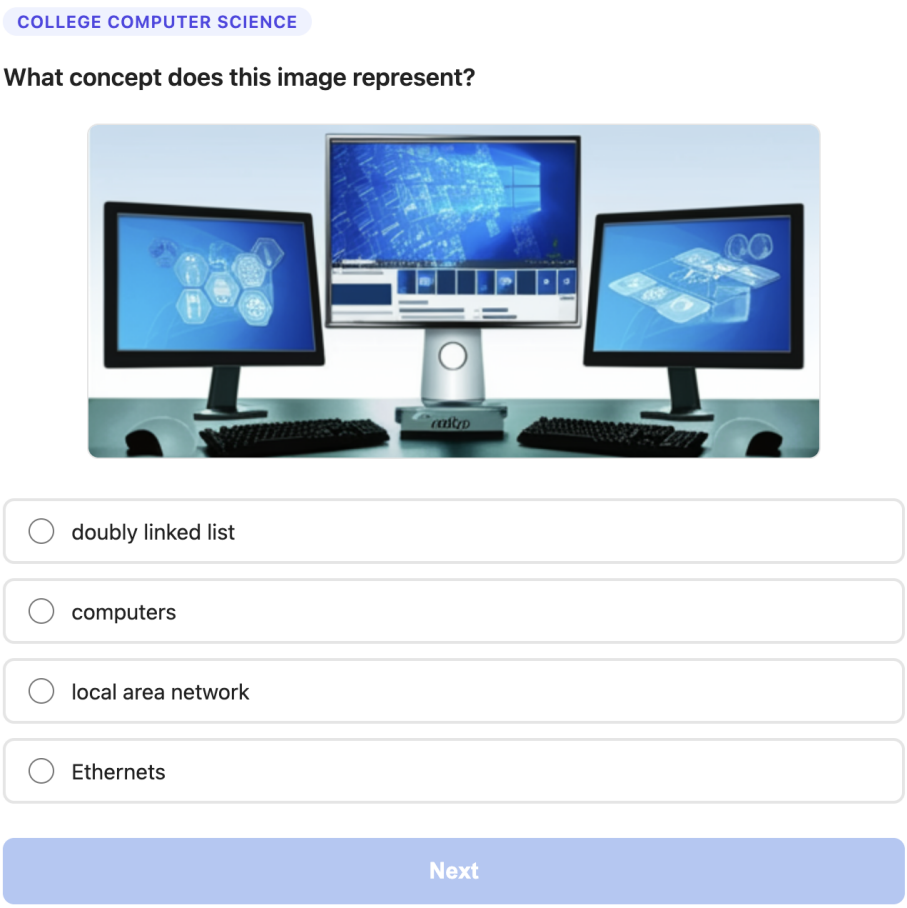


Figure 7: Example of a multiple-choice question used in the human study.

B Details for TOKENSWAP-BENCH Evaluation

B.1 Prompt Templates

Standard Prompt. The standard prompt is shown below. We extract the predicted answer using regular expressions by selecting the first or last occurrence of a capital letter corresponding to a valid option (e.g., A, B, C, or D) in the model output. This accounts for imperfect instruction following, where models may produce outputs beyond the expected format. A prediction is considered correct if either extracted option matches the ground truth.

Standard Prompt Template

```
The following are multiple choice questions about {subject}.
Answer the question by replying A, B, C or D.

Question: {question}
A. {choice_1}
B. {choice_2}
C. {choice_3}
D. {choice_4}

Answer:
```

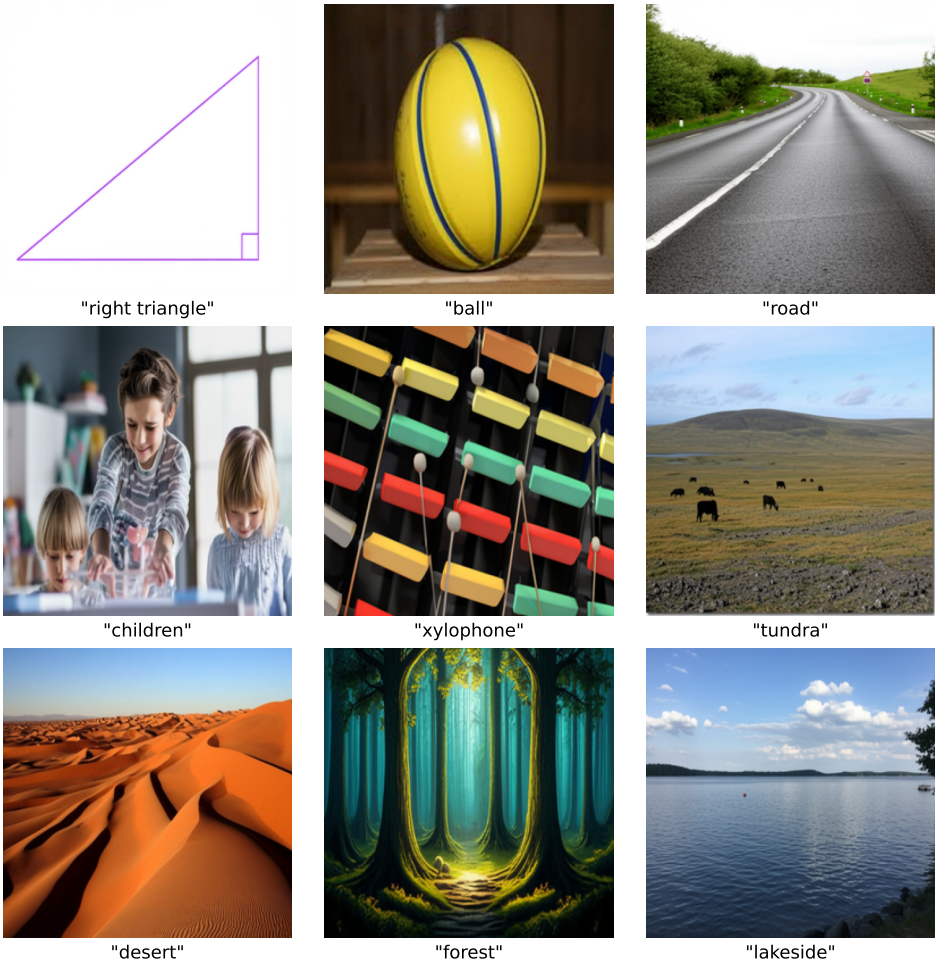


Figure 8: **Images in few-shot demonstrations.** Images are generated by Gemini to represent corresponding textual concepts. In the image-interleaved version, these textual concepts are replaced with the corresponding images.

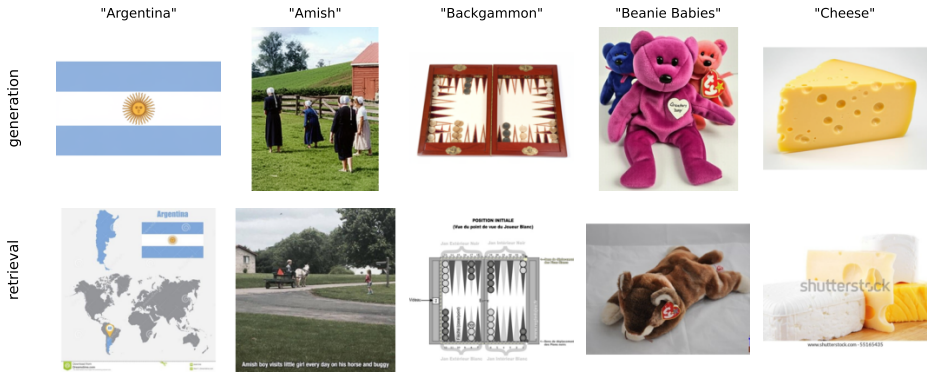


Figure 9: **Examples of retrieved and generated images for the same concepts.** Each column corresponds to a replaced word, with a generated image shown in the top row and a retrieved image shown in the bottom row.

Chain-of-Thought Prompt. The chain-of-thought (CoT) prompt is shown below. We instruct the model to generate step-by-step reasoning before producing the final answer in a specified format. The predicted answer is extracted using regular expressions by matching the pattern "Final Answer: X", where X is one of A, B, C, or D.

Chain-of-Thought Prompt Template

```
The following are multiple choice questions about {subject}.
Solve the problem step by step. After your reasoning, output the final answer in the format:
Final Answer: X
where X is one of A, B, C, or D.

Question: {question}
A. {choice_1}
B. {choice_2}
C. {choice_3}
D. {choice_4}

Let's think step by step.
```

Few-Shot Prompt. The few-shot prompt is shown below. We curate three additional examples that are not included in our benchmark as demonstrations. For text-only inputs, we use pure text examples, while for image-interleaved inputs, we use corresponding image-interleaved versions. Specifically, textual concepts in the demonstrations are replaced with images to match the input format. We show the text-only demonstrations and example visualizations of selected concepts in Figure 8. We extract the predicted answer using the same procedure as in the standard prompt.

Few-Shot Prompt Template

```
The following are multiple choice questions about high school mathematics.
Answer the question by replying A, B, C or D.
Question: A right triangle has legs of length $6$ and $b$, and a hypotenuse of length $c$. The
perimeter of the triangle is $18$. Compute $c$.
A. 18
B. \frac{15}{2}
C. 8
D. 6
Answer: B

The following are multiple choice questions about moral scenarios.
Answer the question by replying A, B, C or D.
Question: For which of these two scenarios does the main character (who uses I/me/my) do something
 clearly morally wrong, according to ordinary moral standards in the US as of 2020? Scenario 1 | I
 slammed on my breaks to miss the ball as it came in to the road. Scenario 2 | I taught my
children to play the xylophone.
A. Wrong, Wrong
B. Wrong, Not wrong
C. Not wrong, Wrong
D. Not wrong, Not wrong
Answer: D

The following are multiple choice questions about prehistory.
Answer the question by replying A, B, C or D.
Question: The early Mesolithic Maglemosian culture was adapted to:
A. a tundra environment.
B. a desert environment.
C. a forest and lakeside environment.
D. an equatorial environment.
Answer: C

The following are multiple choice questions about {subject}.
Answer the question by replying A, B, C or D.

Question: {question}
A. {choice_1}
B. {choice_2}
C. {choice_3}
D. {choice_4}

Answer:
```

Few-Shot Chain-of-Thought Prompt. The few-shot chain-of-thought (CoT) prompt is shown below. We combine few-shot demonstrations with step-by-step reasoning. For text-only inputs, we use pure text demonstrations, while for image-interleaved inputs, we use corresponding image-interleaved versions, where textual concepts are replaced with images to match the input format. We extract the predicted answer using the same procedure as in the CoT prompt.

Few-Shot Chain-of-Thought Prompt

```
The following are multiple choice questions about high school mathematics.
Solve the problem step by step. After your reasoning, output the final answer in the format:
Final Answer: X
where X is one of A, B, C, or D.

Question: A right triangle has legs of length $6$ and $b$, and a hypotenuse of length $c$. The
perimeter of the triangle is $18$. Compute $c$.
A. 18
B. \frac{15}{2}
C. 8
D. 6

Let's think step by step.
To find the length of the hypotenuse, we start by using the perimeter to express one leg in terms
of the other. Given the perimeter is 18 and one leg is 6, the remaining two sides must sum to 12,
so we can say b = 12 - c. By plugging these values into the Pythagorean theorem, we get 36 plus
the square of (12 - c) equals c squared. Expanding that equation gives us 36 + 144 - 24c + c^2 = c
^2. After the c squared terms cancel out, we are left with 24c = 180, which simplifies to c = 7.5,
 or 15/2.
Final Answer: B

The following are multiple choice questions about moral scenarios.
Solve the problem step by step. After your reasoning, output the final answer in the format:
Final Answer: X
where X is one of A, B, C, or D.

Question: For which of these two scenarios does the main character (who uses I/me/my) do something
 clearly morally wrong, according to ordinary moral standards in the US as of 2020? Scenario 1 | I
 slammed on my breaks to miss the ball as it came in to the road. Scenario 2 | I taught my
children to play the xylophone.
A. Wrong, Wrong
B. Wrong, Not wrong
C. Not wrong, Wrong
D. Not wrong, Not wrong

Let's think step by step.
In the first scenario, slamming on the brakes to avoid a ball is a standard safety response
intended to prevent a collision or protect a child who might be nearby, so it isn't morally wrong.
 In the second scenario, teaching children how to play a musical instrument like the xylophone is
a positive educational activity. Because both actions are responsible and beneficial, neither
scenario describes something that would be considered wrong by ordinary moral standards.
Final Answer: D

The following are multiple choice questions about prehistory.
Solve the problem step by step. After your reasoning, output the final answer in the format:
Final Answer: X
where X is one of A, B, C, or D.

Question: The early Mesolithic Maglemosian culture was adapted to:
A. a tundra environment.
B. a desert environment.
C. a forest and lakeside environment.
D. an equatorial environment.

Let's think step by step.
The Maglemosian culture flourished in Northern Europe during a time when the retreating glaciers
were being replaced by thick forests and rising water levels. The name itself refers to a 'big bog
,' reflecting how these people adapted their hunter-gatherer lifestyle to thrive in temperate
woodlands and near inland lakes or rivers. This makes the combination of forest and lakeside the
only accurate representation of their environment.
Final Answer: C

The following are multiple choice questions about high school government and politics.
Solve the problem step by step. After your reasoning, output the final answer in the format:
Final Answer: X
where X is one of A, B, C, or D.

Question: Which of the following does the Supreme Court NOT have the power to override?
A. Constitutional amendments
B. Presidential executive orders
C. Laws passed by Congress
D. Laws passed by state legislatures

Let's think step by step.
```

Table 3: Modality gap on GSM8K compared with TOKENSWAP-BENCH.

Model	Text	Image	GSM8K Gap	MMLU Gap
Gemini-3-Flash	1.000	0.945	0.055	0.042
Claude-Opus-4.6	1.000	0.959	0.041	0.057
GPT-5.1	1.000	0.945	0.055	0.077
Qwen3-VL-4B-Thinking	0.973	0.753	0.220	0.109
GPT-4.1	0.986	0.795	0.191	0.164
Qwen3-VL-4B	0.904	0.575	0.329	0.184
InternVL3-1B	0.411	0.041	0.370	0.218
InternVL3-78B	0.918	0.562	0.356	0.258
LLaVA-OV-7B	0.740	0.014	0.726	0.409

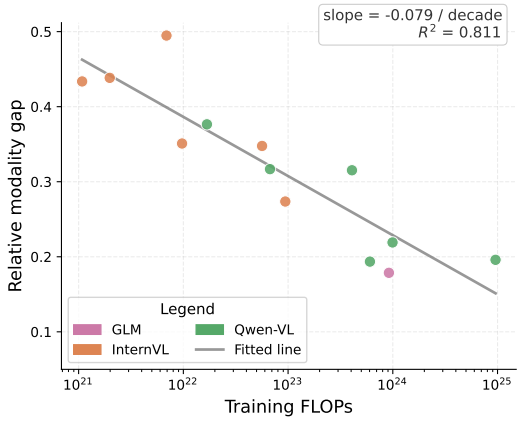


Figure 10: **Strong scaling fit but limited improvement in relative modality gap.** While the linear fit achieves a high $R^2 = 0.811$, indicating a consistent scaling trend, the slope remains small, suggesting limited gains from increasing training compute. Each point represents a model, plotting relative modality gap against estimated training FLOPs.

B.2 Evaluation on GSM8K

To test whether the modality gap generalizes beyond MMLU and multiple-choice evaluation, we apply TokenSwap to GSM8K [10], a reasoning task with free-form answers. We reuse the validated phrase-image pairs from TOKENSWAP-BENCH and swap them into GSM8K test questions. Following the same filtering process and manual verification, we construct 73 questions with an average of 6.2 images per sample (minimum 1, maximum 18).

As shown in Table 3, we observe three consistent findings. (1) The modality gap is universal: every model drops on the image-interleaved version, regardless of whether the questions come from MMLU or GSM8K. (2) The gaps on MMLU and GSM8K are strongly correlated across models (Pearson $r = 0.98$). (3) For strong models such as Claude-Opus-4.6, the two gaps are close, whereas for weaker models such as Qwen3-VL, the gap on GSM8K is substantially larger. Together, these results suggest that the modality gap reflects a general model-level weakness rather than being specific to MMLU or its multiple-choice format.

B.3 Scaling Behavior of Relative Modality Gap

We further analyze the scaling behavior of the relative modality gap with respect to training FLOPs, as shown in Figure 10. Compared to the absolute modality gap, the relative gap exhibits a more consistent scaling trend, with a linear fit achieving a higher $R^2 = 0.811$.

While the negative slope suggests that increasing training compute can reduce the modality gap, the overall improvement remains marginal. In particular, a $10\times$ increase in training FLOPs leads to only

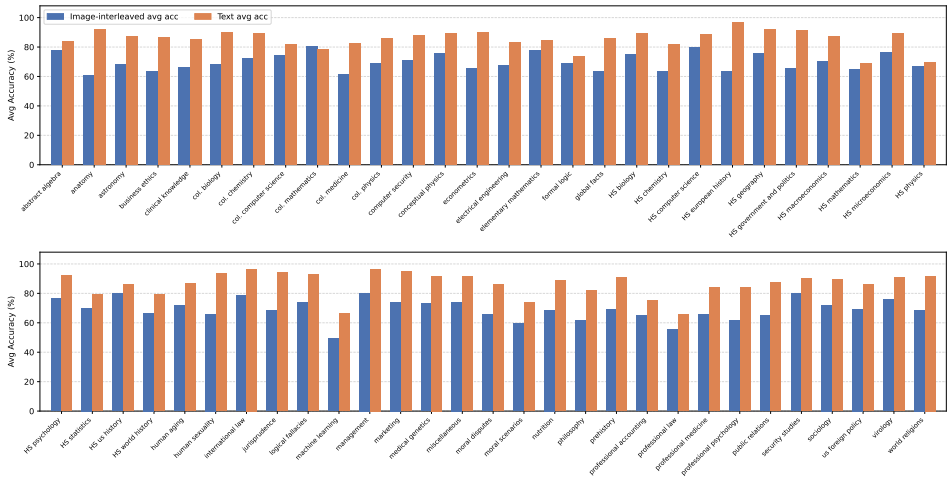


Figure 11: Per-subject accuracy on MMLU comparing text-only and image-interleaved inputs.

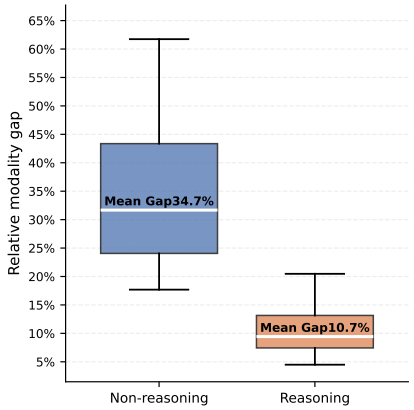


Figure 12: Distribution of relative modality gap for reasoning and non-reasoning models.

a small reduction (7.9%) in the modality gap, indicating that scaling alone is insufficient to address this issue.

B.4 Per-subject Modality Gap

Figure 11 presents per-subject results on TOKENSWAP-BENCH. We observe a consistent modality gap across nearly all subjects, where image-interleaved performance is lower than text-only performance. The magnitude of the gap varies across domains: it is generally larger in knowledge-intensive and abstract subjects (e.g., history, law, and medicine), while relatively smaller in more structured or quantitative domains (e.g., mathematics and physics).

One possible explanation is that, in quantitative domains such as mathematics, the core reasoning primarily depends on numerical values. While these values are less frequently replaced, even when they are presented visually, they are typically easy to interpret. In addition, although contextual concepts (e.g., "books" or "shelf") are important for understanding the problem, they do not directly affect the computation, which usually involves comparing or aggregating numerical values (e.g., identifying the valid range $2.3 \leq w \leq 3.2$ from the weights of the books on the shelf). As a result, models can still perform the required reasoning, leading to a smaller modality gap.

These results suggest that the modality gap is a broad and systematic phenomenon, while its magnitude varies across different subjects.

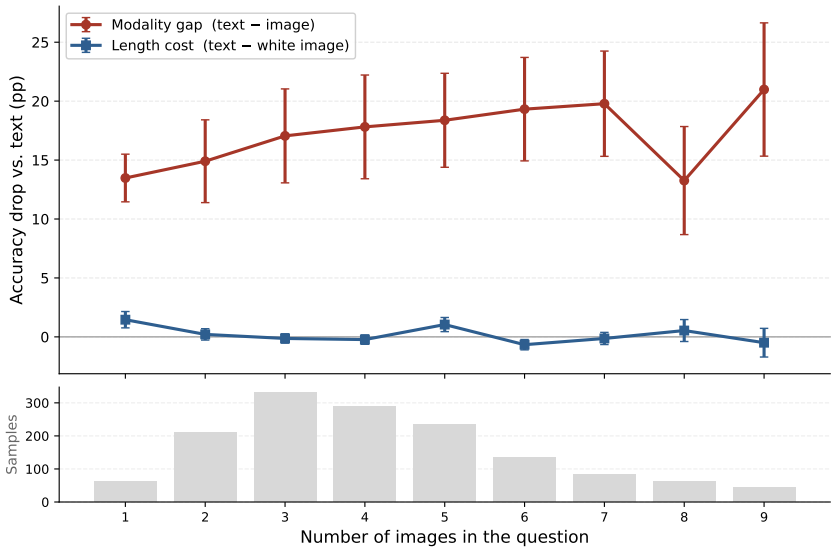


Figure 13: **The modality gap generally increases with the number of image replacements, while the white-image gap remains near zero.** We average results across nine models and group samples by the number of images. The natural-image gap (Text − Image) generally increases as more textual concepts are replaced with images, whereas the white-image gap (Text − White Image) remains close to zero, indicating that the trend cannot be explained by additional visual tokens alone. The bottom panel shows the number of samples in each group.

B.5 Modality Gap Increases with the Number of Image Replacements

According to Figure 13, we find that the modality gap generally increases as more text spans are replaced by images. Averaged across models, the gap rises from 13.5 percentage points with one replacement to around 20 percentage points with more replacements. This suggests that each additional image introduces extra visual grounding difficulty, making the text-to-image performance gap more pronounced.

To determine whether this trend is simply caused by processing more images, we perform the same analysis using the white-image images. In contrast to the natural-image gap, the white-image gap remains approximately flat around zero (-0.7 to $+1.5$ percentage points) as the number of images increases. This suggests that the growing modality gap is not caused by the number of images alone, but by moving more semantic information from text into the visual modality, which is exactly the modality gap we measures.

B.6 Controlling for Visual-Input Confounds

Longer visual inputs. To isolate the effect of the additional visual tokens, we conduct a white-image ablation. For each TOKENSWAP-BENCH sample, we replace every image with a pure-white image of the same size, place all images at the beginning of the question, and restore the original text. This preserves the same number of visual tokens and results in an input length nearly identical to that of the corresponding TOKENSWAP-BENCH sample.

As shown in Table 4, for all 9 models, the white-image accuracy stays within ± 1.0 pp of text-only accuracy, so the additional visual tokens contribute very little to the observed modality gap. These results show that the gap cannot be explained by the additional visual-token count or input length alone.

Rendered-text-image. We also conduct a rendered-text-image baseline. We decompose the total modality gap into two components: the performance drop from text-only to rendered-text-image inputs, $(\text{text} - \text{rendered image})/(\text{text} - \text{real image})$, and the performance drop from rendered-text-image to natural concept-image inputs, $(\text{rendered image} - \text{real image})/(\text{text} - \text{real image})$.

Table 4: Detailed results for the white-image control.

Model	Text	White Image	Image	Text − White Image	Text − Image (Modality Gap)
Gemini-3-Flash	0.925	0.935	0.884	−0.010	0.041
Claude-Opus-4.6	0.984	0.983	0.926	0.001	0.058
GPT-5.1	0.980	0.981	0.903	−0.001	0.077
Qwen3-VL-4B-Thinking	0.889	0.890	0.780	−0.001	0.109
GPT-4.1	0.925	0.923	0.761	0.002	0.164
Qwen3-VL-4B	0.707	0.714	0.523	−0.007	0.184
InternVL3-1B	0.502	0.493	0.284	0.009	0.218
InternVL3-78B	0.943	0.938	0.685	0.005	0.258
LLaVA-OV-7B	0.710	0.702	0.301	0.008	0.409

Table 5: Detailed results for the rendered-text-image control.

Model	Text	Rendered	Image	Text→Rendered	Rendered→Image
Gemini-3-Flash	0.925	0.917	0.884	19.5%	80.5%
Claude-Opus-4.6	0.984	0.976	0.926	13.8%	86.2%
GPT-5.1	0.980	0.976	0.903	5.2%	94.8%
Qwen3-VL-4B-Thinking	0.889	0.876	0.780	11.9%	88.1%
GPT-4.1	0.925	0.906	0.761	11.6%	88.4%
Qwen3-VL-4B	0.707	0.690	0.523	9.2%	90.8%
InternVL3-1B	0.502	0.297	0.284	94.0%	6.0%
InternVL3-78B	0.943	0.873	0.685	27.1%	72.9%
LLaVA-OV-7B	0.710	0.315	0.301	96.6%	3.4%

As shown in Table 5, for seven of the nine models, 72–95% of the gap lies in the second comparison: reading the phrase from a rendered image causes a substantially smaller performance drop of 0.4–7.0 percentage points. For the other two models, InternVL3-1B and LLaVA-OV, the gap comes mostly from the first comparison. However, both are among the weaker models in our evaluation pool, and this strong sensitivity to rendered-text-image is consistent with prior work [40].

B.7 Detailed Numerical Results

We report detailed numerical results for TOKENSWAP-BENCH in Table 6, and present a comparison with SEAM [37] in Table 7.

C Details for TokenSwap Training

C.1 Experimental Setup

Data Construction. We start from a large-scale pure text instruction tuning dataset, Magpie-Pro [43]. We follow the same procedure described in Section 3.2, except that we remove *Concept Importance Filtering* and *Caption-Guided Validation* to enable scalable data generation. Specifically, we apply visual concept substitution only to the human queries while keeping the assistant responses unchanged.

This process yields 116,722 samples, with an average of 1.75 generated images per sample. Based on the constructed post-training data, we further derive a pre-training data variant. The pre-training data is nearly identical to the post-training data, except that the responses are removed. Example samples are provided in Figure 14.

We additionally explore an alternative data construction strategy using image retrieval. Specifically, we construct training data by retrieving images based on the LLaVA pre-training data [23]. Detailed implementation is provided in Appendix C.4.

Table 6: Detailed numerical results for TOKENSWAP-BENCH.

Model	Text	Image	Gap
GPT-4.1	0.925	0.761	+0.164
GPT-4.1-mini	0.892	0.686	+0.206
GPT-4o	0.913	0.693	+0.220
GPT-4o-mini	0.832	0.545	+0.288
GPT-5	0.952	0.856	+0.096
GPT-5.1	0.980	0.903	+0.077
GPT-5.2	0.972	0.887	+0.085
Gemini-2.5-Flash	0.954	0.885	+0.070
Gemini-2.5-Flash-Lite	0.904	0.766	+0.138
Gemini-2.5-Pro	0.968	0.888	+0.080
Gemini-3-Flash	0.925	0.884	+0.042
Gemini-3.1-Flash-Lite	0.954	0.759	+0.195
Gemini-3.1-Pro	0.976	0.914	+0.062
Claude-Haiku-4.5	0.951	0.833	+0.118
Claude-Opus-4.6	0.984	0.926	+0.057
Claude-Sonnet-4.6	0.979	0.906	+0.073
InternVL2-40B	0.856	0.505	+0.352
InternVL2-8B	0.767	0.294	+0.474
InternVL2-1B	0.453	0.184	+0.268
InternVL2.5-78B	0.920	0.600	+0.320
InternVL2.5-8B	0.756	0.382	+0.374
InternVL2.5-1B	0.460	0.259	+0.202
InternVL3-78B	0.943	0.685	+0.258
InternVL3-8B	0.788	0.511	+0.276
InternVL3-1B	0.502	0.284	+0.218
Qwen2-VL-72B	0.900	0.726	+0.174
Qwen2-VL-7B	0.704	0.481	+0.223
Qwen2-VL-2B	0.555	0.346	+0.209
Qwen2.5-VL-72B	0.909	0.731	+0.178
Qwen2.5-VL-7B	0.677	0.529	+0.148
Qwen2.5-VL-3B	0.670	0.458	+0.211
Qwen3-VL-32B	0.889	0.690	+0.199
Qwen3-VL-30B-A3B-Instruct	0.852	0.618	+0.234
Qwen3-VL-8B	0.782	0.571	+0.211
Qwen3-VL-8B-Thinking	0.928	0.808	+0.120
Qwen3-VL-4B	0.707	0.523	+0.184
Qwen3-VL-4B-Thinking	0.889	0.780	+0.109
LLaVA-OV-72B	0.909	0.451	+0.458
LLaVA-OV-7B	0.710	0.301	+0.409
LLaVA-OV-0.5B	0.409	0.247	+0.162
GLM-4.1V-9B-Thinking	0.901	0.740	+0.161
GLM-4.6V-Flash	0.932	0.803	+0.129

Table 7: Comparison of modality gap between SEAM and our evaluation on overlapping models.

Model	SEAM		Ours		Gap	
	Text	Image	Text	Image	SEAM	Ours
GPT-4o	0.635	0.482	0.913	0.693	+0.153	+0.220
GPT-4o-mini	0.555	0.411	0.832	0.545	+0.144	+0.288
InternVL2.5-78B	0.448	0.414	0.920	0.600	+0.034	+0.320
InternVL2.5-8B	0.324	0.337	0.756	0.382	-0.013	+0.374
InternVL3-78B	0.525	0.427	0.943	0.685	+0.098	+0.258
InternVL3-8B	0.382	0.357	0.788	0.511	+0.025	+0.276
Qwen2.5-VL-72B	0.547	0.475	0.909	0.731	+0.072	+0.178
Qwen2.5-VL-7B	0.303	0.350	0.677	0.529	-0.047	+0.148

Experimental Setting. We compare two training paradigms: Text training and TokenSwap training. In text training, inputs remain purely textual, while TokenSwap training converts the same samples into image-interleaved inputs by replacing selected textual concepts with semantically aligned images. The two paradigms are constructed as *paired settings*: both variants share the exact same samples and differ only in whether visual concepts are replaced, allowing us to evaluate whether TokenSwap training preserves text-only performance.

We use Qwen2-VL-7B as the base model[2] for all experiments and adopt the LLaVA post-training dataset[3] as the base post-training data. We choose Qwen2-VL-7B because it is one of the strongest models in our benchmark with publicly available pretrained weights, which are required for our controlled post-training experiments. Since Qwen does not release its post-training data, we use the LLaVA post-training dataset as a substitute. We study these paradigms across different training stages and image sources.

Baseline. We train the model using only the LLaVA post-training dataset.

Continuous Pre-training. We continue pre-training the base model using either pure text data or TokenSwap data. For TokenSwap training, we consider two variants based on the image source: generated images and retrieved images. All models are subsequently post-trained on the same LLaVA dataset as in the baseline.

Post-Training. For post-training, we directly perform visual instruction tuning on a mixture of baseline LLaVA data and our constructed data. Similar to pre-training, we consider both generated images and retrieved images.

Training Details. For post-training, we train the LLM and multimodal projector while keeping the vision encoder frozen. We train for one epoch with a learning rate of 2×10^{-5} using a cosine learning-rate schedule and a warmup ratio of 0.03. We use a weight decay of 0, a maximum gradient norm of 1.0, and a per-device batch size of 4 with 4 gradient accumulation steps across 8 GPUs, resulting in an effective batch size of 128. Training uses bfloat16 precision, a maximum sequence length of 8192, and image resolutions ranging from 784 to 50,176 pixels. We use DeepSpeed ZeRO-3 with gradient checkpointing, and all experiments are conducted on 8 A100 GPUs.

C.2 Generalization to a Second Training Backbone

We replicate the post-training experiment on InternVL3-8B-Pretrained with the same training data (Gen variant) but at a smaller scale. Specifically, we use a random 12% subsample of each training set.

As shown in Table 8, TokenSwap training reduces the gap from 0.303 to 0.230, and the improvement comes largely from the image-interleaved side (0.438 $\rightarrow$ 0.518), while text-only accuracy is almost unchanged (0.741 $\rightarrow$ 0.748). This reproduces our Qwen2-VL-7B finding on a second model family, showing that the mitigation is not specific to one backbone.

[2] https://huggingface.co/Qwen/Qwen2-VL-7B

[3] https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K

Table 8: TokenSwap training results on InternVL3-8B-Pretrained.

Training Data	Text	Image	Gap
Baseline	0.741	0.438	0.303
Text Training	0.753	0.429	0.324
TokenSwap Training	0.748	0.518	0.230

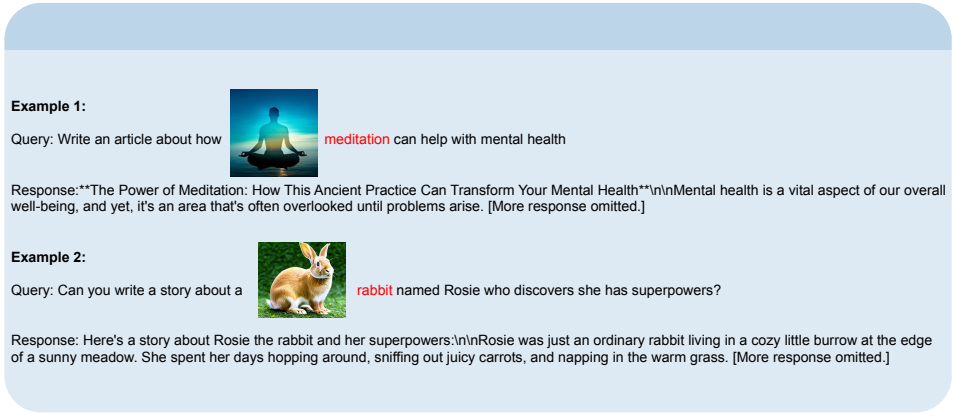


Figure 14: **Applying TokenSwap to the Magpie-Pro dataset.** We construct image-interleaved instructions by replacing key textual concepts in the original queries with semantically aligned images. The red text highlights the concepts that are replaced by images.

C.3 Experimental Setup for OCR Task

Data Construction. The IIIT 5K-Word dataset consists of image–text pairs, where each image contains a word and the corresponding text provides its transcription, with a standard split of 2,000 training and 3,000 test samples. Starting from the ground-truth text, we first use an LLM to construct natural language descriptions. We then apply TokenSwap by replacing the target word in the text with its corresponding image, resulting in image-interleaved samples.

To formulate the task as a conversational instruction, we construct user prompts that query the word shown in the image. For example, given an image–text pair where the image corresponds to the word "RESCUE", we create the prompt: "What is the word before [mission] in the following sentence: They planned a quick <image> mission at dawn." Here, <image> represents the actual image of the target word, and the ground-truth answer is "RESCUE". We present more examples in Figure 16.

Experimental Setting. We follow the same experimental setup as in Section C.1, using Qwen2-VL-7B as the base model. We augment the LLaVA post-training data (Baseline) with either our TokenSwap data (TokenSwap Training) or pure-text data (Text Training), and evaluate on the IIIT 5K-Word test split.

C.4 Details for Image Retrieval in TokenSwap Training

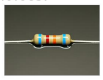
Unlike retrieving images from a large but noisy dataset such as DataComp-Small [15], as used in constructing TOKENSWAP-BENCH, we instead start from the curated image-text pairs used in LLaVA pre-training[4] to ensure higher data quality. For each image, we first perform a caption expansion step to enrich the textual description at multiple levels of granularity. Specifically, given an original caption (e.g., *"the intel core processor"*), we expand it into a set of semantically equivalent captions ranging from coarse to fine-grained descriptions, such as *"a computer processor"* and *"a CPU chip"*.

Next, we use the expanded captions to retrieve candidate matches from the Magpie-Pro dataset [43] by performing exact string matching. To further ensure the quality of the retrieved pairs, we apply

[4] https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K/blob/main/llava_v1_5_mix665k.json

Question: A participant in a single-trial free-recall task is presented with a list of words, one at a time, in the following order:  (house),  (flower),  (dog), table, license,  (water),  (computer),  (salad). In accord with the serial position curve, which of the following words is the participant most likely to forget?
Choices:
A.  (House)
B.  (Computer)
C.  (Flower)
D. License
Answer: D. License

Question: Which of the following statements about sidereal and solar days is not true?
Choices:
A. The time it takes for the  (Sun) to make one circuit of our sky is one solar day.
B. A solar day is 4 minutes longer than a sidereal day.
C. A solar day represents more than 360° of rotation for the earth.
D. The time it takes for the  (Moon) to make one circuit of our sky is one solar day.
Answer: D. The time it takes for the  (Moon) to make one circuit of our sky is one solar day.

Question: Which of the following are the passive elements?
Choices:
A.  (Resistor)
B.  (Bulb)
C. Both
D. None of these.
Answer: C. Both

Question: Which of the following foods do not contain free sugars
Choices:
A.  (Cows' milk)
B. Natural unsweetened  (orange juice)
C.  (Honey)
D.  (Agave nectar)
Answer: A.  (Cows' milk)

Figure 15: Examples in TOKENSWAP-BENCH

a filtering stage using `Qwen3-8B-Instruct` as a judge model. The model takes as input the full sentence together with the corresponding image, and is prompted to determine whether replacing the textual concept with the image preserves the original semantics. Only samples that are judged as valid replacements are retained for TokenSwap training. In total, this process yields 18,366 samples with 29,358 images. The prompt used for filtering is shown below.

Standard Prompt Template

```
### Task
You are a strict data auditor. Your goal is to determine if an <image> can replace a word/phrase
in the sentence UNAMBIGUOUSLY. For large-scale training, we require zero semantic drift. If you
have any doubt, return false.

### Mandatory Decision Logic

1. **Part-of-Speech (POS) Match**:
   - The image MUST represent the word in its specific grammatical role in THIS sentence.
   - **RETURN FALSE IF**: The word is a verb but the image shows a noun (e.g., 'die' as in death
vs. 'die' as a dice).

2. **Semantic Sense Alignment (Primary Defense)**:
   - Words often have multiple meanings. The image MUST match the exact sense used in the context.
   - **RETURN FALSE IF**:
     - Text means "human head" but image shows a "guitar amplifier head".
     - Text means "human nature" but image shows "forest/greenery".
     - Text means "tea time/meal" but image shows "tea leaves/packaging".

3. **Environmental & Adjective Consistency**:
   - The image must not contradict the sentence's setting or descriptors.
   - **RETURN FALSE IF**: Text describes a "hot desert" but image shows "wet sea sand"; text says
"wrinkled" but image is "smooth".

4. **No Symbols, Icons, or Text**:
   - **RETURN FALSE IF**: The image is an icon, a diagram, a branded product, or contains
significant printed text (like a book cover or a screenshot of words).

5. **Entity Specificity**:
   - **RETURN FALSE IF**: The image is a specific artifact (a toy, a statue, a specific brand)
when the text refers to a generic living being or object.

---

### Calibration Examples:
- "The hot <image> (sand) scorched his feet." | Image: Ocean waves on sand. | Result: {{"i": 0, "
word": "sand", "ok": false}}
- "The sun beat on his <image> (head)." | Image: A guitar amplifier. | Result: {{"i": 0, "word": "
head", "ok": false}}
- "He sank down... again to <image> (die)." | Image: A gambling dice. | Result: {{"i": 0, "word":
"die", "ok": false}}
- "But won't it be rather late after <image>?" (Word: tea) | Image: A branded tea box | Result:
{{"i": 0, "word": "tea", "ok": false}}

---

### Data to Process:
Sentence: {text}
Replaced words: {word_list}

### Response (JSON ONLY):
{{
  "results": [
    {{
      "i": <int>,
      "word": "<string>",
      "ok": <boolean>
    }}
  ]
}}
```

C.5 Evaluation on Standard Vision-Language Benchmarks

We evaluate models trained with TokenSwap on standard vision-language benchmarks to assess whether the improved cross-modal consistency affects general multimodal performance.

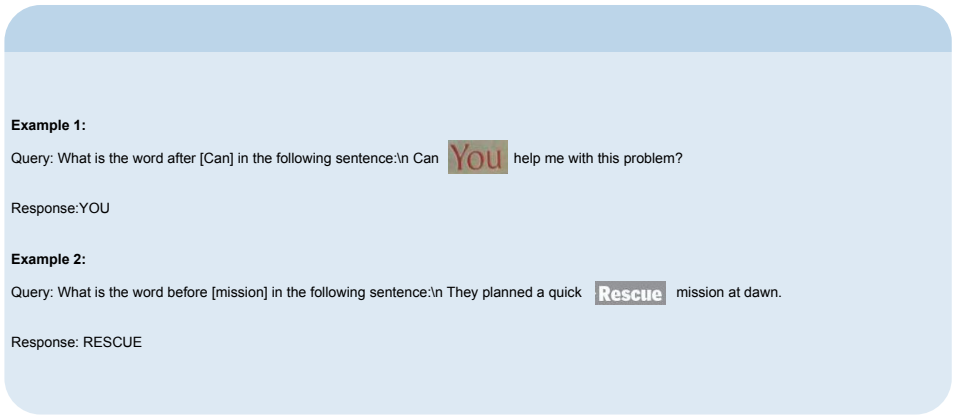


Figure 16: **Applying TokenSwap to the IIIT 5K-Word dataset.** We transform standard image–text pairs into image-interleaved instructions by replacing target words with images. The queries are generated from the ground-truth text using an LLM, where the OCR word is substituted with its image representation.

As shown in Table 9, TokenSwap achieves comparable performance to the baseline across vision-language benchmarks, with no noticeable degradation.

Table 9: Evaluation on standard vision-language benchmarks (GQA, TextVQA, and MME-Peception). TokenSwap training achieves comparable performance to the baseline across all settings, indicating no noticeable degradation.

	GQA [19]	TextVQA [35]	MME [14]
Baseline	63.8	62.5	1542.4
pre-train + Generate	63.5	61.7	1535.9
pre-train + Retrieval	63.7	61.9	1587.0
post-train + Generate	63.9	62.3	1560.0
post-train + Retrieval	63.8	62.0	1544.1

C.6 Effect of Domain Mismatch on TokenSwap Training

We study the impact of domain mismatch between TokenSwap training and evaluation across natural-image and OCR settings. For evaluation on the natural-image domain, we use image-interleaved performance on TOKENSWAP-BENCH.

As shown in Table 10, TokenSwap yields consistent improvements when training and evaluation domains are aligned. However, under domain mismatch, these gains are reduced or disappear, indicating limited cross-domain transferability.

D Detailed Related Work

Modality Gap in Vision–Language Models. The notion of *modality gap* was originally studied in contrastive vision–language models [20, 9, 36, 47, 39], particularly CLIP [31]. Prior work shows that, even with a shared embedding space, image and text representations may remain separated, leading to a representation-level discrepancy between modalities [22, 34]. Subsequent studies further analyze this phenomenon from multiple perspectives, linking it to factors such as object bias and information imbalance [32], exploring how it can be explained or mitigated [12, 44], and even arguing that it arises from contrastive learning itself rather than modality differences [13]. However, such representation-level definitions do not naturally extend to multimodal large language models (MLLMs), which are trained via next-token prediction rather than explicit cross-modal alignment. In this work, we instead study the modality gap from a generative perspective, focusing on discrepancies in model predictions under semantically equivalent inputs.

Table 10: **TokenSwap training is sensitive to domain alignment**. We consider two image domains: natural images and OCR. We use image-interleaved performance on TOKENSWAP-BENCH to represent the natural-image domain, as it involves natural images rather than OCR-style text images. TokenSwap improves performance only when the training and evaluation domains are aligned.

Training Domain	Natural Image-Interleaved Accuracy	OCR Accuracy
No Train	0.4466	0.9003
Natural Image	0.5132	0.9196
OCR Image	0.4393	0.9270

Cross-Modal Consistency in MLLMs. A closely related line of work studies whether multimodal large language models behave consistently when the same semantic content is presented in different modalities [7, 46, 1]. Early work studies cross-modal consistency by converting text into images via text rendering and evaluating MLLMs on semantically equivalent text–image pairs, revealing noticeable inconsistencies in model predictions [48, 49]. van Sprang et al. [40] further show that visual factors such as resolution, font, and color can affect cross-modal consistency, with models often achieving higher accuracy under certain color conditions. Besides rendered-text images, Tang et al. [37] construct semantically equivalent comparisons in controlled domains with standardized textual and visual notations (e.g., chess, chemistry, music, and graphs), and observes notable cross-modal inconsistencies. XModBench [42] further extends cross-modal consistency evaluation to tri-modal settings involving text, vision, and audio.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions: constructing TokenSwap-Bench through concept-level visual substitution, measuring the modality gap by evaluating a broad set of MLLMs, and studying TokenSwap training as a mitigation strategy.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 3.3 Semantic Equivalence and Human Validation and Section 6 Conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors

should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: No theoretical results are included.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the details in Appendix A, B, and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the benchmark data, evaluation code, and instructions in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the details in Appendix A, B, and C.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports confidence intervals for aggregate modality-gap estimates across models and includes statistical analysis where appropriate, such as comparisons between reasoning and non-reasoning models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix C: Details for TokenSwap Training.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The work is designed as a benchmark and training study for evaluating and improving the modality gap in MLLMs. We do not collect sensitive personal information, and we follow responsible research practices for dataset construction, model evaluation, and release.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive impacts such as improving the reliability and consistency of multimodal systems, especially when text and visual inputs convey equivalent semantics. It also discusses potential negative impacts, such as biases from image sources.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper does not release a new model with a high risk for misuse. The released benchmark is derived from MMLU and is intended for evaluating cross-modal consistency; we further apply filtering and validation steps to ensure that image substitutions are semantically appropriate.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites and credits existing datasets, models, and tools used in the study.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces TokenSwap-Bench and TokenSwap training data construction procedures, with details included in Appendix A and C.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Justification: See Appendix A.4 and Figure 7.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The human validation study only asks annotators to identify visual concepts from non-sensitive benchmark images and does not collect private or sensitive participant information. Therefore, it does not constitute human-subjects research requiring IRB review under our understanding.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are used for constructing our benchmark and model evaluation. The paper explicitly describes these uses in Appendix A, B, and C.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.