The Telephone Game: Evaluating Semantic Drift in Unified Models

Sabbir Mollah[⋆], Rohit Gupta[†], Sirnam Swetha[†], Qingyang Liu[‡], Ahnaf Munir[‡], and Mubarak Shah

Institute of Artificial Intelligence, University of Central Florida, USA
{sabbir.mollah, rohit.gupta, Swetha.Sirnam, qingyang.liu2, ahnaf.munir}@ucf.edu, shah@crcv.ucf.edu

[†] Equal contribution as second authors. [‡] Equal contribution as third authors.

Abstract. Unified models (UMs) combine visual understanding (`I2T`) and generation (`T2I`) in a single framework, alongside broader unimodal capabilities. We focus on the cross-modal pair `T2I` and `I2T`, where cross-consistency—what a model understands, it should be able to generate—is a core promise of unification and a practical necessity in applications that compose both capabilities. Yet, existing benchmarks evaluate them in isolation: FID/GenEval for `T2I`; MME/MMBench for `I2T`. We show that this gap is consequential: models that score competitively on these benchmarks can fail severely when their own understanding and generation are composed, progressively losing core entities, attributes, spatial relations, and counts, resulting in *semantic drift*. To quantify drift, we introduce the Semantic Drift Protocol (`SDP`), inspired by the Telephone Game: starting from a caption or image, we iteratively alternate `I2T` and `T2I` over multiple generations and measure how faithfully semantics are preserved. We propose two complementary metrics: Mean Cumulative Drift (`MCD`), an embedding-based measure of overall content retention across three representation spaces, and Multi-Generation GenEval (`MGG`), which extends GenEval's object-level compliance scoring across generations. To stress-test models beyond COCO-style data, we create a benchmark of 400 image-text pairs sampled from NoCaps and DOCCI that emphasizes novel objects and fine-grained descriptions. Applying `SDP` to seven recent unified models reveals that drift behavior varies dramatically and is *not predicted* by single-pass scores: BAGEL retains high semantic fidelity over multiple generations, while VILA-U and Janus variants collapse within five generations, despite comparable isolated metrics. We identify six recurring failure modes and find that degradation is typically catastrophic rather than gradual: once a critical error occurs, subsequent generations compound it. Overall, `SDP` exposes failure modes that single-pass benchmarks miss, enabling a more faithful assessment of unified model reliability. Code and benchmark resources are available at `https://github.com/mollahsabbir/telephone-game-semantic-drift`.

Keywords: Unified Models · Cyclic Evaluation · Semantic Drift

⋆ Corresponding author: Sabbir Mollah.

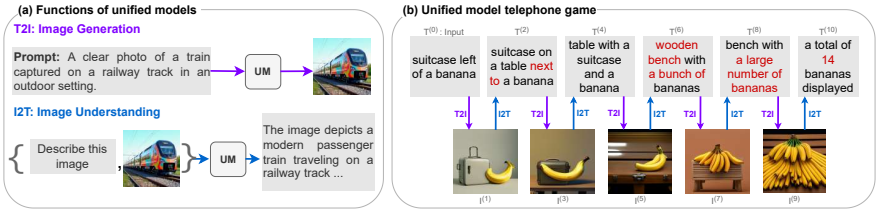


Fig. 1: (a) **Unified models** support both `T2I` and `I2T` tasks. (b) **Telephone Game**: Starting from a textual prompt $T^{(0)}$ describing a suitcase and a banana, a unified model iteratively alternates between `T2I` and `I2T` steps. Semantic drift compounds across rounds: the suitcase degrades and disappears entirely after $I^{(5)}$, while the single banana duplicates and proliferates, reaching 14 by $T^{(10)}$. Red text indicates content that has drifted from the original prompt.

1 Introduction

Multimodal Unified Models (UMs) combine visual understanding and generation within one framework, enabling unimodal tasks (e.g., text-to-text, image-to-image) as well as cross-modal tasks (e.g., image-to-text, text-to-image). By sharing representations across modalities, UMs can demonstrate interesting emerging capabilities such as intelligent photo editing, e.g. BAGEL [8]. Despite rapid progress, UM evaluation mostly remains fragmented. We study two properties of unified models: semantic drift and cross-consistency. Semantic drift refers to the progressive degradation of core scene content (entities, attributes, relations, and counts) relative to the original input as a model repeatedly composes its own `I2T` and `T2I` steps. Cross-consistency measures whether information inferred in `I2T` can also be faithfully realized in `T2I` (and vice versa). Crucially, these properties only become visible under composition: small errors in either direction can be harmless in a single pass yet amplify over repeated `I2T` $\leftrightarrow$`T2I` alternations.

There are several ways to evaluate a model's image generation capabilities. For example, ClipScore [13] uses CLIP to measure semantic alignment of the prompt with generated images. However, its reliance on embeddings may not be reflective of human perceptions [10]. Fréchet Inception Distance (FID) [14] measures the distributional similarity against a reference dataset, but ignores the generated image's faithfulness to the input prompt. A model that ignores the input text and produces high-quality, yet off-prompt images can still score well [10]. GenEval [10] improves prompt compliance via object- and relation-level checks using detectors, but it does not assess realism and, like FID, remains a single-pass measure. A similar limitation holds for image-understanding benchmarks such as MME, MMBench, POPE, and VQA [2, 9, 17, 19], which evaluate `I2T` in isolation without testing unified model's alignment with `T2I`. These single-pass metrics do not assess whether entities, attributes, relations, and counts are retained under repeated `I2T` $\leftrightarrow$`T2I` composition. We therefore focus on the coupled `I2T`/`T2I` setting, where semantic drift is most consequential in composed applications.

Much like the popular children's game *Telephone Game*, where a whispered message drifts in meaning as it passes from person to person, UMs tend to lose or distort semantic meaning when cycling between text and image representations, as shown in Fig. 1(b). Starting from a textual prompt, "a suitcase left of a banana", the model produces an image $I^{(1)}$ correctly, which is then captioned (`I2T`) to form the next prompt $T^{(2)}$, and so on. Although each individual step can look plausible in isolation, semantic drift accumulates across cycles: by generation 5, the image has changed drastically. Notably, a model may score well on isolated single-pass `I2T` or `T2I` metrics while still exhibiting such cross-modal inconsistencies, which current metrics fail to capture. Cross-consistency is illustrated in Fig. 2: even state-of-the-art models like BAGEL [8] can correctly reason about a chessboard in `I2T` (identifying that "the white side wins"), yet fail to generate a faithful `T2I` image of the same winning scenario.

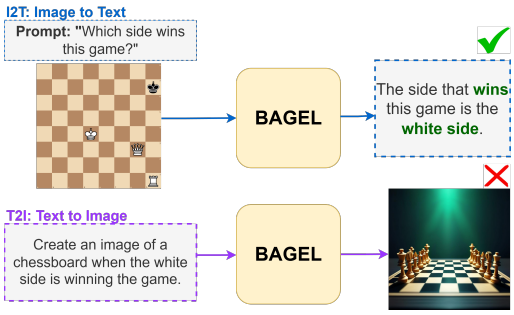


Fig. 2: An example of cross-inconsistency in the BAGEL unified model. **Top:** In `I2T` mode, BAGEL correctly analyzes a chess position and identifies white as the winning side (✓). **Bottom:** In `T2I` mode, when prompted to generate a chessboard where white is winning, the model produces a generic board that does not depict a winning position (✗). This exposes a fundamental inconsistency: a unified model's understanding and generation capabilities are not necessarily aligned, even for the same semantic concept.

These failures suggest that a unified model should maintain a balanced capability in both visual understanding (`I2T`) and visual generation (`T2I`), such that semantics inferred in one direction remain realizable in the other. Our evaluation therefore avoids committing to any single downstream task and instead probes this coupled behavior directly: we alternately compose `I2T` and `T2I` without user intervention, so that small understanding–generation mismatches accumulate into measurable semantic drift. This coupling is implicitly assumed in multi-step workflows that alternate between describing visual states and synthesizing new ones; `SDP` isolates whether divergence arises from the model's own loop rather than from prompting choices. To address this gap, we evaluate unified models for (i) single-step cross-consistency and (ii) multi-step semantic drift under repeated `I2T` $\leftrightarrow$`T2I` composition. We introduce the Semantic Drift Protocol (`SDP`), which generates alternating chains from either text or image and measures how faithfully semantics are preserved as the chain grows (Sec. 3).

Our experiments reveal substantial variation in semantic drift behavior across models. For example, BAGEL [8] maintains strong semantic fidelity across multiple generation cycles, whereas models like VILA-U [41] and Janus [40] degrade rapidly, exposing weaker coupling between their visual understanding and visual

generation capabilities despite competitive single-pass metrics. These findings underscore the need to move beyond isolated `I2T` or `T2I` metrics and toward evaluations that directly measure cross-consistency.

Our contributions are summarized as follows:

- We formalize the cross-consistency and semantic drift problem, showing that single-pass metrics cannot expose gaps between a model's understanding and generation capabilities.
- We propose the Semantic Drift Protocol (`SDP`), which jointly evaluates `I2T` and `T2I` over multiple transitions to track semantic preservation and exposes recurring semantic drift failure modes that emerge during cyclic inference.
- We extend GenEval [10] to a multi-generation setting, which amplifies observable performance differences between models.
- We conduct a human study to determine cross-consistency in existing models and provide a comparative ranking.

2 Related Works

Text-to-image (`T2I`) generation has advanced rapidly with diffusion models such as DALL·E 2 [26], Imagen [29], and Stable Diffusion [28], while visual understanding and captioning progressed from CNN-RNN pipelines [31] to transformer-based architectures trained at web scale [7, 11, 12, 18, 32].

Motivated by applications that compose understanding and generation, recent work explores *unified models* that support both `I2T` and `T2I` within a single system. Early efforts such as Chameleon [36] studied autoregressive mixed-modal generation, while later designs vary in objectives and coupling: Transfusion [45] combines autoregressive and diffusion losses; Show-o [42] uses next-token prediction for text and masked-token prediction [4] for images; VILA-U [41] adopts separate text/vision decoders; and Janus [40]/Janus-Pro [6] decouple visual encoders for understanding and generation. Other approaches relax full sharing, e.g., BLIP3-o [5] adds a dedicated diffusion head, and BAGEL [8] reports strong unified behavior from large-scale interleaved pretraining.

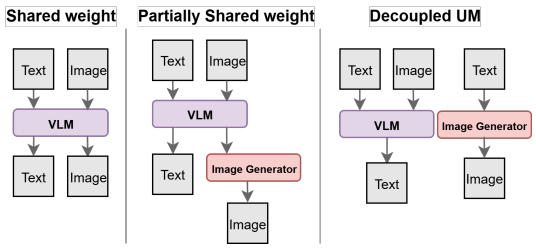


Fig. 3: (Left) A single model handles both understanding and generation. (Middle) The architecture partially shares weights, with a decoder capable of generating text and visual features, the latter is passed to another image generation model. (Right) Understanding and generation processes are fully decoupled, using separate models for each task.

Because these systems differ in *how tightly* `I2T` and `T2I` are coupled, we categorize them by parameter sharing (Fig. 3): **(i) shared-weights** unified

models that use a single backbone for both modalities (e.g., BAGEL [8], Show-o [42], VILA-U [41], Janus [40], Janus-Pro [6]); **(ii) partially-shared** models that retain shared representations but delegate generation to specialized modules (e.g., BLIP3-o [5]); and **(iii) decoupled UM** that emulate unified behavior by composing independently trained models, e.g., pairing LLaVA [18] for `I2T` with SDXL [25] for `T2I`.

With the rapid development of MLLMs [15,16,18,32,35], many benchmarks target understanding: MME [9], MMBench [19], POPE [17], VQA [2] and TLQA [34]; broader reasoning suites include MMMU [44], MM-VET [43], MathVista [20], VRR-QA [33], and robustness or alignment-focused works such as MMVP [37], HumaniBench [27], BBQ-V [23], ALM-Bench [38], and VISAGE [22]. For `T2I`, common metrics include Inception Score [30], FID [14], and GenEval [10]. However, most benchmarks evaluate `I2T` and `T2I` largely in isolation and in a single pass; iterative text–image loops are rarely studied systematically. The work in [3] is closest in spirit, using cycle-consistency to construct preference datasets, but it considers only a single generation step and does not specifically analyze unified models. In contrast, our work evaluates how semantics evolve under repeated composition of `I2T` and `T2I` operations. Uni-MMMU [46] takes an important step toward *coupled* evaluation by designing bidirectionally linked understanding–generation tasks, but it does not measure how faithfully core scene semantics are preserved under *repeated* `I2T` $\leftrightarrow$`T2I` self-composition, leaving semantic drift in multi-round unified-model feedback loops largely unquantified.

3 Semantic Drift Evaluation

We propose a cyclic evaluation protocol, `SDP`, with two metrics for measuring how well a unified model preserves semantic fidelity when alternating between `I2T` and `T2I`. `SDP` proposes to evaluate on multi-generation cycles to provide quantitative measures of semantic drift. In this setting, we treat the $\mathcal{UM}$ as a model composed of at least two functionalities. **Image Generation**: $\mathcal{UM}_{\texttt{T2I}} : \mathcal{T} \to \mathcal{I}$, which synthesizes an image given a textual description. **Image Understanding (`I2T`)**: $\mathcal{UM}_{\texttt{I2T}} : \mathcal{I} \to \mathcal{T}$, which generates a textual description from a given image. Here, $\mathcal{T}$ denotes the set of text representations (e.g., captions, instructions), and $\mathcal{I}$ denotes the set of all possible image representations.

Let $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$ represent a dataset of N paired samples, where each $I_i \in \mathcal{I}$ and each $T_i \in \mathcal{T}$ is its corresponding caption. A *generation step* is defined as the application of either $\mathcal{UM}_{\texttt{T2I}}$ or $\mathcal{UM}_{\texttt{I2T}}$ to transform an input from one modality into the other. We define alternating chains of length G starting from either text or image. Let $g \in \{0, 1, \ldots, G\}$ be the generation step index. Then similar to the chains defined in [3] and as illustrated in Fig. 4, we consider two experimental setups depending on the initial modality:

- **Text-First-Chain**: Starting from $T^{(0)}$, each step applies `T2I` then `I2T`:

$$T^{(0)} \xrightarrow{\texttt{T2I}} I^{(1)} \xrightarrow{\texttt{I2T}} T^{(2)} \xrightarrow{\texttt{T2I}} I^{(3)} \ldots$$

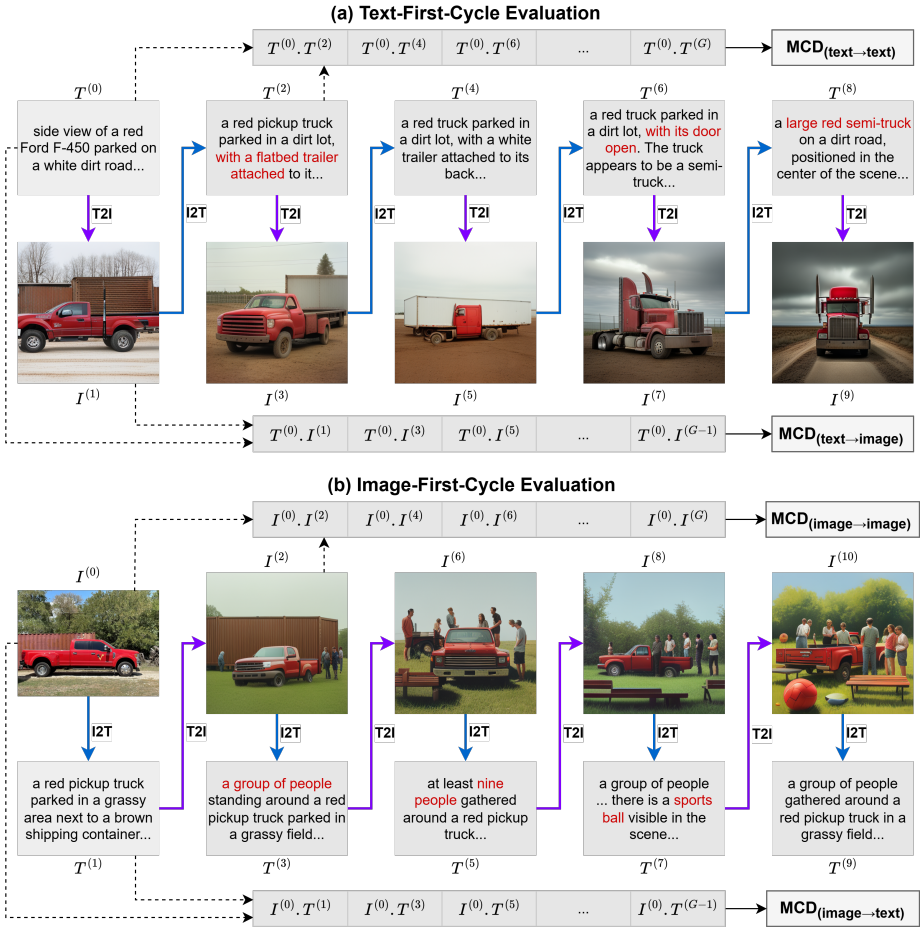


Fig. 4: Semantic Drift Protocol (SDP). We alternate between text-to-image (T2I) and image-to-text (I2T) generations in two setups: Text-First-Chain (a) and Image-First-Chain (b). Blue arrows denote I2T; purple arrows denote T2I; dashed black arrows indicate similarities computed back to the initial input in both same- and cross-modality directions used for MCD. Across generations, concepts drift despite plausible single steps: a "red F-450 truck" evolves into a semi-truck with changing attachments and positions; in the image-first chain, group size inflates and new objects (e.g., a sports ball) appear. The proposed cyclic evaluation reveals cross-modal concept drift that single-pass metrics overlook, enabling direct comparison of unified model's semantic stability

Here, similarity can be measured from initial text against later texts or images, giving the modality mappings $\{\text{text}\rightarrow\text{text}, \text{text}\rightarrow\text{image}\}$.

– **Image-First-Chain**: Starting from $I^{(0)}$, each step applies I2T then T2I:

$$I^{(0)} \xrightarrow{\texttt{I2T}} T^{(1)} \xrightarrow{\texttt{T2I}} I^{(2)} \xrightarrow{\texttt{I2T}} T^{(3)} \dots$$

Here, similarity can be measured from initial image against later images or texts, giving the modality mappings $\{\text{image} \to \text{image}, \text{image} \to \text{text}\}$.

Depending on the modality of initial input and the modality considered for distance calculation, we define a set of modality mappings, $\Delta = \{\text{text} \to \text{text}, \text{image} \to \text{text}, \text{text} \to \text{image}, \text{image} \to \text{image}\}$. The intuition for `SDP` is that a semantically consistent model will preserve the core meaning of the original content across many generations of alternating `T2I` and `I2T`; a weaker model, on the other hand will drift away from the original meaning more quickly. To systematically measure this degradation, in our protocol we propose two distinct metrics. `MCD` provides a holistic measure of drift based on embedding similarity. On the other hand, `MGG` grounds the evaluation in object-level fidelity by extending the GenEval benchmark across multiple generations.

3.1 MCD: Mean Cumulative Drift

`MCD` measures how much meaning a model can retain after multiple `T2I` and `I2T` cycles. To obtain this metric we compare the input with the output of later generations using embedding based similarity scores. For any dataset that has text-image pairs, we can construct two separate chains (Text-First and Image-First chains). Then, for each modality mapping $\delta \in \Delta$ we obtain a sequence of similarity scores across the generations. We then average the sequences at every generation along the entire dataset $\mathcal{D}$,

$$S_\delta(g) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \text{sim}\big(inp_d,\, M_{d,\delta}^{(g)}\big) \tag{1}$$

where $S_\delta(g)$ is the average similarity at generation g for modality mapping δ, $M_{d,\delta}^{(g)}$ is the generated text or image at generation g. Here, sim denotes a similarity function using one of the embedding models (CLIP, DINO, or MPNet). To get overall drift, we compute mean across generations $S_\delta(g)$,

$$\text{MCD}_\delta = \frac{1}{G} \sum_{g=1}^{G} \big(S_\delta(g)\big), \tag{2}$$

where MCD_δ is a single scalar denoting mean cumulative drift for a given modality mapping. To compute across all mappings, we compute mean across all modality mappings to get $\texttt{MCD}_{avg}$. A higher MCD_δ indicates stronger semantic retention across generations (and therefore lower drift), while a lower value indicates higher drift.

As `MCD` relies on embedding similarities, the choice of representation backbone may influence the measured drift. To reduce dependence on a single representation space, we compute drift across multiple embedding spaces rather than relying on one metric. Specifically, we evaluate drift using MPNet (text-text), DINO (image-image), and CLIP (cross-modal), which differ in architecture and training objectives. Further, we examine the effect of representation choices

in Supplementary Secs. 3.2–3.3. Overall, while different embedding spaces may slightly change the ordering of closely performing models, the overall trends remain consistent and align with human judgments, suggesting that SDP captures semantic degradation beyond representation overlap.

3.2 MGG: Multi-Generation GenEval

To complement embedding-based similarities with object-level fidelity, we further extend GenEval [10] to our proposed multi-generation setting. The existing protocol is designed to assess text-to-image fidelity across multiple dimensions of quality. These dimensions include *single object*, *two object*, *counting*, *colors*, and *positions*, and *attributes binding*. For each task, GenEval proposes a diverse set of prompts such as "a photo of a/an [COLOR] [OBJECT]". Once a model has generated images for all the prompts, GenEval uses a pre-trained object detection model to detect and localize objects in the generated images. This process allows us to calculate the accuracy of the model for each task. An average of the task level accuracies is then denoted by GenEval overall accuracy. We build on the existing benchmark by incorporating the rewritten GenEval prompts (provided in the BLIP3-o [5] repository), adopting the newer OwlV2 object detection model [21], and extending evaluation across multiple generations. To calculate MGG, we first calculate the GenEval scores for each generation for all tasks. Then, similar to GenEval overall accuracy, we compute the tasks scores to obtain GenEval overall accuracy for each generation. Finally, we average the generation scores to obtain the MGG score. Higher MGG scores indicate better ability to produce semantically accurate, context-preserving outputs. We provide a human audit of detector reliability in the Supplementary Sec. 3.5, showing that OWLv2 agrees well with human judgments and remains stable across early and late generations.

3.3 Single-Pass Human Evaluation (Cross-Consistency)

We complement our cyclic analysis with a single-step cross-consistency evaluation to highlight cross-modal fidelity issues, sampling 100 Text-First and 100 Image-First chains from the MCD evaluation set for a total of 200 examples. Given a ground-truth pair (I, T), we first generate a caption $T^{(1)} = \mathrm{UM}_{\mathtt{I2T}}(I)$ via I2T and an image $I^{(1)} = \mathrm{UM}_{\mathtt{T2I}}(T)$ via T2I. We then assess whether $T^{(1)}$ and $I^{(1)}$ preserve the semantics of (I, T) along two axes: (a) $I \rightarrow T^{(1)}$ consistency—does $T^{(1)}$ faithfully describe I? and (b) $T \rightarrow I^{(1)}$ consistency—does $I^{(1)}$ depict T? Six human annotators participated in the study; each received a comparable workload, and every example was evaluated independently by two different annotators. Using a web interface, annotators provided two judgments per sample: a three-level fidelity score (Good, Medium, or Poor) and a ranking of model outputs based on semantic correctness relative to the original input. To ensure unbiased evaluation and prevent positional bias, model identities were masked and the output order was randomized for every instance. Each sample page contained two sections: in the **understanding section**, annotators rated

and ranked captions for the input image; in the **generation section**, they rated and ranked generated images for the input text prompt. Finally, rather than averaging annotators' opinions, we keep annotations at the individual level, enabling consistency analysis while preserving annotators' distinct perspectives.

4 Evaluations & Findings

For embedding-based semantic drift analysis (`MCD`), we sample 400 image–text pairs from two challenging vision-language datasets: 200 pairs from NoCaps [1] and 200 pairs from DOCCI [24]. We refer to this 400-pair evaluation set as `Nocaps+Docci400`. These datasets stress complementary aspects of visual semantics. NoCaps contains novel object categories unseen in COCO and visually complex images, making it useful for testing out-of-domain generalization. DOCCI emphasizes fine-grained image–text correspondence, with captions covering attributes, spatial relationships, object counts, text rendering, and world knowledge. For object-level multi-generation evaluation (`MGG`), we use the GenEval protocol [10] with rewritten GenEval prompts from released evaluation resources [5], which provide longer descriptive prompts that better match the caption style produced during cyclic evaluation.

To diagnose where drift occurs, we further evaluate models under controlled common-vs.-novel and simple-vs.-complex settings, using 80 examples for each diagnostic subset. For `MCD`, the common-object subset comes from GenEval-R single-object prompts with COCO-style categories, while the novel-object subset uses NoCaps examples containing out-of-domain objects. For `MGG`, we compare simple prompts involving single objects and colors with compositional prompts involving multiple objects, counting, positioning, and attribute binding. These diagnostic results are reported in Tab. 2 and analyzed separately from the full benchmark. All full-set, diagnostic, and human-evaluation pairs are included with the released code. Our primary full-set ranking uses the seven open systems evaluated consistently across the benchmark; additional small-subset comparisons with Emu3 [39] and proprietary models are provided in Supplementary Sec. 3.1 as diagnostic results rather than part of the primary ranking.

Choosing the generation length G is an important design choice for cyclic evaluation, as it determines how long semantic drift can accumulate. If G is too small, drift may not fully manifest; if too large, most models may already have diverged from the original semantics. Hence, the maximum number of generations G should be large enough for drift to reliably emerge, but not so large that the evaluation is dominated by saturated, post-failure trajectories. We set G=20 based on the weakest model in our study (VILA-U), which reaches a near-zero `MGG` score by generation 19 in Fig. 6. This choice provides enough horizon to separate strong and weak models before saturation, and it yields strong agreement with our human evaluations.

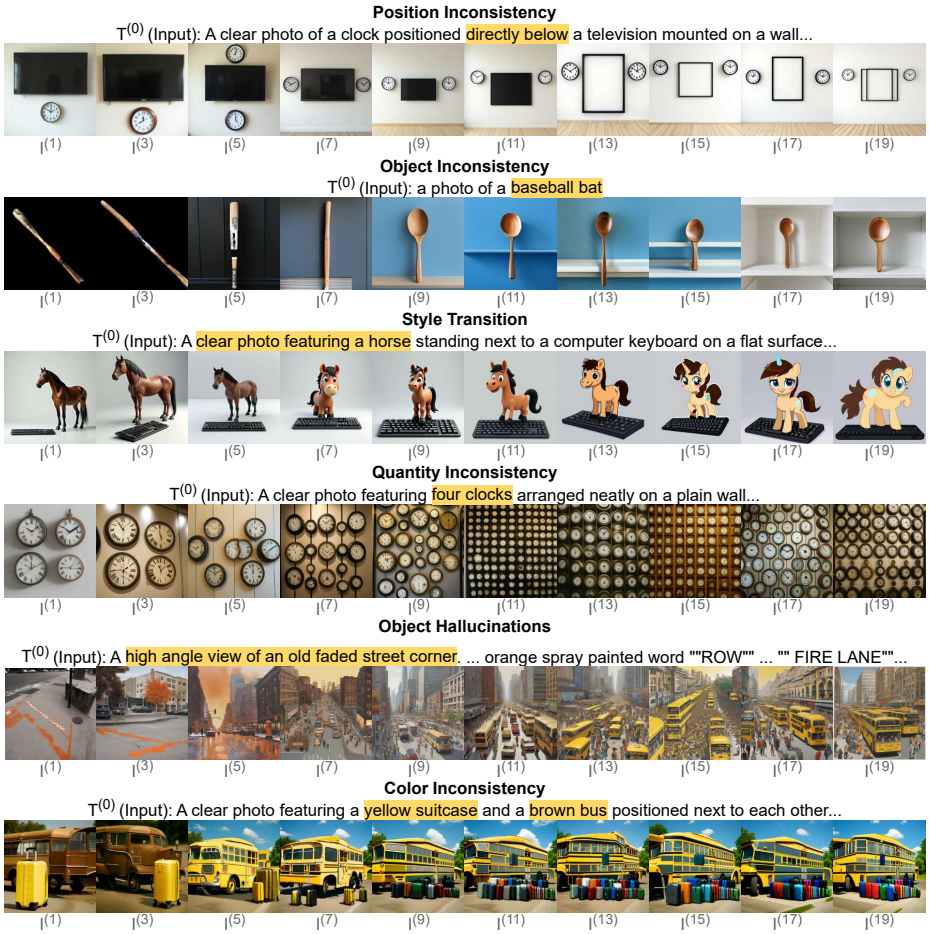


Fig. 5: Failure modes during cyclic generation. Starting from the input prompt $T^{(0)}$, the model repeatedly generates images $I^{(g)}$ through alternating `T2I`–`I2T` cycles. Across generations, semantic information can drift in multiple ways, including spatial relation errors, object changes, style shifts, count errors, hallucinated scene content, and color changes. These examples show that drift is not limited to object presence, but can affect several semantic attributes needed for faithful cyclic generation.

4.1 Semantic Drift Protocol Findings

We observe several qualitative failure patterns in cyclic generation. Fig. 5 illustrates six representative ways in which unified models lose information under alternating `T2I` $\leftrightarrow$ `I2T` cycles:

1. **Position inconsistency**, where the spatial relation "clock directly below the television" is gradually ignored; 2. **Object misidentification**, where a baseball bat is rendered or re-captioned as a spoon; 3. **Style transition**, where the visual style shifts from a realistic photograph to a cartoon-like depiction; 4.

Quantity inconsistency, where the specified count of four clocks is not preserved; 5. **Object hallucination**, where the scene expands to include a city environment not present in the prompt; and 6. **Color inconsistency**, where the brown bus changes into a yellow bus.

The overall model-level evaluation scores in Tab. 1 show that models with stronger object-level preservation generally also achieve higher embedding-based semantic retention across generations. A notable exception is the decoupled LLaVA+SDXL system, which scores relatively well on MGG but lower on $\texttt{MCD}_{\text{avg}}$, suggesting that it can render specific objects while losing holistic scene semantics. Across all evaluations, BAGEL shows the greatest resilience to semantic drift, likely due to its scale, unified architecture, and diverse interleaved training. Supplementary Sec. 2 provides a complete summary table with individual MCD components, MCD_{avg}, MGG, and human rankings.

Table 1: Semantic-drift results. BAGEL performs best on both embedding-based retention ($\texttt{MCD}_{\text{avg}}$) and object-level multi-generation preservation (MGG).

Model	MCD_{avg}	MGG
BAGEL	0.4414	78.2
BLIP3-o	0.3984	62.2
Show-o	0.3691	57.8
Janus-7B	0.3778	52.5
LLaVA+SDXL	0.3298	32.3
Janus-1.3B	0.3470	27.0
VILA-U	0.2829	19.7

Next, we present the empirical results in Fig. 7, which shows the scores from Eq. (1) for all four modality mappings, $\delta \in \{\text{text} \to \text{text}, \text{image} \to \text{text}, \text{text} \to \text{image}, \text{image} \to \text{image}\}$. These scores are later used to compute MCD. Ideally, similarities should remain nearly constant across generations. Instead, we observe consistent degradation in semantic fidelity, with modality-dependent asymmetries. Fig. 7(a) compares the original caption with text generated in the Text-First chain. Top-performing models start with high similarity (∼0.65–0.70), but only BAGEL maintains it relatively well, ending around 0.65. In contrast, VILA-U and Janus 1.3B decline more steeply, with VILA-U dropping below 0.40, indicating that its generations quickly lose connection to the original prompt. Fig. 7(b,d) provide a cross-modal view, evaluating text → image and image → text, respectively. In both cases, BAGEL maintains

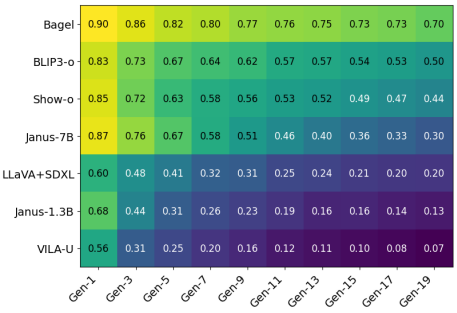
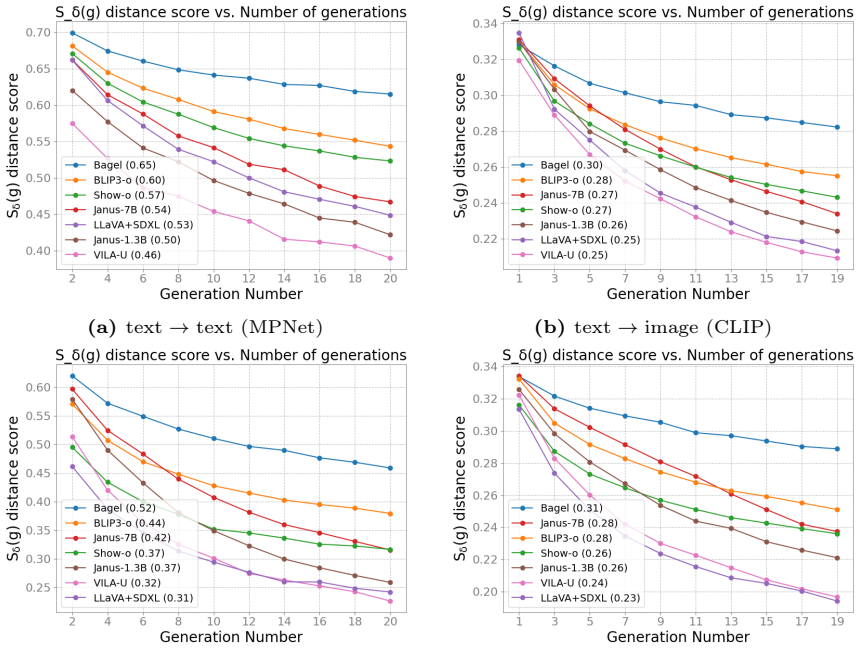


Fig. 6: Several models begin with high first-generation compliance, but repeated self-composition causes semantic drift: BAGEL remains comparatively stable, while weaker models collapse over later generations.

a clear lead, while VILA-U drifts severely at later stages. The final-generation ranking is identical across both plots. Fig. 7(c) measures visual fidelity by comparing the original image with generated images in the Image-First chain. While leading models follow similar trends, Janus 1.3B starts high in the first generation (0.6) but degrades substantially by the last generation. Overall, models may perform well initially yet lose context over later generations, a behavior not reliably captured by conventional single-pass metrics.



(a) text → text (MPNet) **(b)** text → image (CLIP)

(c) image → image (DINO) **(d)** image → text (CLIP)

Fig. 7: Similarity scores $S_\delta(g)$ computed using Eq. (1) across generations. Panels (a,b) show Text-First chains, while panels (c,d) show Image-First chains. Lower scores indicate greater semantic drift, and legend values report the average score across generations for each modality mapping.

Fig. 6 shows that while initial `MGG` scores are high, they can mask differences between models. For instance, BAGEL produces more faithful generations than SHOW-O even with similar initial scores, a divergence that only becomes numerically apparent in later generations as semantic drift occurs. This underscores that cyclic evaluation reveals quality differences that single-pass metrics obscure. Furthermore, performance collapses dramatically on compositional tasks like positioning and attribute binding (this trend is shown in the per-task MGG breakdown in Supplementary Sec. 3.6), suggesting this weakness is a main cause of semantic drift.

Table 2: Common/novel columns report MCD, while simple/complex columns report MGG. Models degrade more sharply on compositional constraints than on simpler contexts, indicating that drift is driven largely by loss of structured semantics.

	MCD		MGG	
Model	Common	Novel	Simple	Complex
BAGEL	0.3178	0.2969	92.4	75.9
BLIP3-o	0.3293	0.2897	87.2	49.5
Show-o	0.3152	0.2853	89.8	43.4
Janus-7B	0.3041	0.2776	55.0	14.9
LLaVA+SDXL	0.2453	0.2554	70.2	14.0
Janus-1.3B	0.2833	0.2675	78.9	39.6
VILA-U	0.2600	0.2540	49.5	8.6

Table 3: Average rank for understanding fidelity (I2T) and generation fidelity (T2I): 1 indicates the best-ranked model and 7 the worst among the models.

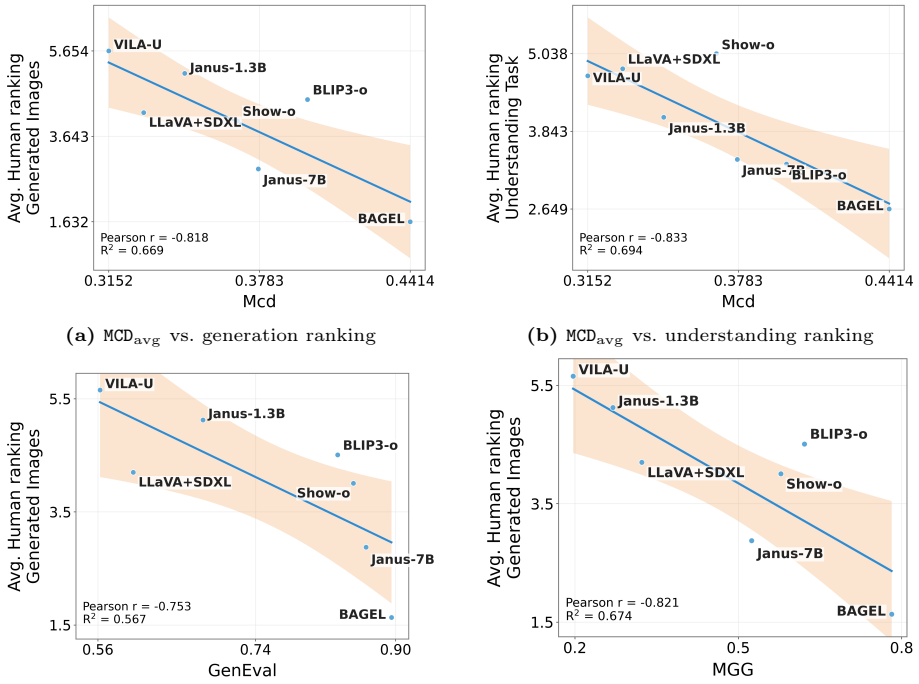
Model	I2T ↓	T2I ↓
BAGEL	2.65	1.63
BLIP3-o	3.34	4.51
Show-o	5.04	4.01
Janus-7B	3.41	2.88
LLaVA+SDXL	4.81	4.20
Janus-1.3B	4.06	5.12
VILA-U	4.70	5.65

Tab. 2 shows where drift is most severe. Models retain simple semantics more reliably, while complex constraints degrade much faster. The large simple–complex MGG gap shows that drift comes not only from losing objects, but also from failing to preserve counts, spatial relations, and attribute bindings. The common-vs.-novel MCD split shows a similar pattern: models stable on common objects also tend to be stable on novel objects. These findings suggest that drift is often catastrophic: once a critical mistake occurs (e.g., the brown bus turning yellow in Fig. 5), the original semantic content is effectively unrecoverable. Supplementary Sec. 3.7 quantifies this with a survival-style semantic-collapse analysis, reporting when each semantic property is first lost. After such loss, later generations mostly compound the error, and scores quickly saturate—especially for weaker models—indicating that the chain has already diverged. While secondary details may still erode, the metric largely plateaus once core entities, attributes, relations, or counts are compromised. To verify that these trends are not tied to a particular embedding choice, we also report CLIP-based text→text and image→image drift in Supplementary Sec. 3.2, and a Show-o representation-backbone ablation in Supplementary Sec. 3.3.

4.2 Human Evaluation Results

We summarize human preferences in Tab. 3, where rankings are averaged across samples separately for understanding and generation, and lower values indicate stronger preference. Here, BAGEL receives the best generation ranking and remains among the strongest models for understanding, while weaker automated scores generally correspond to worse human rankings. Additional qualitative examples of each cross-inconsistency type are included in Supplementary Sec. 4. We then compare these human rankings with our automated metrics to assess whether MCD and MGG reflect human-perceived semantic fidelity. Fig. 8(a,b) shows that $\texttt{MCD}_{\text{avg}}$ correlates strongly with human rankings, with Pearson correlations of $r = -0.818$ for T2I and $r = -0.833$ for I2T. Fig. 8(c,d) compares GenEval and

MGG against human rankings for image generation. GenEval obtains a Pearson correlation of $r = -0.753$, while MGG shows stronger agreement at $r = -0.821$, suggesting that multi-generation evaluation better captures human-perceived generation quality.



(a) MCD_{avg} vs. generation ranking

(b) MCD_{avg} vs. understanding ranking

(c) GenEval vs. generation ranking

(d) MGG vs. generation ranking

Fig. 8: Correlation between automated metrics and human judgments. Each point is a model, and lower human ranking is better. Both MCD and MGG align well with human preferences, supporting their use as proxies for semantic drift.

Beyond ranking, human evaluation further confirms that cross-consistency failures are asymmetric across modalities. We consider a model consistent when the caption and generated image for the same (I, T) pair receive the same fidelity rating (e.g., Good); otherwise, the mismatch indicates inconsistency and reveals which failure type is more prevalent. As shown in Fig. 9, most unified models perform better in image understanding than generation, leading primarily to Poor-Generation/Good-Understanding cross-inconsistency. BAGEL shows the strongest cross-consistency in Fig. 9(c), with fewer mismatches between understanding and generation fidelity.

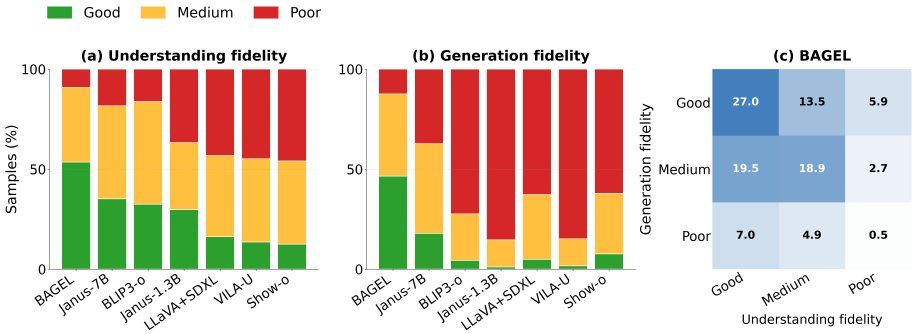


Fig. 9: Human evaluation of cross-consistency. First two plots show the percentage of samples (y-axis) rated with a fidelity score (color) for different models. We see that most models gain a high proportion of Poor fidelity score in image generation, whereas understanding is comparatively balanced, with BAGEL mostly getting Medium or better. The third plot illustrates a finer look at the responses for the BAGEL model. For BAGEL, only 7% of samples show Good-Understanding/Poor-Generation inconsistency, while 5.9% show Good-Generation/Poor-Understanding inconsistency, indicating stronger cross-consistency than other models.

5 Conclusion

We introduced the Semantic Drift Protocol (`SDP`), a cyclic evaluation framework that alternates image-to-text (`I2T`) and text-to-image (`T2I`) to measure how unified models preserve meaning over repeated modality shifts. By combining embedding-based metrics (`MCD`) and object-level fidelity (`MGG`), `SDP` reveals semantic degradation that single-pass evaluations often fail to capture. Evaluating seven recent models on the sampled `Nocaps+Docci400` benchmark shows substantial variability: BAGEL maintains the strongest cross-modal stability, VILA-U and Janus variants drift quickly, and Show-o, while not always leading initially, degrades more gracefully across generations. Human evaluations confirm these findings, showing that automated metrics like MCD strongly align with human judgments. These results demonstrate that single-pass benchmarks can overstate robustness, whereas our cyclic evaluation validated by human judgment reveals hidden inconsistencies between image understanding and image generation. These findings highlight the importance of cyclic evaluation for reliably assessing unified multimodal models.

Acknowledgments

The authors would like to thank Mamshad Nayeem Rizve for useful discussions.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Bahng, H., Chan, C., Durand, F., Isola, P.: Cycle consistency as reward: Learning image-text alignment without human preferences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22934–22946 (2025)
4. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
5. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
7. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
8. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), `https://arxiv.org/abs/2306.13394`
10. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
11. Gupta, R., Rizve, M.N., Unnikrishnan, J., Tawari, A., Tran, S., Shah, M., Yao, B., Chilimbi, T.: Open vocabulary multi-label video classification. In: European Conference on Computer Vision. pp. 276–293. Springer (2024)
12. Gupta, R., Unnikrishnan, J., Fei, F., Liu, S., Tran, S., Shah, M.: ViLL-E: Video LLM embeddings for retrieval. In: Liakata, M., Moreira, V.P., Zhang, J., Jurgens, D. (eds.) Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 43239–43258. Association for Computational Linguistics, San Diego, California, United States (Jul 2026), `https://aclanthology.org/2026.acl-long.2003/`
13. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

15. Kim, Y., Swetha, S., Kagdi, F., Shah, M.: Safe-llava: A privacy-preserving vision language dataset and benchmark for biometric safety. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 2100–2110 (June 2026)
16. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
17. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
19. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
20. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=KUNzEQMWU7
21. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36**, 72983–73007 (2023)
22. Narnaware, V., Gupta, A., Zhai, K., Wang, Z., Shah, M.: Seeing to ground: Visual attention for hallucination-resilient mdllms (2026), https://arxiv.org/abs/2603.25711
23. Narnaware, V., Vayani, A., Gupta, R., Swetha, S., Shah, M.: Bbq-v: Benchmarking visual stereotype bias in large multimodal models (2026), https://arxiv.org/abs/2502.08779
24. Onoe, Y., Rane, S., Berger, Z., Bitton, Y., Cho, J., Garg, R., Ku, A., Parekh, Z., Pont-Tuset, J., Tanzer, G., et al.: Docci: Descriptions of connected and contrasting images. In: European Conference on Computer Vision. pp. 291–309. Springer (2024)
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
27. Raza, S., Narayanan, A., Khazaie, V.R., Vayani, A., Radwan, A.Y., Chettiar, M.S., Singh, A., Shah, M., Pandya, D.: Humanibench: A human-centric framework for large multimodal models evaluation. arXiv preprint arXiv:2505.11454 (2025)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
29. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
30. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans (2016), https://arxiv.org/abs/1606.03498

31. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016)
32. Sirnam, S., Yang, J., Neiman, T., Rizve, M.N., Tran, S., Yao, B., Chilimbi, T., Shah, M.: X-former: Unifying contrastive and reconstruction learning for mllms. In: European Conference on Computer Vision. pp. 146–162. Springer (2024)
33. Swetha, S., Gupta, R., Kulkarni, P.P., Shatwell, D.G., A Chan Santiago, J., Siddiqui, N., Fioresi, J., Shah, M.: Vrr-qa: Visual relational reasoning in videos beyond explicit cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 32840–32849 (2026)
34. Swetha, S., Kuehne, H., Shah, M.: Timelogic: A temporal logic benchmark for video qa. arXiv preprint arXiv:2501.07214 (2025)
35. Swetha, S., Meng, R., Ram, S., Neiman, T., Tran, S., Shah, M.: Smpro: Self-supervised visual preference alignment via differentiable multi-preference multi-group ranking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 37951–37960 (2026)
36. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
37. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
38. Vayani, A., Dissanayake, D., Watawana, H., Ahsan, N., Sasikumar, N., Thawakar, O., Ademtew, H.B., Hmaiti, Y., Kumar, A., Kukreja, K., et al.: All languages matter: Evaluating lmms on culturally diverse 100 languages. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19565–19575 (2025)
39. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
40. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
41. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., Han, S., Lu, Y.: VILA-u: a unified foundation model integrating visual understanding and generation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=02haSpO453
42. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=o6Ynz6OIQ6
43. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
44. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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. pp. 9556–9567 (2024)
45. Zhou, C., YU, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and dif-

fuse images with one multi-modal model. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=SI2hI0frk6

46. Zou, K., Huang, Z., Dong, Y., Tian, S., Zheng, D., Liu, H., He, J., Liu, B., Qiao, Y., Liu, Z.: Uni-mmmu: A massive multi-discipline multimodal unified benchmark. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 908–924 (2026)