The Telephone Game: Evaluating Semantic Drift in Unified Models

Sabbir Mollah Rohit Gupta* Sirnam Swetha* Qingyang Liu† Ahnaf Munir† Mubarak Shah
Center For Research in Computer Vision, University of Central Florida, USA
{sabbir.mollah, rohit.gupta, Swetha.Sirnam, qingyang.liu2, ahnaf.munir}@ucf.edu, shah@crcv.ucf.edu

Abstract

Employing a single, unified model (UM) for both visual understanding (image-totext: `I2T`) and and visual generation (text-to-image: `T2I`) has opened a new direction in Visual Language Model (VLM) research. While UMs can also support broader unimodal tasks (e.g., text-to-text, image-to-image), we focus on the core cross-modal pair `T2I` and `I2T`, as consistency between understanding and generation is critical for downstream use. Existing evaluations consider these capabilities in isolation: FID and GenEval for `T2I`, and benchmarks such as MME, MMBench for `I2T`. These single-pass metrics do not reveal whether a model that "understands" a concept can also "render" it, nor whether meaning is preserved when cycling between image and text modalities. To address this, we introduce the Unified Consistency Framework for Unified Models (`UCF-UM`), a cyclic evaluation protocol that alternates `I2T` and `T2I` over multiple generations to quantify semantic drift. UCF formulates 3 metrics: (i) Mean Cumulative Drift (`MCD`), an embedding-based measure of overall semantic loss; (ii) Semantic Drift Rate (`SDR`), that summarizes semantic decay rate; and (iii) Multi-Generation GenEval (`MGG`), an object-level compliance score extending GenEval. To assess generalization beyond COCO, which is widely used in training; we create a new benchmark `ND400`, sampled from NoCaps and DOCCI and evaluate on seven recent models. `UCF-UM` reveals substantial variation in cross-modal stability: some models like BAGEL maintain semantics over many alternations, whereas others like Vila-u drift quickly despite strong single-pass scores. Our results highlight cyclic consistency as a necessary complement to standard `I2T` and `T2I` evaluations, and provide practical metrics to consistently assess unified model's cross-modal stability and strength of their shared representations. Code will be provided in this repository: https://github.com/mollahsabbir/Semantic-Drift-in-Unified-Models

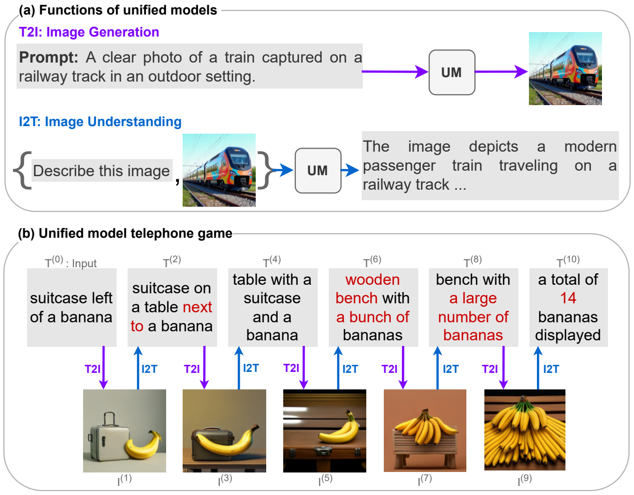

(a) Functions of unified models
T2I: Image Generation
Prompt: A clear photo of a train captured on a railway track in an outdoor setting.
UM
I2T: Image Understanding
Describe this image
UM
The image depicts a modern passenger train traveling on a railway track ...
(b) Unified model telephone game
T(0) : Input
suitcase left of a banana
T(2)
suitcase on a table next to a banana
T(4)
table with a suitcase and a banana
T(6)
wooden bench with a bunch of bananas
T(8)
bench with a large number of bananas
T(10)
a total of 14 bananas displayed
T2I
I2T
I(1)
I(3)
I(5)
I(7)
I(9)

Figure 1. (a) Illustrates image generation and image understanding functionalities of a unified model. (b) Example evaluation using proposed `UCF-UM` framework. In this example the unified model starts from a textual prompt $T^{(0)}$ about a suitcase and a banana. After successive `T2I` and `I2T` steps we observe in generation 5 the model fails to generate a convincing suitcase, and subsequently the suitcase disappeared from future generations. Also, at generation 5 it created two bananas instead of one, which culminated in lots of bananas in generation 7 and forward.

1. Introduction

Multimodal Unified Models (UMs) combine visual understanding and generation within a single framework, enabling them to perform a wide range of 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]. However, despite rapid model progress, UM evaluation remains fragmented. Existing metrics assess image understanding and image generation in isolation; *e.g.* MME, MMBench, POPE, VQA [1, 9, 12, 14] are used for evaluating understanding (`I2T`), and Inception score, CLIPScore, FID, GenEval [11, 20, 22] are used for evaluating image synthesis (`T2I`), while overlooking the reten-

[*] Equally contributing second author.
[†] Equally contributing third author.

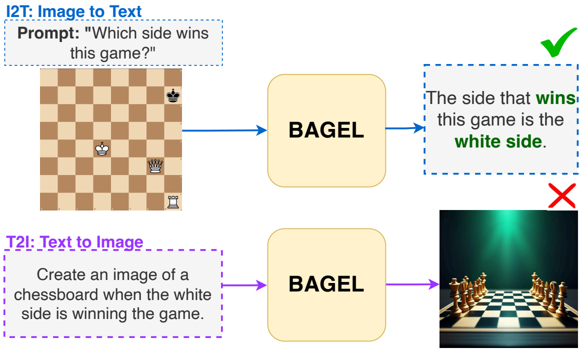

I2T: Image to Text
Prompt: "Which side wins this game?"
BAGEL
The side that wins this game is the white side.
T2I: Text to Image
Create an image of a chessboard when the white side is winning the game.
BAGEL

Figure 2. An example of unified inconsistency in the BAGEL unified model. Given an image of a chess board along with a question (top), BAGEL performs I2T correctly answering “white side wins”. By creating another caption for the T2I prompt (bottom), BAGEL should generate a chess board image consistent with the same semantic predicate (white winning side). However, the model generates a generic, mismatched chessboard image. This exposes a UM inconsistency: BAGEL’s correct visual reasoning (I2T) does not carry over to generation (T2I) for the concept “winning side in chess”.

tion of important information during T2I or I2T multi-turn conversion. In other words, current single-pass metrics do not assess the retention of entities, attributes, relations, and counts under alternating I2T ↔ T2I conversions. We defer unimodal tasks and center our analysis on I2T and T2I tasks as the potential for semantic divergence and its impact on real use is most pronounced on the cross-modal tasks.

Fig. 1 demonstrates this limitation. Much like the popular children’s game called *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 the cycles: by generation 5, the suitcase is no longer convincingly rendered and soon disappears entirely; meanwhile, the banana count inflates from one, to two, and eventually to many. Notably, a model may score well on single-pass I2T or T2I metrics, while still exhibiting these cross-modal inconsistencies, which the current metrics fail to capture.

There are several ways to evaluate a model’s image generation capabilities. For example, ClipScore [10] uses clip embeddings to measure semantic alignment of the prompt with generated images. However, it strongly relies on clip embeddings, which may not always be reflective with human perceptions [22]. Fréchet Inception Distance (FID) [11] measures the distributional similarity between the generated images and real images, 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 [22]. GenEval [22] improves on prompt alignment by checking object and relation-level compliance with detection models, but, by design, does not assess overall visual quality or realism, and like FID, remains a single-pass measure.

A similar limitation appears on the image-understanding benchmarks such as MME and MMBench [9, 14] which assess I2T skills in isolation, without testing whether the model’s understanding capability aligns with its generation capability. As illustrated in Fig. 2, even state-of-the-art unified models like BAGEL [8] can correctly reason about a chessboard image in I2T identifying that “the white side wins” yet fail to produce a faithful T2I image of the same winning scenario. This mismatch highlights the need for metrics that directly measure semantic drift across modalities rather than a single pass metric.

To address this gap, we propose the Unified Consistency Framework for Unified Models (UCF-UM), a cyclic evaluation protocol designed to quantify how well UMs preserve semantic meaning under repeated T2I and I2T conversions. Starting from an initial input $T^{(0)}$ (text) or $I^{(0)}$ (image), the model alternates T2I or I2T to produce a sequence $\{I^{(g)}, T^{(g)}\}$, where g denotes generation step. At each generation g, UCF-UM measures semantic similarity back to the initial input and across steps, capturing drift directions and exposing misalignment between a model’s understanding and generation spaces. We employ CLIP [19], DINO [4], and MPNet [26] embeddings for text–image, image–image, and text–text comparisons, respectively. For rigorous testing, we design three different metrics: Mean Cumulative Drift (MCD), Semantic Drift Rate (SDR), and Multi-Generation Geneval (MGG). In MCD, we use raw embedding distance scores to quantify cumulative information retention, SDR tells us properties of the drift through parameters such as decay rate, and MGG extends the GenEval benchmark for multiple generations. We propose a benchmark dataset ND400, sampling 200 image-text pairs from NoCaps [2] and 200 image-text pairs DOCCI [17] datasets. These two datasets were selected for their novel objects and fine-grained visual details that better probe generalization. We benchmark 7 recent models spanning shared-weight, partially shared, and decoupled architectures, to analyze how architectural design choices influence semantic stability.

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 [30] and Janus [29] degrade rapidly, exposing weaker coupling between their visual understanding and visual generation capabilities de-

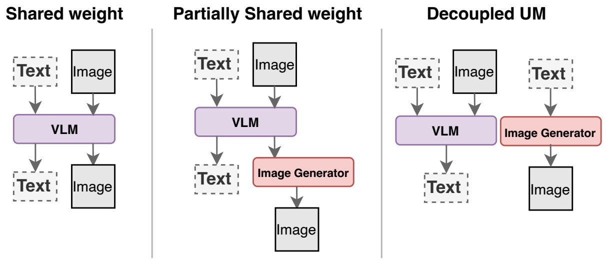


Figure 3. On the left, a single model handles both understanding and generation. In the middle, the architecture partially shares weights, with a decoder capable of generating text and visual features, the latter is passed to another image generation model. On the right, the understanding and generation processes are fully decoupled, using separate models for each task.

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

Our contributions are summarized as follows:

- We formalize semantic drift problem and show that single-pass understanding or generation metrics cannot expose gaps between a model's visual understanding and its image generation capabilities.
- We propose the Unified Consistency Framework for Unified Models (UCF-UM), which jointly evaluates image understanding (I2T) and image generation (T2I) through 3 different metrics, and tracks semantic preservation over multiple cross-modal transitions.
- We further extend GenEval [22] to a multi-generation setting, enabling it to capture semantic preservation beyond single prompt–generation pairs. In this setup, overall performance differences between models amplifies and becomes clearly observable.

2. Unified Models

T2I generation has advanced with diffusion-based models such as DALL·E 2 [21], Imagen [24], and Stable Diffusion [23], which synthesize high-fidelity images from textual prompts. Image captioning, on the other hand, has evolved from CNN-RNN pipelines [25] to transformer-based decoders [7, 13] trained with large web-scale data. Recent works in unified models have started investigating how to unite understanding and generation under one architecture.

Unified models employ visual and textual modalities as both input and output. The motivation is that these universal models facilitate richer semantic interoperability among the two tasks, I2T and T2I. Also, having a single architecture can reduce training time, and simplify deployment. Chameleon [27] is one of the early works in this domain which aimed to auto-regressively generate text tokens and image embeddings. Later, Transfusion [34] fused the autoregressive and diffusion loss within a single architecture. Show-o [31] has also used two different objectives, next token prediction for text generation, and masked token prediction [5] for image generation. Vila-u [30] uses next token prediction with different text and vision decoders. Janus and Janus-pro [29] employ separate encoders for image input during understanding and generation. The idea is that a model might require different level of information for understanding and generation. Other works like Blip-3o [6] demonstrates good quality of image generation by leveraging a separate diffusion transformer head. A recent work, BAGEL [8] demonstrates some unique capabilities of unified models by training on a large-scale interleaved dataset.

While most prior works focus on building a single model for both tasks, we propose a broader categorization that encompasses unified models as well as models that can emulate unified behavior.

Shared-Weights Unified Models This category has received the most attention in recent research. These models leverage a single model, typically a transformer decoder, to perform a wide spectrum of unimodal and cross-modal tasks, with T2I and I2T generation being prominent examples. The encoder component can vary where some models employ a shared visual encoder across tasks, while others use distinct encoders for generation and understanding. In our experiments, we use 5 such models: BAGEL [8], Janus 1.3B [29], Janus Pro 7B [29], Show-o [31], and Vila-u [30].

Partially Shared Models Models in this category retain a degree of parameter sharing, while delegating specific responsibilities to task-specific modules. This design allows more flexibility in handling modality-specific complexities while preserving shared knowledge across tasks. We use *Blip-3o* [6] which incorporates a dedicated diffusion model for image generation.

Decoupled Models Models in the third category are formed by constructing a unified pipeline by composing independently trained models, which in tandem can emulate unified behavior. The example we have used is pairing a VLM like *LLaVA* [13] for I2T with a *Stable Diffusion* [18] model for T2I. This setup enables task interoperability without requiring joint training or weight sharing.

3. Unified Consistency Evaluation

We propose a cyclic evaluation framework UCF-UM which provides three different metrics to measures how well a unified model preserves semantic fidelity when alternating between I2T and T2I. As illustrated by the single alternation in Fig. 2 where correct I2T reasoning about a

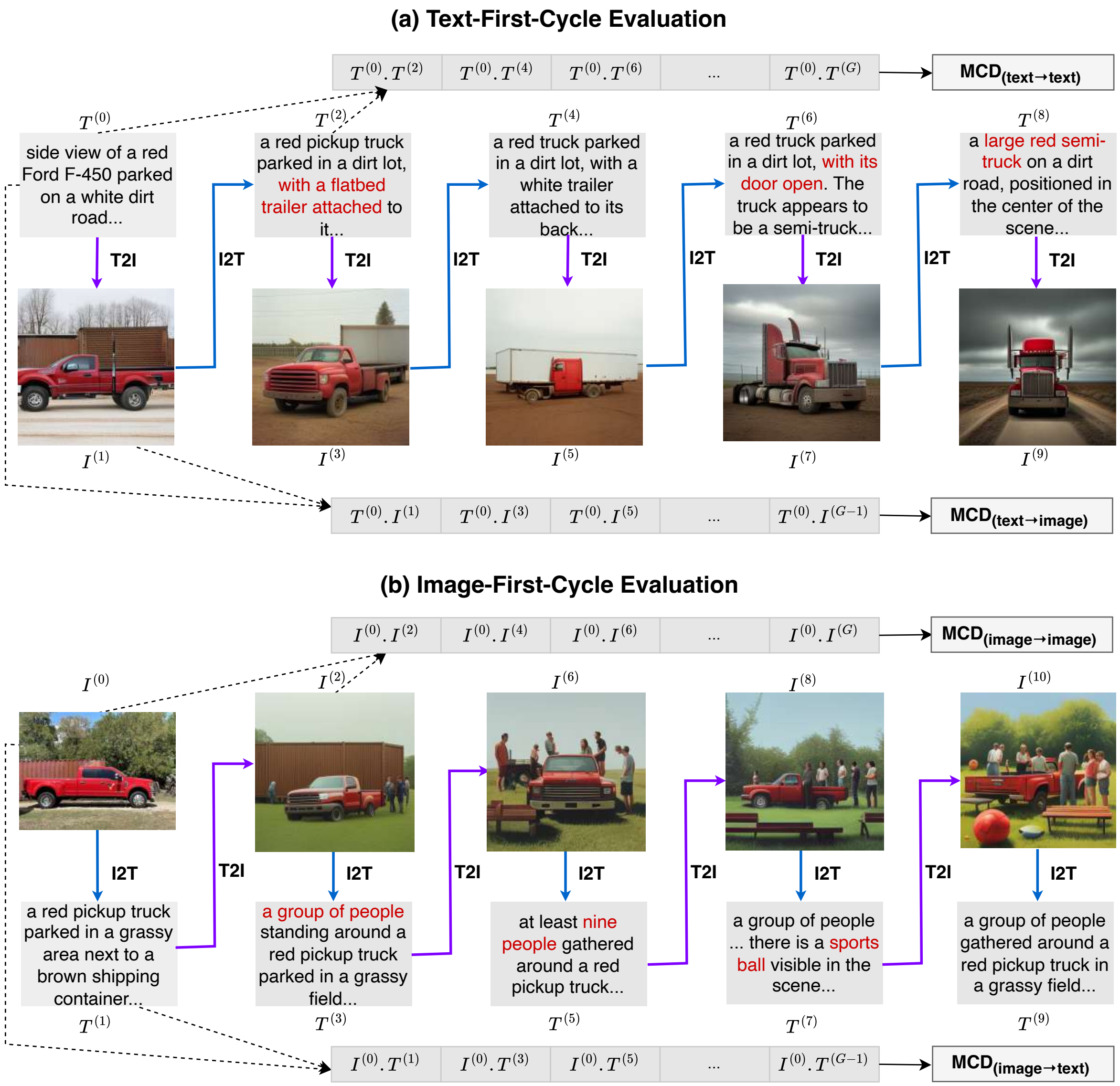

(a) Text-First-Cycle Evaluation
$T^{(0)}.T^{(2)}$
$T^{(0)}.T^{(4)}$
$T^{(0)}.T^{(6)}$
...
$T^{(0)}.T^{(G)}$
MCD(text→text)
$T^{(0)}$
side view of a red Ford F-450 parked on a white dirt road...
$T^{(2)}$
a red pickup truck parked in a dirt lot, with a flatbed trailer attached to it...
$T^{(4)}$
a red truck parked in a dirt lot, with a white trailer attached to its back...
$T^{(6)}$
a red truck parked in a dirt lot, with its door open. The truck appears to be a semi-truck...
$T^{(8)}$
a large red semi-truck on a dirt road, positioned in the center of the scene...
T2I
I2T
$I^{(1)}$
$I^{(3)}$
$I^{(5)}$
$I^{(7)}$
$I^{(9)}$
$T^{(0)}.I^{(1)}$
$T^{(0)}.I^{(3)}$
$T^{(0)}.I^{(5)}$
...
$T^{(0)}.I^{(G-1)}$
MCD(text→image)
(b) Image-First-Cycle Evaluation
$I^{(0)}.I^{(2)}$
$I^{(0)}.I^{(4)}$
$I^{(0)}.I^{(6)}$
...
$I^{(0)}.I^{(G)}$
MCD(image→image)
$I^{(0)}$
$I^{(2)}$
$I^{(6)}$
$I^{(8)}$
$I^{(10)}$
I2T
T2I
a red pickup truck parked in a grassy area next to a brown shipping container...
a group of people standing around a red pickup truck parked in a grassy field...
at least nine people gathered around a red pickup truck...
a group of people ... there is a sports ball visible in the scene...
a group of people gathered around a red pickup truck in a grassy field...
$T^{(1)}$
$T^{(3)}$
$T^{(5)}$
$T^{(7)}$
$T^{(9)}$
$I^{(0)}.T^{(1)}$
$I^{(0)}.T^{(3)}$
$I^{(0)}.T^{(5)}$
...
$I^{(0)}.T^{(G-1)}$
MCD(image→text)

Figure 4. Unified Consistency Evaluation framework (UCF-UM). 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.

chessboard does not translate into a faithful T2I rendering; UCF-UM generalizes this intuition to multi-generation cycles and provides quantitative measures of semantic drift. This framework considers how quickly the information from the initial input is drifted in subsequent generations. In this setting, we treat the $\mathcal{UM}$ as a model composed of at least two functionalities.

- **Image Generation (T2I)**: $\mathcal{UM}_{\texttt{T2I}} : \mathcal{T} \rightarrow \mathcal{I}$, which synthesizes an image given a textual description.
- **Image Understanding (I2T)**: $\mathcal{UM}_{\texttt{I2T}} : \mathcal{I} \rightarrow \mathcal{T}$, which generates a textual description from a given image and a prompt (*e.g.*'Describe this image')

Here, $\mathcal{T}$ denotes the set of all possible 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], we consider two experimental setups depending on the initial modality:

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

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

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

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

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

  Here, similarity can be measured from initial image against later images or texts, giving the distance 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 distance 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 `UCF-UM` 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 will drift away from the original meaning more quickly. To systematically measure this degradation, in our framework we propose three distinct metrics. `MCD` provides a holistic measure of drift based on embedding similarity. `SDR` offers a granular analysis of the decay rate by fitting it to a power-law function. Finally, `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 distance mapping $\delta \in \Delta$ we obtain a sequence of distance scores across the generations. We then average the sequences at every generation along the entire dataset $\mathcal{D}$,

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

where $S_\delta(g)$ is the average similarity at generation g for distance mapping δ, $M_{d,\delta}^{(g)}$ is the generated text or image at generation g, and sim denotes the similarity function from Tab. 1. 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 integer denoting mean cumulative drift for a given distance mapping. To compute across all mappings, we compute mean across all distance mappings to get $\texttt{MCD}_{avg}$. A higher `MCD` means the chain retains its semantic meaning more consistently across generations, while a lower value indicates higher drift.

3.2. SDR: Semantic Drift Rate

While `MCD` provides a holistic measure of semantic preservation, the Semantic Drift Rate (`SDR`) quantifies how rapidly a model's outputs diverge from the original input. In `MCD`, a single distance mapping with high similarity values can dominate the overall score, potentially masking drift in other mappings. SDR addresses this by fitting a power curve across all directions, and obtain an averaged power curve, producing three parameters: α, β, γ, which capture scaling, decay rate, and baseline similarity, respectively.

The power-law decay curve has the form $y = \alpha \cdot g^{-\beta} + \gamma$, where y denotes similarity at generation g, α represents the extrapolated initial similarity, and γ denotes the asymptotic baseline, the value at which a model eventually collapses to. Here, β denotes the decay rate. β is followed by a negative sign, which ensures that g^{-b} decreases as g increases, reflecting the fact that similarity diminishes over successive generations. More details on the power-law curve is provided in Appendix Fig. **??**. The power-law form is chosen because it models the empirically observed nonlinear decay: a sharp drop in early generations followed by a slower decline in later cycles.

To compute the `SDR` parameters, we use the dataset-level similarity scores $S_\delta(g)$ obtained in Eq. 1. We fit a power-law decay model, across the scores yielding the parameters α_δ, β_δ, and γ_δ for each $\delta \in \cdot$.

Finally we average each parameter across all distance mapping δ to obtain the final α, β, and γ which are used to construct the `SDR` power-law curve.

3.3. MGG: Multi-Generation GenEval

To complement embedding-based similarities with object-level fidelity, we further extend GenEval [22] to our proposed multi-generation setting. The existing framework [22] 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*.

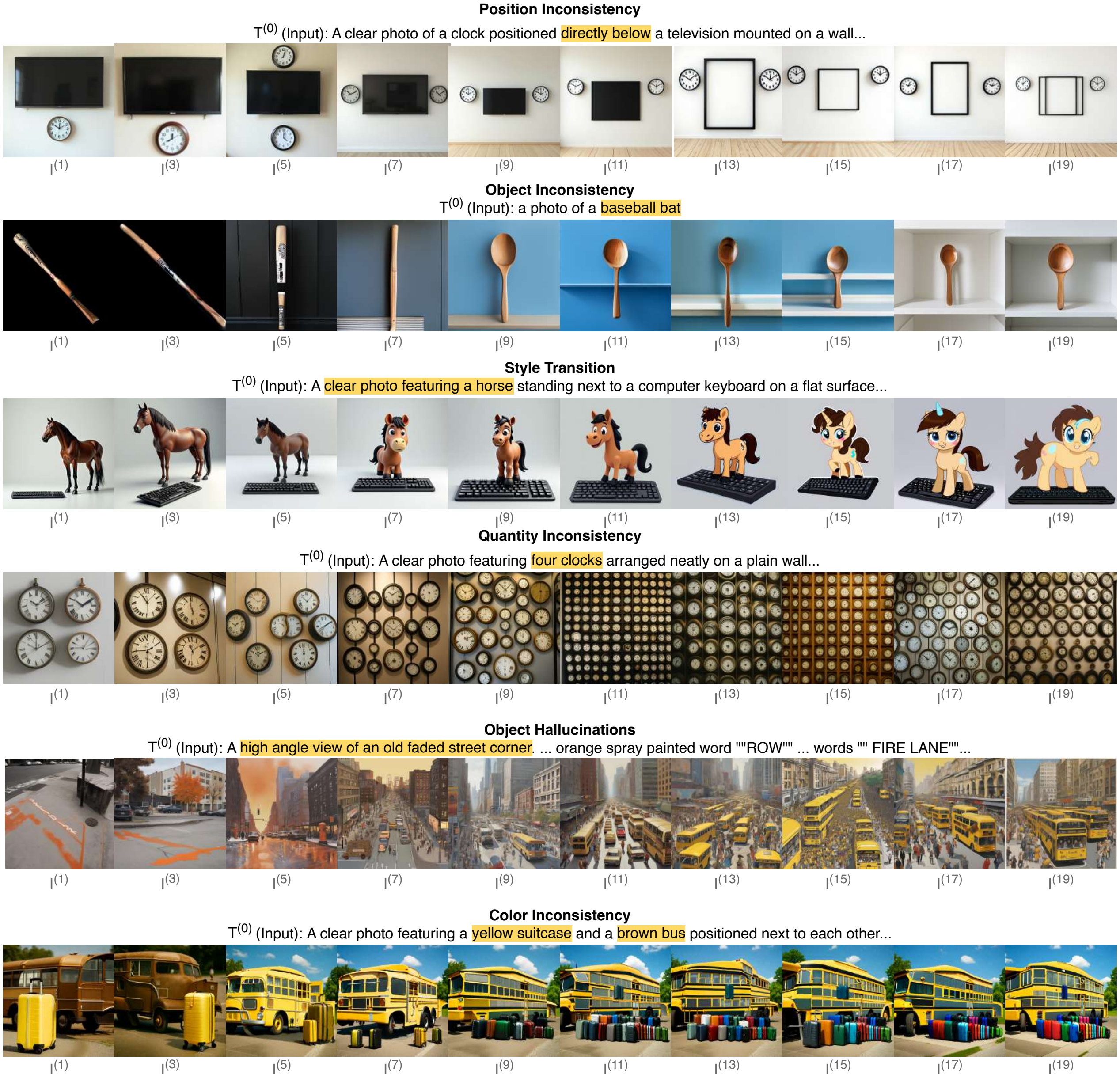

Position Inconsistency
$T^{(0)}$ (Input): A clear photo of a clock positioned directly below a television mounted on a wall...
$I^{(1)}$ $I^{(3)}$ $I^{(5)}$ $I^{(7)}$ $I^{(9)}$ $I^{(11)}$ $I^{(13)}$ $I^{(15)}$ $I^{(17)}$ $I^{(19)}$
Object Inconsistency
$T^{(0)}$ (Input): a photo of a baseball bat
$I^{(1)}$ $I^{(3)}$ $I^{(5)}$ $I^{(7)}$ $I^{(9)}$ $I^{(11)}$ $I^{(13)}$ $I^{(15)}$ $I^{(17)}$ $I^{(19)}$
Style Transition
$T^{(0)}$ (Input): A clear photo featuring a horse standing next to a computer keyboard on a flat surface...
$I^{(1)}$ $I^{(3)}$ $I^{(5)}$ $I^{(7)}$ $I^{(9)}$ $I^{(11)}$ $I^{(13)}$ $I^{(15)}$ $I^{(17)}$ $I^{(19)}$
Quantity Inconsistency
$T^{(0)}$ (Input): A clear photo featuring four clocks arranged neatly on a plain wall...
$I^{(1)}$ $I^{(3)}$ $I^{(5)}$ $I^{(7)}$ $I^{(9)}$ $I^{(11)}$ $I^{(13)}$ $I^{(15)}$ $I^{(17)}$ $I^{(19)}$
Object Hallucinations
$T^{(0)}$ (Input): A high angle view of an old faded street corner. ... orange spray painted word ""ROW"" ... words "" FIRE LANE""...
$I^{(1)}$ $I^{(3)}$ $I^{(5)}$ $I^{(7)}$ $I^{(9)}$ $I^{(11)}$ $I^{(13)}$ $I^{(15)}$ $I^{(17)}$ $I^{(19)}$
Color Inconsistency
$T^{(0)}$ (Input): A clear photo featuring a yellow suitcase and a brown bus positioned next to each other...
$I^{(1)}$ $I^{(3)}$ $I^{(5)}$ $I^{(7)}$ $I^{(9)}$ $I^{(11)}$ $I^{(13)}$ $I^{(15)}$ $I^{(17)}$ $I^{(19)}$

Figure 5. Information can be lost in different ways during a cyclic inference. In the first row, the model ignores the position of the clock, which is a crucial detail. In the second row, the model changes a baseball bat into a spoon. A model can also change the style from realistic to cartoon, as shown in the third row. In the fourth row the model loses count of four clocks and generates lots of clocks instead. In the fifth row a whole city is hallucinated around an empty road. In the sixth row, the model changes a brown bus into a yellow bus.

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 GenEval Rewritten dataset [6], adopting the newer OwlV2 object detection model [16], 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 and, context-preserving outputs.

4. Evaluations & Findings

4.1. Datasets

For embedding based semantic drift analysis (MCD, SDR), we randomly sample 200 image-text pairs from each of the two challenging vision-language datasets, Nocaps [2] and DOCCI [17]. We denote this sample dataset as ND400. These corpora stress both novel objects and fine-grained details, making them well-suited to reveal drift that single-pass metrics do not capture. NoCaps introduces nearly 400 novel objects unseen in COCO and features more visually complex images. The novel objects enables testing models on out-of-domain. DOCCI was specifically curated to evaluate fine-grained reasoning in image-text models. The image captions cover attributes, spatial relationships, object counts, text rendering, and world knowledge. These data allow will allow us to evaluate models in their descriptive understanding or generation capabilities. For multi-generation GenEval evaluations (MGG), we employ the *GenEval Rewritten* dataset [6], which extends the short GenEval prompts into long descriptive texts which better match models' outputs.

4.2. Evaluation of Unified Consistency

To conduct our evaluations, we construct three independent evaluation chains as shown in Tab. 1. Chain-1 (Exp. 1) is a Text-First-Chain that uses GenEval-Rewritten dataset. This chain is used exclusively for the multi-generation GenEval score (MGG). Chain-2 (Exps. 2–4) is also a Text-First-Chain constructed by initializing from the captions provided in ND400. Chain-3 (Exps. 5–7) follows the Image-First-Chain and is initialized from the images in ND400. Chain-2 and Chain-3 are used to calculate the similarity S used to compute our embedding-based metrics: Mean Cumulative Drift (MCD) and Semantic Drift Rate (SDR).

4.3. Findings

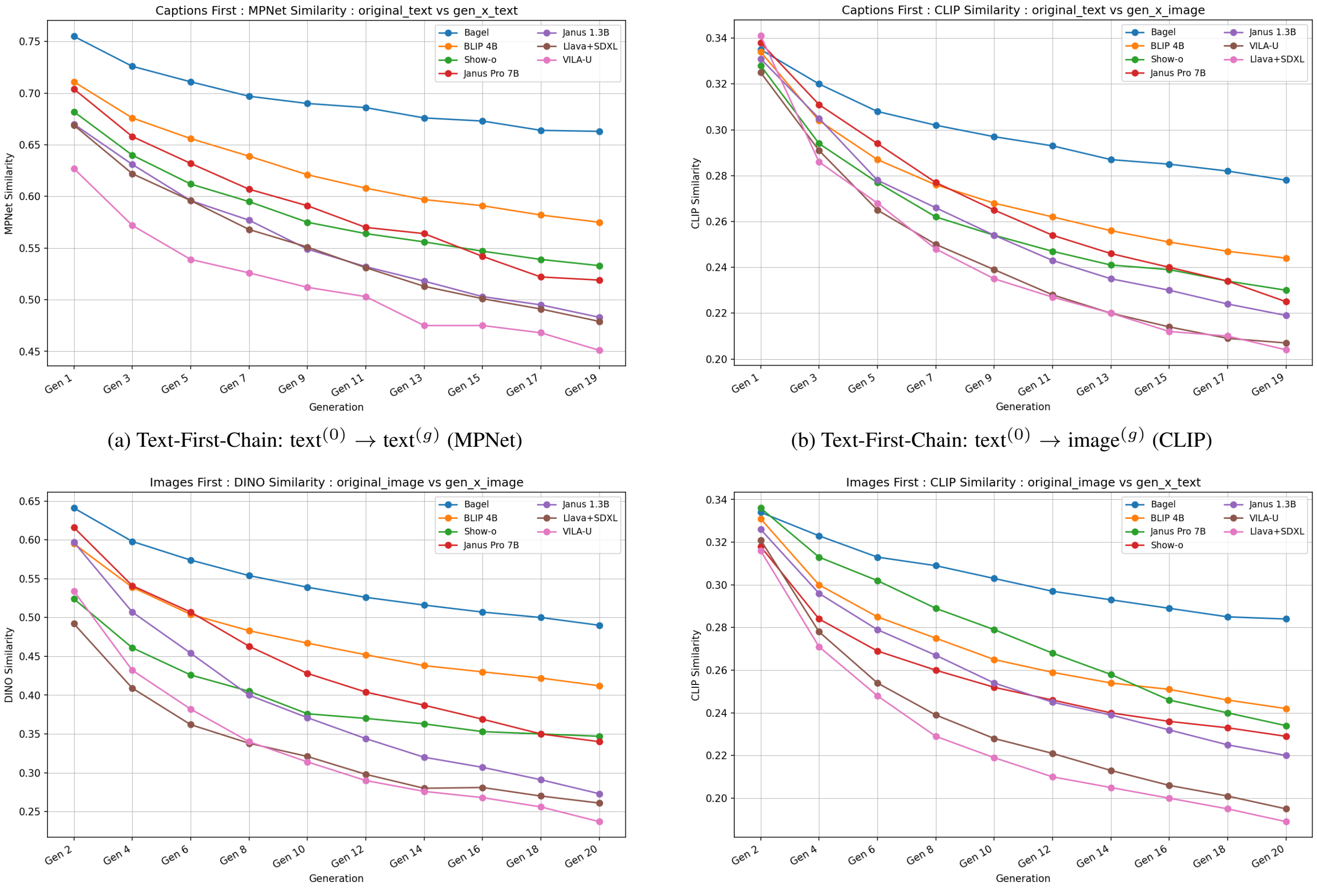
From our evaluations, we observe several interesting qualitative patterns. Fig. 5 illustrates six of such different ways in which unified models lose information under alternating T2I ↔ I2T cycles: 1. **Position Inconsistency**: the model fails to preserve spatial relationships that are central to the scene, 2. **Object Misidentification**: low-fidelity renderings lead to incorrect re-captioning, 3. **Style Transition**: the model may change the style of an image, particularly for rare object pairings (e.g., a horse on a keyboard), 4. **Quantity Inconsistency**: numerical counts may be inflated, 5. **Object Hallucinations**: new elements are introduced, 6. **Color Inconsistency**: important colors are not retaine

Exp.	Dataset	Chain (start)	Direction (δ)	Similarity backbone	Metric(s)
1	GenEval-R	Text-First	-	-	MGG
2	ND400	Text-First	text → text	MPNet	MCD, SDR
3	ND400	Text-First	text → text	CLIP	MCD, SDR
4	ND400	Text-First	text → image	CLIP	MCD, SDR
5	ND400	Image-First	image → image	DINO	MCD, SDR
6	ND400	Image-First	image → image	CLIP	MCD, SDR
7	ND400	Image-First	image → text	CLIP	MCD, SDR

Table 1. Summary of experimental settings. ND400 contains image-caption pairs, whereas GenEval Rewritten uses captions only. Each experiment is defined by its starting modality (text-first or image-first), the transformation direction (δ), and the similarity backbone used for evaluation. CLIP for text-to-text, text-to-image, and image-to-image; DINO for image-to-image; and a MPNet for text-to-text. Experiment 1 corresponds to the multi-step GenEval configuration, and Experiments 2-7 capture semantic drift on ND400 through embedding-level similarity based metrics (MCD, SDR). GenEval-R: GenEval-Rewritten

Next, we present the empirical results in Fig. 6 which shows the scores obtained from Eq. 1 for all distance mappings, $\{\text{text} \rightarrow \text{text}, \text{image} \rightarrow \text{text}, \text{text} \rightarrow \text{image}, \text{image} \rightarrow \text{image}\}$. These scores are later used to obtain both MCD and SDR. In the ideal case, the similarities should remain nearly constant across generations. Instead, as shown in these plots we observe consistent degradation in semantic fidelity, with modality dependent asymmetries. Plot 6(a) measure the similarity between the original caption and the text generated in Text-First-Chain. Top-performing models such as BAGEL and Blip-3o start with a high similarity ($\sim$ 0.70-0.75) and maintains it relatively well, ending around 0.60. In contrast, models like VILA-U and Janus 1.3B exhibit a much steeper decline, with VILA-U's similarity dropping below 0.50, indicating that its generated texts or images quickly lose connection to the original prompt. Plot 6(b) and Plot 6(d) offers a cross-modal perspective, evaluating the text → image, and image → text respectively. In both scenarios, BAGEL maintains a clear lead, while VILA-U's generations drift so severely that their relevance to the original text becomes minimal at later stages. Across both plots, the overall model ranking at the last step is exactly same. Plot 6(c) measure visual fidelity by comparing the original image to the generated images at subsequent steps in Image-First-Chain. While the leading models perform similar to prior trends discussed above, we notice Janus 1.3B scoring high in the first generation (0.6), but eventually degrading to a low score in the last generation. Overall, this behavior of models performing well in the first generation, but eventually losing context along the generations is a characteristic not reliably captured by conventional single-pass metrics.

Fig. 7, gives us a fine-grained look at the decay. The parameters used to plot these curves are shown in Tab. 3. The β parameter tells us the decay rate for each model. BAGEL's curves are visibly flatter, which is due to its lower β value, indicating slower drift and greater cross-modal sta-



(a) Text-First-Chain: text$^{(0)}$ $\rightarrow$ text$^{(g)}$ (MPNet)

(b) Text-First-Chain: text$^{(0)}$ $\rightarrow$ image$^{(g)}$ (CLIP)

(c) Image-First-Chain: image$^{(0)}$ $\rightarrow$ image$^{(g)}$ (DINO)

(d) Image-First-Chain: image$^{(0)}$ $\rightarrow$ text$^{(g)}$ (CLIP)

Figure 6. $S_{\delta}(g)$ distance scores computed using Eq. 1. Plots showing Text-First (a)(b) and Image-First (c)(d) chains that illustrate semantic drift across generations.

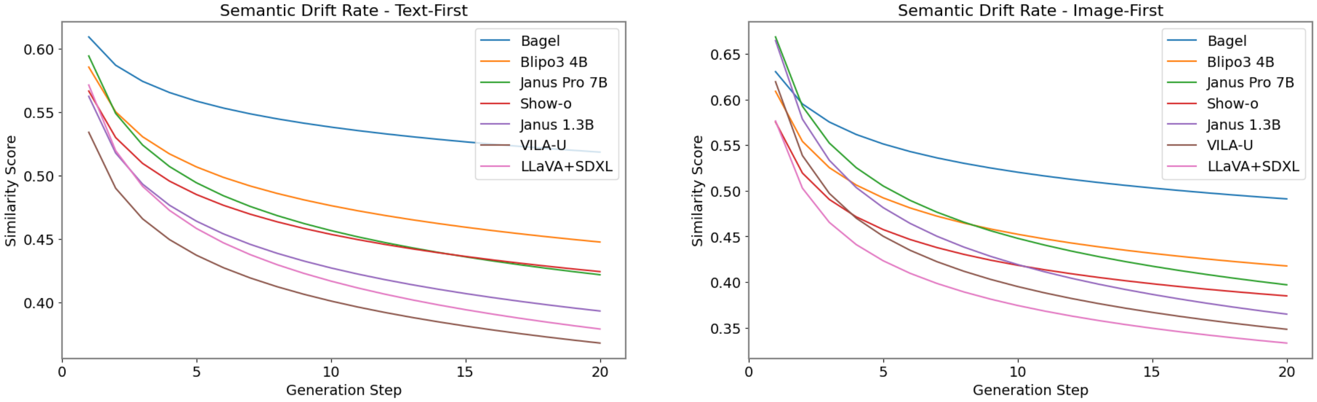
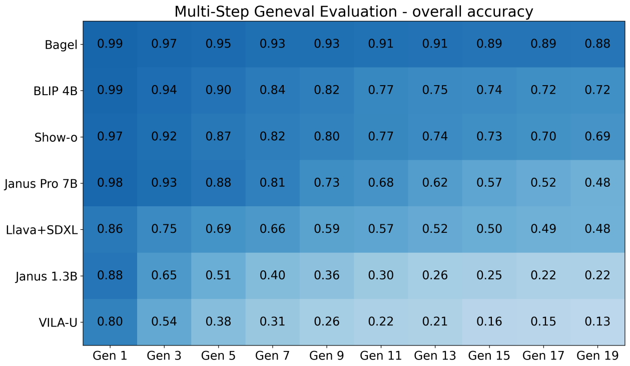


Figure 7. Results for SDR obtained using the technique in Eq. 3.2 on both text-first and image-first settings. The results show a similar ranking across both settings. Although Show-o seems to be doing worse than Janus 1.3B in the text-first setup, the decay rate hints that it might perform better in the later generations.

bility. In contrast, LLaVA+SDXL and Janus 1.3B decline more steeply, reflecting faster semantic decay as shown in Fig. 7. Although Show-o appears weaker than Janus 1.3B in early text-first steps, its slower decay rate (β) suggests relatively better behavior in later generations.

Fig. 8 reports MGG overall performance across cate-



	Gen 1	Gen 3	Gen 5	Gen 7	Gen 9	Gen 11	Gen 13	Gen 15	Gen 17	Gen 19
Bagel	0.99	0.97	0.95	0.93	0.93	0.91	0.91	0.89	0.89	0.88
BLIP 4B	0.99	0.94	0.90	0.84	0.82	0.77	0.75	0.74	0.72	0.72
Show-o	0.97	0.92	0.87	0.82	0.80	0.77	0.74	0.73	0.70	0.69
Janus Pro 7B	0.98	0.93	0.88	0.81	0.73	0.68	0.62	0.57	0.52	0.48
Llava+SDXL	0.86	0.75	0.69	0.66	0.59	0.57	0.52	0.50	0.49	0.48
Janus 1.3B	0.88	0.65	0.51	0.40	0.36	0.30	0.26	0.25	0.22	0.22
VILA-U	0.80	0.54	0.38	0.31	0.26	0.22	0.21	0.16	0.15	0.13

Figure 8. MGG results on the GenEval Rewritten dataset. This heatmap shows the overall performance across the six tasks described in the GenEval [22] benchmark. On average, BAGEL consistently drifts the least from the semantic meaning of the original caption. In contrast, VILA-U and Janus 1.3B lose more than half of their first-generation score within just a few generations.

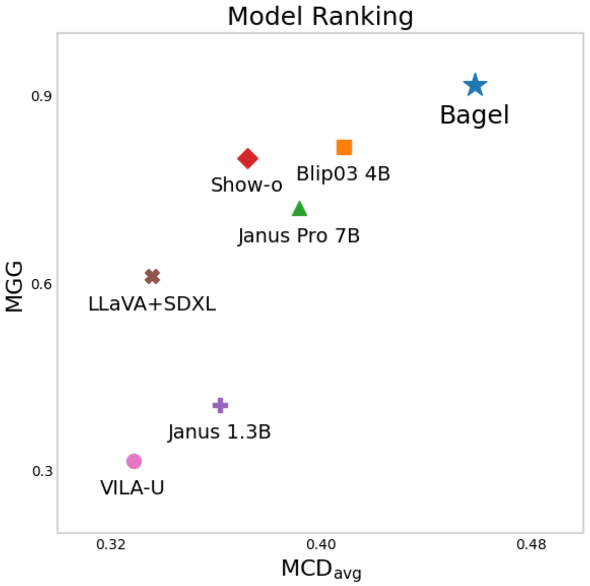


Figure 9. Comparison across MCD and MGG shows that BAGEL achieves the highest performance on both metrics, while VILA-U lags in both. LLaVA+SDXL and Janus 1.3B demonstrate strength in one metric but underperform in the other.

gories. First generation scores are above 0.8 and >0.95 for stronger models, which can mask qualitative differences; BAGEL, for instance, produces more faithful, higher-quality objects than peers despite similar initial scores. As generations accumulate, the distribution widens and robustness differences emerge. As shown in Fig. 8, the results highlights that early quantitative parity does not guarantee long-run stability: during Generations 1-2, SHOW-O attains overall GenEval and DINO similarities comparable to BAGEL, yet side-by-side inspection shows BAGEL is consistently more coherent and faithful. This divergence becomes numerically apparent only in later generations as SHOW-O's decay accelerates, underscoring that cyclic evaluation reveals quality differences that single-pass metrics can obscure. The results in separate GenEval tasks (Fig. 11) reveal a critical vulnerability in these models: their performance collapses most dramatically on compositional tasks like positioning and attribute binding. This may suggest weakness in composite attribute binding to be a cause for semantic drift.

Finally, Fig. 9 summarizes model performance by plotting MGG (y-axis) against MCD_{avg} (x-axis), aligning embedding-level and object-level metrics. The rankings reveal a correlation between the two metrics for certain models (e.g., VILA-U, BAGEL). The decoupled LLaVA+SDXL system exhibits a unique performance profile, where it performs well across MGG, but struggles in MCD. It struggles to preserve the holistic semantics or the overall "vibe" of a scene, as reflected in its low embedding-level scores. However, its powerful generator allows it to perform comparatively better on object-level tasks, successfully rendering the primary subjects from a caption even as the broader context degrades. This hints at a disconnect between maintaining the content of an image and preserving its essential meaning. Across all evaluations, BAGEL consistently demonstrates the greatest resilience to semantic drift. This superior performance is likely due to its larger scale, as well as its architecture and training methodology. Further, BAGEL has a lot of carefully chosen architectural components, and it has been trained on a diverse set of filtered data, including various interleaved multimodal datasets. This may enable BAGEL to preserve important details, making it uniquely robust against compounding errors observed in our MGG evaluation framework.

5. Conclusion

We presented the Unified Consistency Framework (UCF), a cyclic evaluation that alternates image-to-text (I2T) and text-to-image (T2I) with complementary metrics (MCD for overall drift, SDR for decay rate, and MGG for multi-generation GenEval) to quantify how unified models preserve semantics over repeated modality shifts. Evaluating seven recent models on the sampled ND400 dataset 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. These results demonstrate that single-pass benchmarks can overstate robustness, whereas cyclic evaluation reveals hidden inconsistencies between understanding and generation. We demonstrate that cyclic consistency is essential for reliable assessment, decay rates matter as much as initial scores.

Acknowledgements We would like to thank Dr. Mamshad Nayeem Rizve for the insightful discussions.

A. Appendix

This appendix provides additional details and extended analyses that complement the results presented in the main paper. We first describe the models used in our experiments, including their parameterization and image generation settings. We then report further evaluations using CLIP embeddings, provide fitted parameters for our decay function, and present comprehensive results from the extended multi-generation GenEval benchmark.

A.1. Models & Parameters

Tab. 2 lists the models included in our evaluations, along with their parameter counts and image resolutions used during generation.

Name	Parameters	Image Resolution
BAGEL	14B - Mixture of Transformers (7B Active)	1024×1024
Show-o	1.3B	512×512
Janus	1.3B	1024×1024
Janus Pro	7B	1024×1024
VILA-U	7B	256×256
Blip-3o	4B	1024×1024
LLaVA 1.5 + SDXL	7B + 3.5B	1024×1024

Table 2. Overview of models used in our experiments, including parameter counts and image resolution. The BAGEL model is a mixture-of-transformers architecture, where 7B parameters are active during inference.

A.2. Related Works

Prior Evaluations A variety of benchmarks have been proposed to evaluate the multimodal capabilities of vision-language models. MME [9] assesses basic perception and reasoning through fine-grained tasks such as object existence, color, and OCR. MMBench [14] introduces more complex queries, especially in spatial reasoning. MMMU [33] focuses on college-level academic problems in fields such as science and art. MM-VET [32] covers diverse skills, including math, OCR, and spatial understanding. MathVista [15] targets mathematical reasoning in visual contexts such as graphs. MMVP [28] highlights flaws in existing benchmarks using CLIP-similar but human-atypical images. FID [11] provides a metric-based evaluation of image generation quality, while Geneval [22] benchmarks generative vision language models in instruction follow-up and visual grounding. Iterative text-image generation loops have rarely been studied in systematic depth. The work in [3] is the closest in spirit where they they use cycle-consistency to create a preference dataset. However, this work only looks at one generation and is limited to VLM models in general and does not consider unified models.

A.3. More Results Using CLIP Embeddings

The main paper Fig. 6 presents $S_\delta(g)$ results for text → text and image → image settings using MPNet (for textual embeddings) and DINO (for visual embeddings). Here, we extend this analysis by incorporating CLIP as an additional backbone, shown in Fig. 10.

For the Text-First-Chain, **text** → **text** comparison shown in Fig. 10 (a), CLIP similarities are consistently lower than those produced with MPNet as shown in Fig. 6 (a). Despite this, the overall ranking of models is preserved as BAGEL continues to outperform others.

For the Image-First-Chain, $image \rightarrow image$ comparison shown in Fig. 10 (b), the models have higher similarities in the first generation compared to DINO in Fig. 6 (c). The relative order of model performance remains consistent with DINO.

A.4. SDR Parameters

A lower b, makes the graph more flat, indicating less semantic drift. Tab. 3 gives us the parameter values from our experiments.

Model	Text-First			Image-First		
	α	β	γ	α	β	γ
Bagel	0.6092	0.0538	0.0000	0.6305	0.0834	0.0001
Blip-3o	0.5854	0.0896	0.0000	0.4272	0.1984	0.1818
Janus Pro 7B	0.5942	0.1143	0.0000	0.6687	0.1740	0.0000
Show-o	0.5665	0.0965	0.0000	0.3919	0.2224	0.1836
Janus 1.3B	0.5624	0.1193	0.0000	0.6647	0.2002	0.0000
VILA-U	0.5341	0.1243	0.0000	0.5323	0.2378	0.0873
LLaVA+SDXL	0.5713	0.1369	0.0000	0.4586	0.2525	0.1180

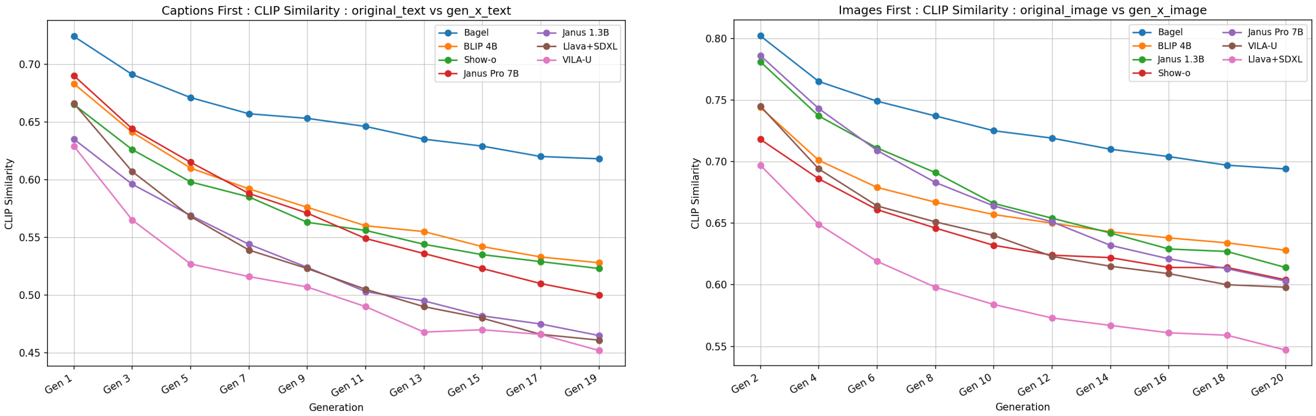
Table 3. Fitted power-law parameters (α, β, γ) for Text-First and Image-First experiments.

A.5. Analysis of Multi-Generation GenEval Results

Fig. 11 shows multi-generation performance in the six tasks from GenEval benchmark. In these heatmaps, darker shades represent higher accuracy. Results from later generations reveal that a model's proficiency in complex tasks is highly susceptible to generational decay, a weakness that single-step evaluations fail to capture.

Plot 11(a) Single Object: The simplest task, requiring generation of a single specified object. Nearly every model achieves near-perfect accuracy in the first generation, but consistency issues appear quickly. VILA-U shows clear degradation, struggling to maintain even one concept.

Plot 11(b) Two Objects: This task assesses handling two entities. The performance drop-off is more pronounced than in the single-object case. Models like Janus 1.3B and LLaVA+SDXL lose the ability to consistently generate both objects after only a few generations.



(a) Text-First-Chain: text$^{(0)}$ → text$^{(g)}$ (CLIP)

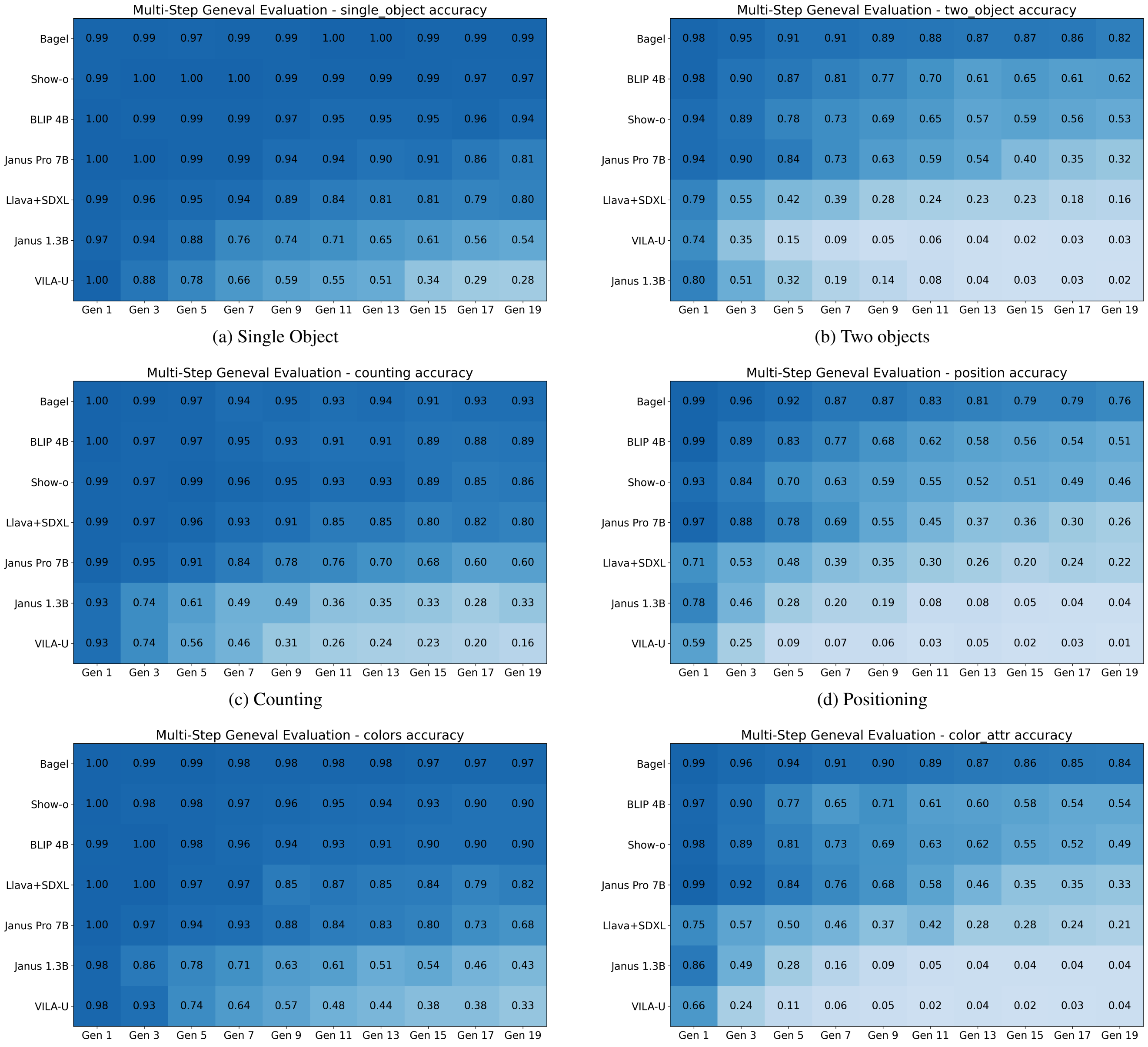
(b) Image-First-Chain: image$^{(0)}$ → image$^{(g)}$ (CLIP)

Figure 10. We show $S_\delta(g)$ **distance scores** computed using CLIP for both text → text and image → image.

Plot 11(c) Counting: Tests counting capabilities. Initial accuracy is high, but many models fail rapidly, replacing precise numbers (e.g., “three dogs”) with vague quantities (e.g., “some dogs”), leading to cascading errors in subsequent generations.

Plot 11(d) Positioning: Evaluates spatial reasoning (e.g., “a cup to the left of a plate”). Accuracy plummets after the first generation for most models. Preserving spatial relationships proves extremely difficult. BAGEL maintains accuracy longer than other models.

Plots 11(e) Colors & 11(f) Color Attribute: These assess attribute binding. “Colors” is simpler, while “Color Attribute” requires binding colors to specific objects. Both show rapid decay, particularly (f). Models often forget or swap colors. Only top performers retain any meaningful accuracy beyond the initial generations.



Multi-Step Geneval Evaluation - single_object accuracy

	Gen 1	Gen 3	Gen 5	Gen 7	Gen 9	Gen 11	Gen 13	Gen 15	Gen 17	Gen 19
Bagel	0.99	0.99	0.97	0.99	0.99	1.00	1.00	0.99	0.99	0.99
Show-o	0.99	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.97	0.97
BLIP 4B	1.00	0.99	0.99	0.99	0.97	0.95	0.95	0.95	0.96	0.94
Janus Pro 7B	1.00	1.00	0.99	0.99	0.94	0.94	0.90	0.91	0.86	0.81
Llava+SDXL	0.99	0.96	0.95	0.94	0.89	0.84	0.81	0.81	0.79	0.80
Janus 1.3B	0.97	0.94	0.88	0.76	0.74	0.71	0.65	0.61	0.56	0.54
VILA-U	1.00	0.88	0.78	0.66	0.59	0.55	0.51	0.34	0.29	0.28

(a) Single Object

Multi-Step Geneval Evaluation - two_object accuracy

	Gen 1	Gen 3	Gen 5	Gen 7	Gen 9	Gen 11	Gen 13	Gen 15	Gen 17	Gen 19
Bagel	0.98	0.95	0.91	0.91	0.89	0.88	0.87	0.87	0.86	0.82
BLIP 4B	0.98	0.90	0.87	0.81	0.77	0.70	0.61	0.65	0.61	0.62
Show-o	0.94	0.89	0.78	0.73	0.69	0.65	0.57	0.59	0.56	0.53
Janus Pro 7B	0.94	0.90	0.84	0.73	0.63	0.59	0.54	0.40	0.35	0.32
Llava+SDXL	0.79	0.55	0.42	0.39	0.28	0.24	0.23	0.23	0.18	0.16
VILA-U	0.74	0.35	0.15	0.09	0.05	0.06	0.04	0.02	0.03	0.03
Janus 1.3B	0.80	0.51	0.32	0.19	0.14	0.08	0.04	0.03	0.03	0.02

(b) Two objects

Multi-Step Geneval Evaluation - counting accuracy

	Gen 1	Gen 3	Gen 5	Gen 7	Gen 9	Gen 11	Gen 13	Gen 15	Gen 17	Gen 19
Bagel	1.00	0.99	0.97	0.94	0.95	0.93	0.94	0.91	0.93	0.93
BLIP 4B	1.00	0.97	0.97	0.95	0.93	0.91	0.91	0.89	0.88	0.89
Show-o	0.99	0.97	0.99	0.96	0.95	0.93	0.93	0.89	0.85	0.86
Llava+SDXL	0.99	0.97	0.96	0.93	0.91	0.85	0.85	0.80	0.82	0.80
Janus Pro 7B	0.99	0.95	0.91	0.84	0.78	0.76	0.70	0.68	0.60	0.60
Janus 1.3B	0.93	0.74	0.61	0.49	0.49	0.36	0.35	0.33	0.28	0.33
VILA-U	0.93	0.74	0.56	0.46	0.31	0.26	0.24	0.23	0.20	0.16

(c) Counting

Multi-Step Geneval Evaluation - position accuracy

	Gen 1	Gen 3	Gen 5	Gen 7	Gen 9	Gen 11	Gen 13	Gen 15	Gen 17	Gen 19
Bagel	0.99	0.96	0.92	0.87	0.87	0.83	0.81	0.79	0.79	0.76
BLIP 4B	0.99	0.89	0.83	0.77	0.68	0.62	0.58	0.56	0.54	0.51
Show-o	0.93	0.84	0.70	0.63	0.59	0.55	0.52	0.51	0.49	0.46
Janus Pro 7B	0.97	0.88	0.78	0.69	0.55	0.45	0.37	0.36	0.30	0.26
Llava+SDXL	0.71	0.53	0.48	0.39	0.35	0.30	0.26	0.20	0.24	0.22
Janus 1.3B	0.78	0.46	0.28	0.20	0.19	0.08	0.08	0.05	0.04	0.04
VILA-U	0.59	0.25	0.09	0.07	0.06	0.03	0.05	0.02	0.03	0.01

(d) Positioning

Multi-Step Geneval Evaluation - colors accuracy

	Gen 1	Gen 3	Gen 5	Gen 7	Gen 9	Gen 11	Gen 13	Gen 15	Gen 17	Gen 19
Bagel	1.00	0.99	0.99	0.98	0.98	0.98	0.98	0.97	0.97	0.97
Show-o	1.00	0.98	0.98	0.97	0.96	0.95	0.94	0.93	0.90	0.90
BLIP 4B	0.99	1.00	0.98	0.96	0.94	0.93	0.91	0.90	0.90	0.90
Llava+SDXL	1.00	1.00	0.97	0.97	0.85	0.87	0.85	0.84	0.79	0.82
Janus Pro 7B	1.00	0.97	0.94	0.93	0.88	0.84	0.83	0.80	0.73	0.68
Janus 1.3B	0.98	0.86	0.78	0.71	0.63	0.61	0.51	0.54	0.46	0.43
VILA-U	0.98	0.93	0.74	0.64	0.57	0.48	0.44	0.38	0.38	0.33

(e) Colors

Multi-Step Geneval Evaluation - color_attr accuracy

	Gen 1	Gen 3	Gen 5	Gen 7	Gen 9	Gen 11	Gen 13	Gen 15	Gen 17	Gen 19
Bagel	0.99	0.96	0.94	0.91	0.90	0.89	0.87	0.86	0.85	0.84
BLIP 4B	0.97	0.90	0.77	0.65	0.71	0.61	0.60	0.58	0.54	0.54
Show-o	0.98	0.89	0.81	0.73	0.69	0.63	0.62	0.55	0.52	0.49
Janus Pro 7B	0.99	0.92	0.84	0.76	0.68	0.58	0.46	0.35	0.35	0.33
Llava+SDXL	0.75	0.57	0.50	0.46	0.37	0.42	0.28	0.28	0.24	0.21
Janus 1.3B	0.86	0.49	0.28	0.16	0.09	0.05	0.04	0.04	0.04	0.04
VILA-U	0.66	0.24	0.11	0.06	0.05	0.02	0.04	0.02	0.03	0.04

(f) Color attribute

Figure 11. **Detailed Multi-Generation GenEval (MGG) Results.** Performance of unified models usingMGG across 20 generations for six different evaluation categories: (a) Single Object, (b) Two Objects, (c) Counting, (d) Positioning, (e) Colors, and (f) Color Attribute. Darker colors indicate higher accuracy. The results show that while initial performance is high for many models, consistency varies significantly over successive generations, especially for complex tasks.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. 1

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019. 2, 7

[3] Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences, 2025. 5, 10

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 2

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer, 2022. 3

[6] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 3, 6, 7

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning, 2020. 3

[8] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. 1, 2, 3

[9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 1, 2, 10

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1, 2, 10

[12] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 1

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[14] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 1, 2, 10

[15] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. 10

[16] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2024. 6

[17] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. Docci: Descriptions of connected and contrasting images, 2024. 2, 7

[18] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 3

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2

[20] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 1

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 3

[22] Ehud Reiter and Anja Belz. GENEVAL: A proposal for shared-task evaluation in NLG. In Nathalie Colineau, Cécile Paris, Stephen Wan, and Robert Dale, editors, *Proceedings of the Fourth International Natural Language Generation Conference*, pages 136–138, Sydney, Australia, July 2006. Association for Computational Linguistics. 1, 2, 3, 5, 9, 10

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 3

[25] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, 2015. 3

[26] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. 2

[27] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. 3

[28] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. 10

[29] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024. 2, 3

[30] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2025. 2, 3

[31] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation, 2024. 3

[32] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. 10

[33] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. 10

[34] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. 3