The Telephone Game: Evaluating Semantic Drift in Unified Models

Sabbir Mollah Rohit Gupta† Sirnam Swetha† Qingyang Liu‡ Ahnaf Munir‡ Mubarak Shah

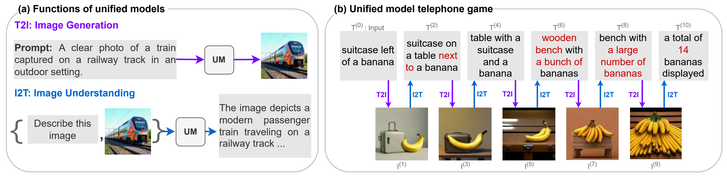
Institute of Artificial Intelligence, University of Central Florida

ECCV 2026 | MALMÖ | SEPT 8–12

1 Unified Models Need Cyclic Evaluation

- **Unified models support** image understanding (I2T) and image generation (T2I).

Like a telephone game, image-text cycles reveal semantic drift.

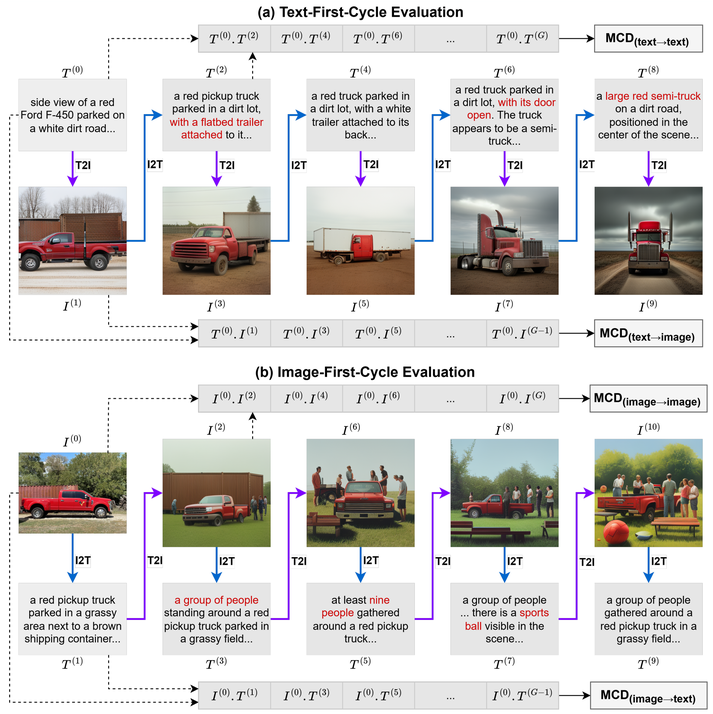


Single-Pass Evaluation Falls Short

- **T2I metrics** judge either realism or object/attribute presence in one generation.
- **I2T benchmarks** test understanding alone, without alignment to generation.
- **Unified evaluation should test** whether meaning survives I2T–T2I composition.

Repeated cycles expose generation–understanding mismatch and semantic drift.

2 Semantic Drift Protocol



- **Cyclic evaluation tests** whether meaning survives repeated I2T–T2I use.
- **Mean Cumulative Drift (MCD)** measures embedding-level retention.
  - MCD summarizes semantic retention over repeated cycles.

$$\mathrm{MCD}_\delta = \operatorname*{avg}_{g,d} \operatorname{sim}\left(\mathrm{inp}_d, M^{(g)}_{d,\delta}\right)$$

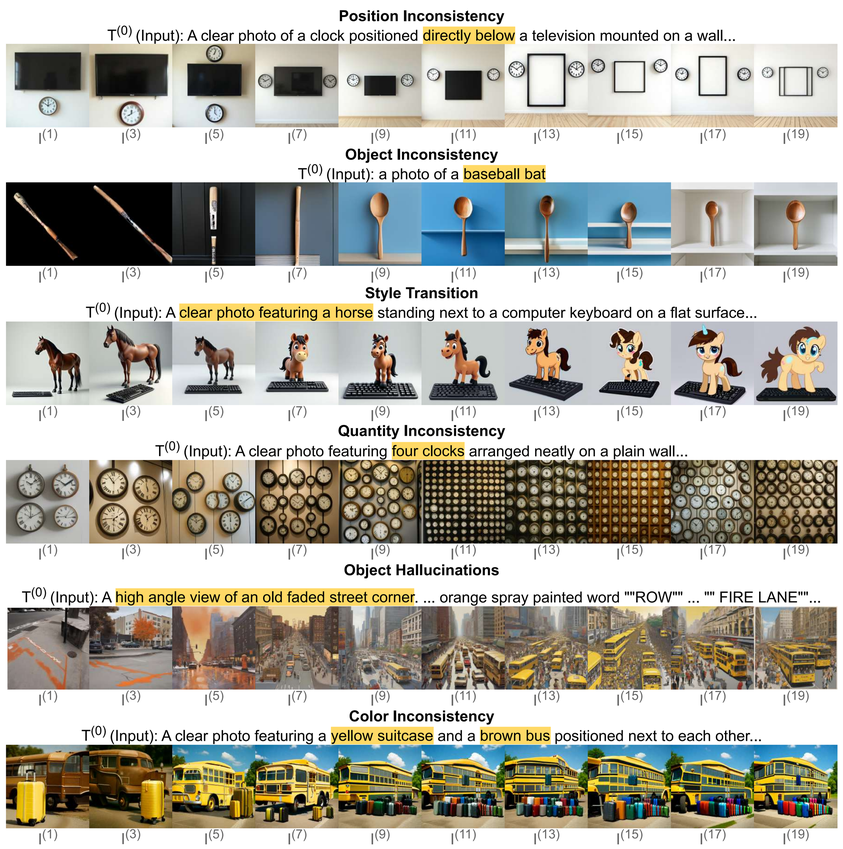
- **Multi-Generation GenEval (MGG)** measures object-level generation fidelity.
  - MGG extends GenEval across generations to track prompt fidelity.

$$\mathrm{MGG} = \operatorname*{avg}_{g} \mathrm{GenEval}\left(I^{(g)}, T^{(0)}\right)$$

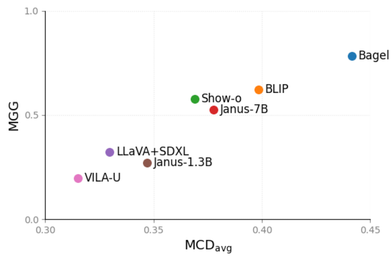
3 Qualitative Examples

Repeated cycles reveal drift modes that single-pass scores obscure.

Drift appears as object substitution, attribute loss, relation changes, and caption narrowing.



4 Comparison Across Unified Models



- BAGEL leads both retention measures.
- Every model drifts over repeated cycles.
- LLaVA+SDXL keeps objects but loses meaning.
- Single-pass scores do not predict cyclic stability.

Prompt Ablation: Rankings Remain Stable

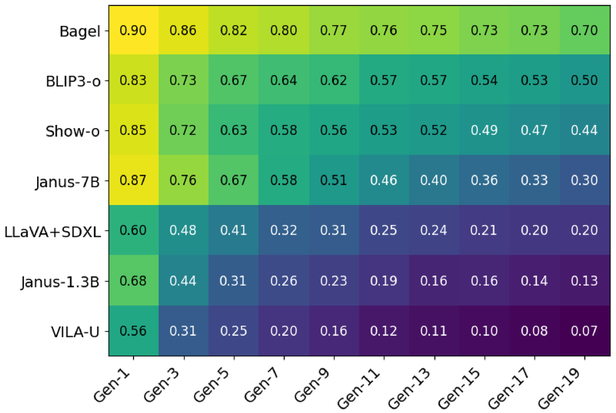
Model	Prompt 1	Prompt 2	Prompt 3
BAGEL	78.14	76.62	78.38
Show-o	57.81	58.03	58.05
VILA-U	19.66	24.65	24.59

- **Prompt 1:** concise caption of the core scene.
- **Prompt 2:** adds counts, relations, and bindings.
- **Prompt 3:** adds attributes, colors, and states.

Values shift slightly; BAGEL stays first, Show-o second, and VILA-U third.

5 MGG Scores Across Generations

Generation quality decays unevenly, revealing long-horizon stability differences.



- First-generation success can overstate long-horizon reliability.
- Repeated self-composition reveals sharply diverging trajectories.

6 MGG Category Persistence

Generation when each property is first lost (↑ lasts longer)

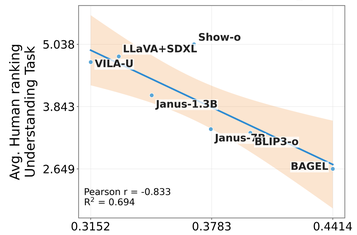
	Single	Two	Count	Pos.	Color	Bind.
BAGEL	20	15	10	11	19	17
VILA-U	10	2	2	2	10	2

- Presence and color persist longest.
- Counts, positions, bindings fail.
- Lost properties rarely return later.

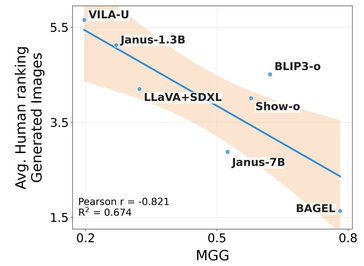
7 Human Alignment and Capability Mismatch

MCD and MGG closely track human rankings across both tasks ($|r| > 0.82$).

MCD vs. human understanding rank



MGG vs. human generation rank



Understanding succeeds more often when generation fails than the reverse.

Outcome	BLIP3-o	Janus-7B	Show-o
Good understanding, poor generation	22.7	13.0	6.5
Poor understanding, good generation	2.2	0.5	4.3

Project page

Paper, code, benchmark, and qualitative results

`mollahsabbir.com/projects/semantic-drift.html`