PRESENCE IS NOT FAITHFULNESS: FIGURATIVE VEHICLE INTRUSION IN TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (TTI) models increasingly generate high-quality images from natural-language prompts, yet figurative language exposes a failure: a vehicle that should guide the depiction of a tenor may instead be rendered as a visible object. We call this failure **Figurative Vehicle Intrusion**: the intruding content is textually licensed, but it is assigned the wrong visual role, showing that visual presence is not always faithfulness and that presence-oriented evaluation can miss such errors. To study it systematically, we introduce **V**ehicle **I**ntrusion and **S**emantic **T**enor **A**ssessment (**VISTA**), a multilingual benchmark of figurative prompts organized by Figurative Form and Mapping Mechanism. We further propose **V-Score**, a diagnostic question-answering metric that evaluates role-aware figurative faithfulness in generated images. Evaluations on recent high-performing TTI models show that vehicle intrusion persists across languages and figurative categories. As a lightweight mitigation, we introduce **VISTA-Guard**, which partially reduces vehicle intrusion and suggests a practical path toward more figuratively faithful TTI generation. All resources will be released publicly.

1 INTRODUCTION

Should Generative Models Draw Every Word in Prompts?

Modern text-to-image (TTI) models, such as GPT-Image (OpenAI, 2026b) and Nano Banana Pro (Raisinghani, 2025), can now generate high-quality images that closely follow natural-language prompts, grounding objects and relations with impressive fidelity. Yet this strength also exposes a subtle failure when language is figurative: models may mistake a descriptive comparison for a drawable entity. As shown in Fig. 1, birds described as "black arrows" may become arrows, while twilight likened to wine may bring wine into the scene. These cases reveal that **visual presence is not necessarily faithfulness**. The model does not fail by omitting prompt content, but by placing it in the **wrong visual role**: a figurative vehicle is rendered as a scene entity rather than used to shape the intended depiction.



Figure 1: Figurative vehicles literalized by models.

This failure matters because visual-generation workflows are increasingly moving beyond manually cleaned prompts and operating directly on source prose. In applications such as illustration, storyboarding, and narrative content creation, users often expect models to turn original passages into visual scenes (Zhou et al., 2024; Shahmohammadi et al., 2023). For example, Fanqie Novel (Novel, 2023) supports illustration generation from selected novel passages, while Douyin's Xiaoyunque (Volcengine, 2026) parses long scripts for short-drama production. Yet narrative prose is not a list of drawable entities; it often uses figurative constructions to convey imagery, affect, and abstraction (Lakoff & Johnson, 2008; Shutova, 2015; Saakyan et al., 2025). In these settings, faithful generation requires deciding not only what is mentioned, but also what should become visually present.

When models make this visibility decision incorrectly, the result is a role assignment error. We define it as **Figurative Vehicle Intrusion**: a textually licensed vehicle that should characterize the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tenor is instead instantiated as an unintended scene entity. Unlike hallucination, the intruding content is supported by the prompt; unlike omission, the model grounds too much text rather than too little. However, this role-level failure has remained largely outside the scope of both TTI generation objectives and evaluation protocols. Recent models and alignment methods often improve image–text compatibility by encouraging generated images to follow detailed captions or preference signals more closely (Saharia et al., 2022; OpenAI, 2023; Hessel et al., 2021; Kirstain et al., 2023; Xu et al., 2023). Compositional benchmarks further reward correct rendering of specified objects, attributes, counts, and relations (Ghosh et al., 2023; Huang et al., 2023; Li et al., 2026), while QA-based protocols ask whether prompt-derived facts are recoverable from the image (Hu et al., 2023; Cho et al., 2024; Lim et al., 2025). These criteria are effective for detecting omissions and binding errors, but they also reinforce a ***presence is faithfulness*** assumption that breaks down for figurative prompts. As illustrated in Fig. 2, in the prompt "*The island was shaped like a fat dragon standing up*," the intended tenor is the island and the vehicle is the dragon, which means that the dragon should only describe the island's shape rather than appear as a literal object. A TIFA-style (Hu et al., 2023) check may nevertheless reward the generated image because the dragon is a prompt-mentioned entity that becomes visible, whereas a semantically faithful image should preserve the island as the scene subject and use the dragon only to guide its shape. This example shows why evaluation must go beyond prompt coverage and test whether a mentioned vehicle is used as semantic guidance or rendered as an object.

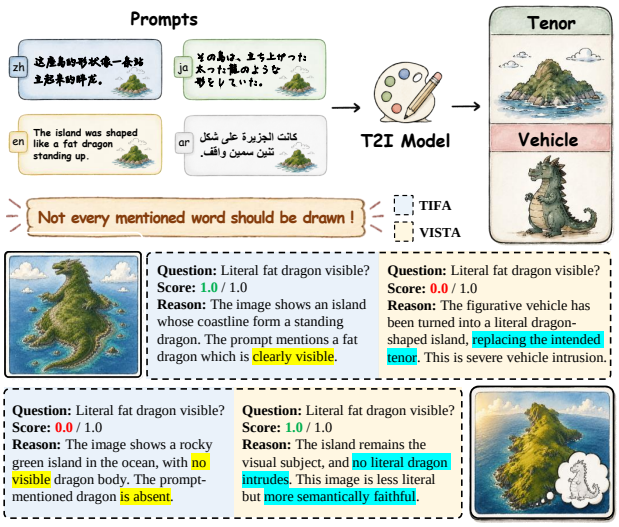


Figure 2: TIFA-style evaluation rewards vehicle intrusion.

It is therefore urgent to more carefully consider what TTI models should be faithful to when prompts contain figurative language: every mentioned word, or the visual scene implied by the sentence. To answer this question systematically, we define the role-level failure, build a diagnostic benchmark and metric, and test a lightweight mitigation strategy. Specifically, we introduce **V**ehicle **I**ntrusion and **S**emantic **T**enor **A**ssessment (**VISTA**), a benchmark composed of figurative prompts. Because figurative language appears across languages and varies in how comparisons are expressed and what motivates tenor–vehicle transfer, VISTA uses multilingual prompts, *Figurative Form* and *Mapping Mechanism* annotations, and explicit tenor–vehicle role labels. To turn annotations into an interpretable evaluation, we further propose **V-Score**, a diagnostic QA-based metric that evaluates tenor preservation, vehicle avoidance, scene coherence, and figurative-meaning preservation, so that a model is not rewarded merely for deleting expressive language. Using VISTA and V-Score, we evaluate five contemporary TTI models and show that vehicle intrusion persists across languages and figurative categories, while comparison with presence-oriented evaluation confirms that conventional prompt-coverage protocols can miss this failure. Finally, as a proof-of-concept mitigation, we introduce **VISTA-Guard**, a skill-based agent (Anthropic, 2026) that reduces vehicle intrusion and improves overall semantic faithfulness. To summarize, our contributions are as follows:

- We identify and formalize **Figurative Vehicle Intrusion**, a role-level grounding failure in figurative prompts that challenges presence-oriented assumptions about visual faithfulness.
- We introduce **VISTA**, a multilingual benchmark organized by figurative form and semantic mapping mechanism, together with **V-Score**, providing diagnostic guidance for evaluating and improving TTI models under figurative prompts.
- We evaluate contemporary TTI models and show that vehicle intrusion persists across languages and figurative categories, while presence-oriented evaluation can miss this failure.
- We propose **VISTA-Guard**, a lightweight skill-based agent that significantly reduces vehicle intrusion in existing TTI models without modifying their underlying generators.

2 Related Work

2.1 Advances in Text-to-Image Generation

Modern TTI research has advanced along two coupled goals: visual fidelity and semantic control. Imagen (Saharia et al., 2022) improves photorealism and language understanding with large pretrained language encoders, while Parti (Yu et al., 2022) studies autoregressive scaling for content-rich prompts. Recent diffusion-transformer systems further improve high-resolution synthesis, text comprehension, and typography (Esser et al., 2024); DALL-E-3 (OpenAI, 2023) uses synthetic re-captioning to follow detailed descriptions more reliably. As image quality has improved, work has increasingly targeted compositional correctness: Structured Diffusion (Feng et al., 2023) reduces missing objects and attribute-binding errors, whereas RPG (Yang et al., 2024) plans region-wise generation for multiple objects and relations. These advances largely treat fidelity as accurately rendering specified content. We instead ask whether a mentioned concept should be rendered at all, since figurative prompts may use it as semantic guidance rather than as a drawable object.

2.2 Text-to-Image Evaluation Benchmarks and Metrics

TTI evaluation has evolved from broad capability coverage to finer-grained diagnosis. Prompt suites such as DrawBench (Saharia et al., 2022) test general prompt following, while holistic frameworks like HEIM (Lee et al., 2023) and Gecko (Wiles et al., 2025) cover alignment, aesthetics, reasoning, multilinguality, and safety-related criteria. Compositional benchmarks (Ghosh et al., 2023; Li et al., 2026) instead ask whether specified objects, attributes, counts, and relations are rendered correctly. Scalar metrics as CLIPScore (Hessel et al., 2021) and PickScore (Kirstain et al., 2023) provide another view by estimating image–text compatibility or human preference. QA-based metrics translate prompts into verifiable visual propositions: TIFA (Hu et al., 2023) evaluates prompt-derived question–answer pairs, DSG (Cho et al., 2024) organizes atomic questions into dependency graphs, and I-HallA (Lim et al., 2025) extends this paradigm to factual image hallucination. These evaluations mainly reward the presence of prompt-derived content, whereas VISTA asks whether each concept should be visually instantiated according to its semantic role.

2.3 Figurative Language in Visual Generation

Figurative language conveys meaning through relations between target and source concepts rather than literal reference alone (Lakoff & Johnson, 2008). Recent work shows that common TTI captions underrepresent subjective and figurative language, limiting the coverage of nonliteral prompting in standard training data (Kundu et al., 2025; Saakyan et al., 2025). Complementary multimodal benchmarks study literal–figurative distinctions, visual entailment, and visual metaphor understanding (Zhang et al., 2021; Yarom et al., 2023; Kleinlein et al., 2022). On the generation side, prior approaches translate nonliteral text into richer visual descriptions (Shahmohammadi et al., 2023; Chakrabarty et al., 2023) or introduce explicit metaphor mappings and binding mechanisms to guide synthesis (Huang et al., 2026). VISTA instead evaluates whether a figurative vehicle is assigned the wrong visual role, and VISTA-Guard provides a lightweight intervention for this role-aware setting.

3 Methodology

3.1 Figurative Vehicle Intrusion

Modern TTI models are often optimized to render prompt-mentioned entities as visible content. However, this presence-oriented assumption does not hold for figurative descriptions such as those illustrated in Fig. 2. Linguistically, a figurative expression relates a ***Tenor*** to a ***Vehicle*** (Lakoff & Johnson, 2008): the tenor refers to the entity that exists in the described scene and should be visualized, whereas the vehicle does not denote an independent scene entity but provides a concept for characterizing the tenor. When a model violates this role distinction by rendering the vehicle as a literal entity, we refer to the resulting error as **Figurative Vehicle Intrusion (FVI)**.

Formally, let $\mathcal{C}_x$ denote the concepts mentioned in prompt x. For any concept c, let $O_x(c)$ indicate that c should exist as an independent visual object in the intended scene, and let $G_I(c)$ indicate

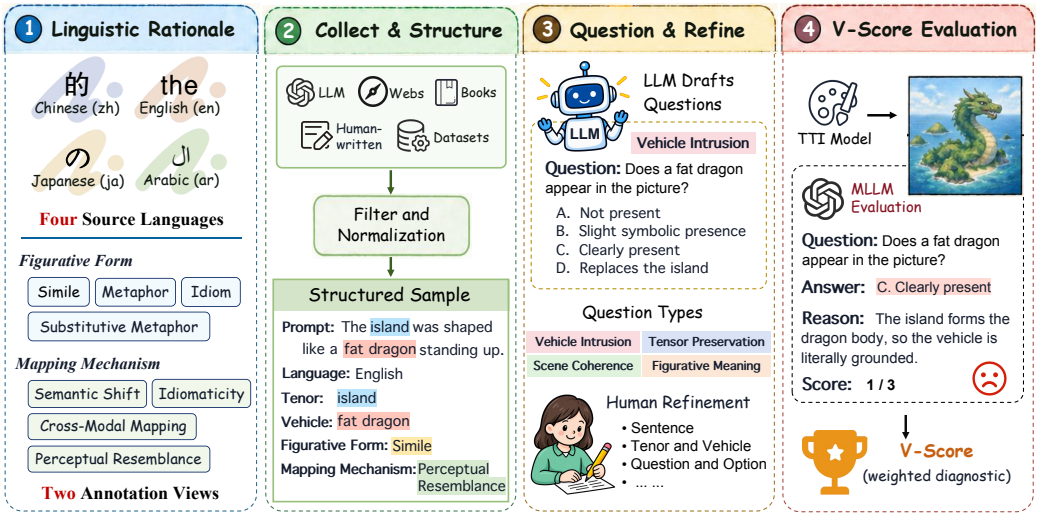


Figure 3: Overview of VISTA Benchmark Construction and Evaluation.

that c is recognizably instantiated in the generated image I. Presence-oriented evaluation typically requires only $\forall c \in \mathcal{C}_x,\ G_I(c)$, whereas FVI specifies the condition:

$$\mathrm{FVI}(x, I) \iff \exists c \in \mathcal{C}_x : \neg O_x(c) \ \wedge \ G_I(c). \tag{1}$$

Thus, FVI occurs when a vehicle concept that should function only as figurative guidance is granted object-level presence in the generated image.

3.2 VISTA: Benchmark for Vehicle Intrusion and Semantic Tenor Assessment

Evaluating FVI requires prompts where figurative vehicles are separated from concepts that should appear as visual objects, a role distinction not targeted by existing TTI benchmarks. To support this evaluation, we introduce VISTA and its QA-based V-Score, as illustrated in Fig. 3.

3.2.1 Dataset Construction

Linguistic Rationale. Figurative language appears across languages (Dryer & Haspelmath, 2013) and is realized through different linguistic forms and tenor–vehicle mappings (Lakoff & Johnson, 2008). Accordingly, VISTA organizes the benchmark by source language and annotates each prompt from two complementary views, as shown in panel 1 of Fig. 3.

For language coverage, we select *Chinese*, *English*, *Japanese*, and *Arabic*, spanning Sinitic, Germanic, Japonic, and Semitic languages and covering typological differences in script, segmentation, morphology, and writing direction (Dryer & Haspelmath, 2013; Hammarström et al.). This selection introduces typological variation while retaining the same core semantic problem (Ploeger et al., 2024), and extends evaluation beyond English-centered prompts (Joshi et al., 2020).

For annotation, VISTA uses two complementary views of the tenor–vehicle relation without reducing figurative language to a single taxonomy (Lakoff & Johnson, 2008). **Figurative Form** follows metaphor-identification and multiword-expression work by recording how the relation is linguistically expressed; its four labels are *Simile*, *Metaphor*, *Substitutive Metaphor*, and *Idiom*. **Mapping Mechanism** records what motivates the transfer from vehicle to tenor; its four labels are *Semantic Shift*, *Perceptual Resemblance*, *Cross-modal Mapping*, and *Idiomaticity*. Detailed definitions and decision boundaries for the eight labels are provided in Appendix A.1.

Data Collection and Construction. We collect figurative candidates from five sources: LLM-assisted generation, websites, books, human-written cases, and existing datasets. As shown in panel 2 of Fig. 3, all candidates are filtered and normalized, retaining only prompts with identifiable

Table 1: VISTA sample counts by language and annotation view.

Annotation	zh	en	ja	ar	Total
Simile	130	220	190	120	660
Metaphor	120	110	230	120	580
Substitutive Metaphor	120	90	20	160	390
Idiom	70	70	60	100	300
Semantic Shift	180	320	270	200	970
Perceptual Resemblance	100	60	110	100	370
Cross-modal Mapping	90	40	60	100	290
Idiomaticity	70	70	60	100	300
Language Total	**440**	**490**	**500**	**500**	**1930**

tenor–vehicle roles and no literal obligation for the vehicle to appear. Each sample records the prompt x_i, language ℓ_i, tenor t_i, vehicle v_i, Figurative Form f_i, and Mapping Mechanism m_i:

$$b_i = (x_i, \ell_i, t_i, v_i, f_i, m_i). \quad (2)$$

Table 1 reports the sample distribution in VISTA, which preserves natural variation from language-specific conventions and source availability.

Question Generation Because FVI concerns both role assignment and overall semantic faithfulness, we generate four multiple-choice questions for each sample along two diagnostic axes, as shown in panel 3 of Fig. 3. The role-level questions assess whether the intended tenor is preserved and whether the vehicle avoids object-level realization, while the semantic-level questions assess scene coherence and preservation of the figurative meaning. Given the prompt and its tenor–vehicle annotations, an LLM drafts each question and its ordered response options in the source language, so that the QA items test semantic-role-aware grounding rather than generic object presence.

Human Refinement To make the generated QA sets reliable as visual evaluation items, human annotators review the original sentence, the tenor–vehicle labels, and the question–option set together. They verify that each question matches the intended diagnostic axis, can be answered from the image, and has mutually exclusive ordered options, revising ambiguous wording and incorrect role assignments. Annotators also check whether the source-language wording preserves the original figurative relation and whether the ordered options reflect meaningful severity levels rather than stylistic variants. The resulting verified QA set is used for V-Score evaluation.

3.2.2 V-SCORE: AN EVALUATION METRIC USING QUESTION-ANSWERING

Given the constructed benchmark items and their human-verified QA sets, we define V-Score to evaluate whether a generated image grounds a figurative prompt according to semantic roles rather than prompt coverage alone. It follows the interpretable QA paradigm used in TTI evaluation (Hu et al., 2023; Li et al., 2024), but adapts it to role-aware figurative faithfulness. As shown in panel 4 of Fig. 3, for each benchmark prompt p, a TTI model produces an image I^p, and an MLLM evaluator selects the option best supported by visible evidence while returning a short rationale.

Each sample contains four ordered multiple-choice questions. Two questions diagnose role assignment: *Tenor Preservation* checks whether the intended subject remains visible, and *Vehicle Avoidance* checks whether the figurative vehicle avoids object-level realization. The other two questions serve as semantic safeguards: *Scene Coherence* checks whether the image remains plausible, and *Figurative Meaning* checks whether the intended figurative interpretation is still conveyed. For each question, options are ordered from the most to the least faithful outcome and mapped to scores in $\{3, 2, 1, 0\}$. To calculate V-Score, we input each generated image with its four questions and answer choices into the MLLM evaluator and map the selected options through their ordered scoring rules; the evaluator template is provided in Appendix A.2.2. Let $\mathcal{P}$ be the set of prompts and I^p the image generated for prompt p. For each p, let $\mathcal{Q}^p = \{(q_k^p, C_k^p, r_k^p)\}_{k=1}^4$ contain its questions, option sets, and scoring rules. The score is defined as:

$$\text{V-Score}(\mathcal{P}) = \frac{100}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{k=1}^{4} \alpha_k \, \rho(\text{VQA}(I^p, q_k^p, C_k^p), r_k^p), \quad \sum_{k=1}^{4} \alpha_k = 1. \quad (3)$$

Here, $\text{VQA}(\cdot)$ is the evaluator's selected option and $\rho(\cdot, r_k^p) \in [0, 1]$ maps it to the normalized score specified by r_k^p. We set $\alpha_1 = \alpha_2 = 0.30$ and $\alpha_3 = \alpha_4 = 0.20$, prioritizing tenor preservation and vehicle-intrusion avoidance while retaining scene plausibility and figurative meaning. Beyond the aggregate score, V-Score preserves the four sub-scores and evaluator rationales, enabling analysis by language, figurative form, mapping mechanism, and error type.

3.3 VISTA-GUARD: A SKILL-BASED AGENT FOR MITIGATING VEHICLE INTRUSION

As a proof-of-concept mitigation for FVI, we design VISTA-Guard as a lightweight skill-based agent. Since LLMs can identify the figurative expression, tenor, and vehicle from prompts, we package this role-explicitation process as a reusable skill (Appendix A.3), following the general paradigm of tool-using and skill-based agents (Schick et al., 2023; Anthropic, 2026). The skill is specified by a description d and workflow w. Given a prompt x, the agent first decides whether to invoke the skill, constructs an auxiliary context only when needed, and then calls the generator:

$$a = R_\phi(x, d) \in \{0, 1\}, \qquad z = \begin{cases} w(x), & a = 1, \\ \varnothing, & a = 0, \end{cases} \qquad I = \mathcal{M}_m(x, z)\,. \tag{4}$$

Here, R_ϕ is a single-pass LLM controller that decides whether the prompt matches the skill description, z is the context produced by the workflow when the skill is invoked, and $\mathcal{M}_m$ denotes inference by the TTI model conditioned on the raw prompt and any auxiliary context. When invoked, the workflow infers the source language, figurative expression, tenor, and vehicle from x alone, then produces a guarded context that preserves the tenor as the visual subject, suppresses the vehicle as a literal entity, and conveys vehicle-derived qualities indirectly through visual attributes. This allows VISTA-Guard to reduce vehicle intrusion with minimal integration cost across generator architectures, without retraining the underlying models.

4 EXPERIMENTS AND ANALYSIS

4.1 EXPERIMENTAL SETUP

We evaluated five text-to-image (TTI) models: Nano Banana Pro (Raisinghani, 2025), GPT-Image-2 (OpenAI, 2026b), DALL-E-2 (Ramesh et al., 2022), DALL-E-3 (OpenAI, 2023), and Grok Imagine (xAI, 2026). To control evaluation cost while retaining manually verified coverage, we select 193 samples from the full VISTA pool by $1/10$ stratified sampling according to the original language and annotation composition (Appendix A.4). For each of the 193 VISTA samples [1], every model generated one image from the original sentence using the same source-language instruction template; the complete templates are provided in Appendix A.2.1. We evaluated each generated image with GPT-5.5 (OpenAI, 2026a) by presenting the four multiple-choice questions in the sample's source language. The evaluator was instructed to rely only on visible evidence and to select the more conservative option when the image was ambiguous. The four questions measure Tenor Preservation (Q1), Vehicle Intrusion (Q2), Scene Coherence (Q3), and Figurative Meaning (Q4); their ordered options are mapped to scores from 3 to 0 and aggregated into V-Score.

4.2 EVALUATION OF TEXT-TO-IMAGE MODELS

Following the V-Score setting defined above, we first evaluate the five TTI models with the default weights $\alpha_1 = \alpha_2 = 0.30$ and $\alpha_3 = \alpha_4 = 0.20$; a weight sensitivity analysis is reported in Appendix A.5. As shown in Fig. 4(a), mean V-Scores occupy a narrow range, from 77.38 for DALL-E-2 to 79.31 for Grok Imagine, but the sample-level distributions remain broad. This indicates that similar aggregate scores mask substantial prompt-level variation rather than uniformly reliable behavior. Table 2 identifies the main source of this instability: more than 70% of outputs receive the highest score on Q1, Q3, and Q4, but only 33.6% do so on Q2, and 57.9% receive 0 or 1. Thus, all evaluated models preserve subjects, scene coherence, and figurative meaning more reliably than they avoid literalizing the vehicle, establishing FVI as a recurrent cross-model failure.

We next stratify V-Score by Language, Figurative Form, and Mapping Mechanism in Fig. 4(b–d). FVI-related errors appear in all four languages, indicating that the failure is not an English-specific artifact of prompt wording. Arabic and Japanese are more difficult for several systems, possibly reflecting uneven multilingual coverage in image–caption data and weaker grounding for less represented scripts or linguistic structures (Li et al., 2023). Across Figurative Forms, idioms are consistently difficult, plausibly because their meanings are non-compositional and cannot be recovered from literal constituents alone (Sag et al., 2002). The Mapping Mechanism view shows a related

[1] Sample #178 may have triggered the DALL-E content policy, leaving 192 valid samples for them.

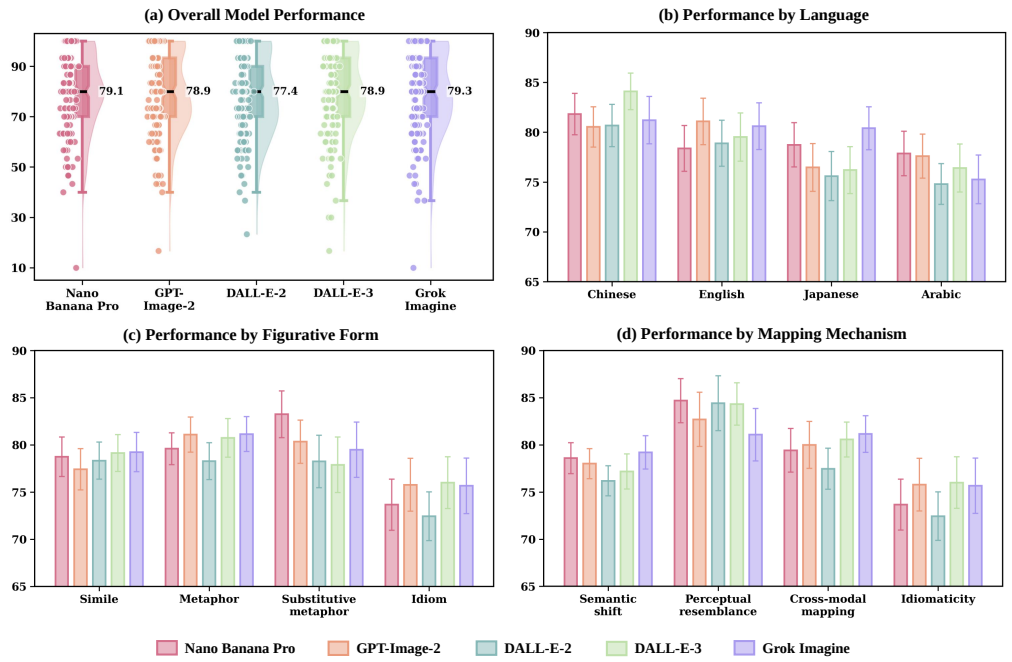


Figure 4: Base performance on VISTA. (a) Sample-level V-Score distributions for each model. (b–d) Mean V-Score with standard errors across Languages, Figurative Forms, and Mapping Mechanisms.

Table 2: Option-score distributions for each model on Base generation. Values are percentages of valid samples assigned to scores 3–0 for each diagnostic question.

Model	Tenor Preservation				Vehicle Avoidance				Scene Coherence				Figurative Meaning			
	3	2	1	0	3	2	1	0	3	2	1	0	3	2	1	0
Nano Banana Pro	72.5	23.3	4.1	0.0	31.6	11.4	41.5	15.5	75.1	19.7	4.7	0.5	79.3	18.1	2.1	0.5
GPT-Image-2	76.7	19.2	4.1	0.0	33.2	5.7	40.4	20.7	71.5	22.8	4.1	1.6	86.0	11.4	2.1	0.5
DALL-E-2	69.3	26.6	4.2	0.0	32.3	6.8	37.5	23.4	69.8	22.9	7.3	0.0	81.8	15.6	2.1	0.5
DALL-E-3	72.4	22.4	5.2	0.0	35.4	7.3	37.0	20.3	72.9	18.8	7.3	1.0	84.4	13.5	2.1	0.0
Grok Imagine	77.2	18.7	3.6	0.5	35.8	10.9	35.8	17.6	69.9	22.8	5.7	1.6	77.7	17.6	4.1	0.5

pattern: perceptual resemblance is relatively easier, likely because it offers visible correspondences such as shape or color (Perniss et al., 2010), whereas idiomaticity depends more on lexicalized conventional knowledge. Together, these results indicate that FVI is not confined to a particular language or surface form, but varies with how explicit, conventionalized, and visually recoverable the tenor–vehicle relation is. To reduce potential evaluator bias from GPT-5.5, we further assess cross-evaluator robustness on a held-out subset in Appendix A.6.

4.3 Comparison with Presence-Oriented Evaluation

To determine whether conventional prompt-faithfulness evaluation exposes FVI, we compared V-Score with a TIFA-like protocol on 40 prompts sampled from the common valid set of all five models. The subset contains ten prompts per language and approximately preserves the distributions of Figurative Form and Mapping Mechanism. Following TIFA (Hu et al., 2023), GPT-5.5

Table 3: TIFA-like and V-Score evaluation results.

Model	TIFA-like		V-Score	
	Score	Rank	Score	Rank
Nano Banana Pro	93.50	4	77.17	2
GPT-Image-2	**98.50**	**1**	76.25	4
DALL-E-2	94.50	2	73.83	5
DALL-E-3	94.50	2	76.50	3
Grok Imagine	92.00	5	**78.50**	**1**

generated five source-language multiple-choice questions from each raw prompt, covering entities, attributes, relations, actions, and scene properties without access to tenor–vehicle annotations. Table 3 first compares the model-level results under the two protocols. TIFA-like scores concentrate between 92.0 and 98.5, yet the resulting rankings differ markedly from those produced by V-Score. GPT-Image-2 ranks first under TIFA-like but fourth under V-Score, whereas Grok Imagine ranks fifth under TIFA-like but first under V-Score. Thus, even when evaluated on the same prompts and images, the two protocols favor different generation behavior. We further quantified their agreement at the image level (Appendix Fig. 8). Of the 200 images, 168 (84%) receive a TIFA-like score of 100, while their V-Scores remain widely distributed. Correspondingly, the association between the two metrics is negligible (Spearman's $\rho = -0.042$, $N = 200$). This disagreement follows from their evaluation criteria: TIFA-like questions treat prompt-mentioned concepts as target content, so a literally rendered vehicle can satisfy the generated QA despite occupying the wrong semantic role. Presence-based faithfulness and V-Score therefore capture complementary properties: the former measures whether described content is recoverable, whereas the latter determines whether that content has been grounded appropriately.

4.4 Effectiveness of VISTA-Guard

To mitigate FVI, we evaluated VISTA-Guard across all five generators by pairing each guarded output with the Base image from the same prompt. As shown in Fig. 5, VISTA-Guard consistently shifts the score distributions upward, with mean paired gains of 5.2–9.9 points. The remaining overlap and lower-score tails indicate that it reduces, rather than eliminates, difficult failures.

We then compared VISTA-Guard with two prompt controls to determine whether its gains could be explained by prompt elaboration alone. Human-Interact (Human) mimics an ordinary user with a minimal request to visualize the sentence, whereas Generic Anti-Hallucination (Generic Anti-Hall) adds general constraints against unsupported content without identifying tenor–vehicle roles. As shown in Table 4, Human-Interact remains close to Base, while Generic Anti-Hallucination improves the overall score by only 1.3 points on average and produces almost no change in Q2. In contrast, VISTA-Guard improves the overall score and Q2 by 7.90 and 1.08 points on average, respectively. These results suggest that explicit tenor–vehicle role assignment is a useful and low-cost direction for mitigating FVI, beyond generic caution against hallucination. Prompt templates for all strategies are provided in Appendix A.2.1.

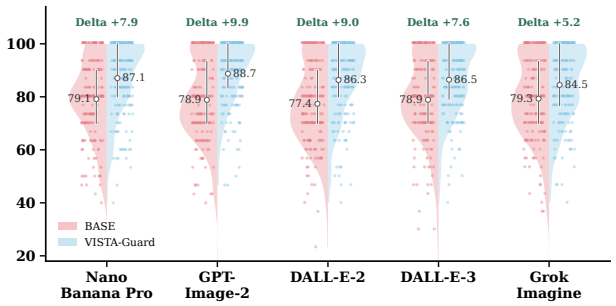


Figure 5: Sample-level V-Score distributions for Base and VISTA-Guard across five generators.

Table 4: V-Score under four prompting strategies. Δ Score and Δ Q2 denote changes in the overall score and vehicle-avoidance score relative to Base, respectively.

Model	Human	Base	Generic Anti-Hall			VISTA-Guard		
			Score	Δ Score	Δ Q2	Score	Δ Score	Δ Q2
Nano Banana Pro	78.2	79.1	79.5	+0.42	-0.13	**87.1**	**+7.94**	+1.08
GPT-Image-2	77.3	78.9	80.2	+1.33	+0.21	**88.7**	**+9.86**	+1.18
DALL-E-2	78.2	77.4	79.7	+2.37	-0.09	**86.3**	**+8.96**	+1.25
DALL-E-3	78.8	78.9	79.4	+0.54	-0.27	**86.5**	**+7.57**	+1.13
Grok Imagine	79.9	79.3	81.2	+1.90	+0.06	**84.5**	**+5.18**	+0.76
Mean	78.5	78.7	80.0	+1.31	-0.04	**86.6**	**+7.90**	+1.08

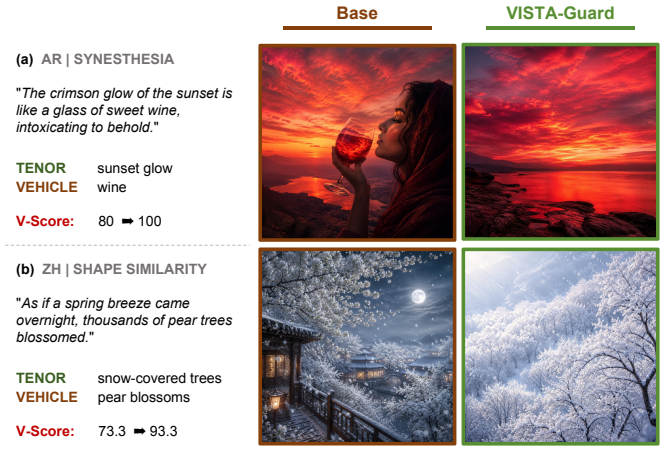


Figure 6: Two representative cases comparing Base and VISTA-Guard. Prompts are translated into English.

To make these changes more concrete, Fig. 6 provides qualitative examples of the same effect. In the Arabic synesthetic simile, the intended tenor is the sunset and the vehicle is wine: Base preserves the reddish atmosphere but turns the vehicle into a foreground glass, shifting attention away from the sky. VISTA-Guard instead keeps the sunset as the scene subject and expresses the wine-like quality through crimson color and glow, raising V-Score from 80.0 to 100.0. In the Chinese shape-based metaphor, the intended tenor is snow-covered scenery and the vehicle is pear blossoms. Base literalizes the blossoms on the branches, whereas VISTA-Guard renders white snow-laden trees without adding flowers, increasing V-Score from 73.3 to 93.3.

4.5 Reliability of V-Score through Human Evaluation

To examine whether V-Score reflects human judgment, we recruited 12 participants to independently evaluate 30 blind images using the same four question dimensions and option-to-score mapping. Each image received 12 complete ratings, and human reference scores were obtained by averaging the participant scores before comparison with the MLLM evaluator. The human ratings show strong internal reliability, with an absolute-agreement ICC of 0.784 for single raters and 0.978 for the 12-rater average. The results are shown in Fig. 7, where each point compares the MLLM-derived V-Score with the mean human score and horizontal error bars indicate uncertainty across raters. The marginal density curves further show that the two score distributions occupy similar ranges, rather than matching only a few isolated examples. The points closely follow the diagonal trend from low- to high-scoring images, and the MLLM-derived V-Score correlates strongly with mean human judgment (Spearman's $\rho = 0.939$, MAE = 6.3). The results support V-Score as a proxy for human assessment of FVI-sensitive image faithfulness.

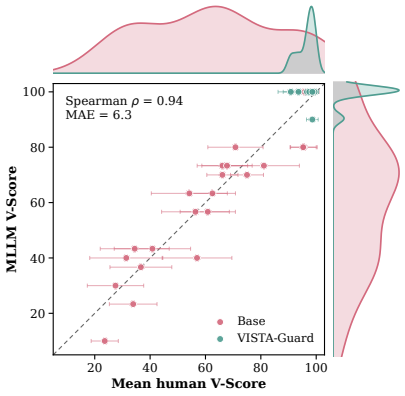


Figure 7: Human–MLLM agreement on 30 blind images.

5 Conclusion

This work shows that figurative prompts require TTI models to follow visual roles, not simply every mentioned word. We identify figurative vehicle intrusion, a role-level grounding failure in which a textually supported vehicle is rendered as an unintended scene entity. To make this failure measurable, VISTA and V-Score organize figurative prompts by language, Figurative Form, and Mapping Mechanism, and evaluate whether images preserve the intended tenor while avoiding literal vehicle intrusion. Experiments show that FVI persists in recent high-performing TTI models, is missed by presence-oriented evaluation, and aligns well with human judgment under our QA-based protocol. VISTA-Guard further shows that making tenor–vehicle roles explicit before generation can reduce this failure. Together, these resources provide a benchmark and practical direction for TTI models faithful not only to prompt words, but to their intended visual roles.

AI USE STATEMENT

In this work, we used generative AI tools to assist with synthetic benchmark item generation, dataset cleaning and reformatting, MLLM-based evaluation, literature search and summarization, and manuscript language polishing. AI-assisted data and evaluation outputs were manually reviewed by the authors, including checks of tenor–vehicle annotations, question–answer sets, scoring outputs, and cited literature. Generative AI tools were not used to replace author judgment in defining the research problem, validating the experimental conclusions, or determining the final claims. The authors take full responsibility for the final content of this submission, including all text, data, evaluations, and artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. Extend claude with skills. Claude Code Docs, 2026. URL https://code.claude.com/docs/en/skills. Accessed 26 August 2026.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7370–7388, 2023.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, volume 2024, pp. 15625–15645, 2024.

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, et al. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841):20200390, 2021.

Matthew S. Dryer and Martin Haspelmath (eds.). *WALS Online (v2020.4)*. Zenodo, 2013. doi: 10.5281/zenodo.13950591. URL https://doi.org/10.5281/zenodo.13950591.

Patricia L Duffy and Julia Simner. Synaesthesia in fiction. *cortex*, 46(2):277–278, 2010.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PUIqjT4rzq7.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Pragglejaz Group. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39, 2007.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. glottolog/glottolog: Glottolog database 5.3. Zenodo. URL http://glottolog.org.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 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.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 20349–20360. IEEE, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Qingbao Huang, Cheng Yang, Jiawei Yao, Zhiyue Liu, Yi Cai, and Xingmao Zhang. Cmig: Conceptual metaphor theory-inspired framework for metaphorical image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23748–23761, 2026.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6282–6293, 2020.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Ricardo Kleinlein, Cristina Luna-Jiménez, and Fernando Fernández-Martínez. Language does more than describe: On the lack of figurative speech in text-to-image models. *arXiv preprint arXiv:2210.10578*, 2022.

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. Looking beyond the pixels: Evaluating visual metaphor understanding in vlms. In *EMNLP (Findings)*, pp. 23137–23158, 2025.

George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 2008.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5290–5301. IEEE, 2024.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *International Conference on Learning Representations*, volume 2026, pp. 86729–86758, 2026.

Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulić, and Anna Korhonen. Translation-enhanced multilingual text-to-image generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9174–9193, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.510. URL https://aclanthology.org/2023.acl-long.510/.

Youngsun Lim, Hojun Choi, and Hyunjung Shim. Evaluating image hallucination in text-to-image generation with question-answering. In *Association for the Advancement of Artificial Intelligence*, 2025.

Fanqie Novel. Ai illustration feature user guide. Official Fanqie Novel Creator Center, 2023. URL https://fanqienovel.com/writer/zone/article/7270037871984115770. Accessed 26 August 2026.

OpenAI. Dall·e 3 system card. OpenAI, 2023. URL https://openai.com/index/dall-e-3-system-card/. Accessed 19 September 2026.

OpenAI. Gpt-5.5 instant: smarter, clearer, and more personalized. OpenAI, 2026a. URL https://openai.com/index/gpt-5-5-instant/. Accessed 19 September 2026.

OpenAI. Introducing chatgpt images 2.0. OpenAI, 2026b. URL https://openai.com/zh-Hans-CN/index/introducing-chatgpt-images-2-0/. Accessed 26 August 2026.

Pamela Perniss, Robin L Thompson, and Gabriella Vigliocco. Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in psychology*, 1:227, 2010.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. What is "typological diversity" in nlp? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5681–5700, 2024.

Naina Raisinghani. Introducing Nano Banana Pro. Google Blog, 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/. Accessed 26 August 2026.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 10.48550. *arXiv preprint arXiv.2204.06125*, 2022.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1–23, 2025.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pp. 1–15. Springer, 2002.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Hassan Shahmohammadi, Adhiraj Ghosh, and Hendrik Lensch. Vipe: Visualise pretty-much everything. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5477–5494, 2023.

Ekaterina Shutova. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623, 2015.

Volcengine. Xiaoyunque–intelligent video generation agent 2.0–product introduction. Jimeng AI Documentation, 2026. URL https://docs.volcengine.com/docs/85621/2371410?lang=zh. Accessed 26 August 2026.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, et al. Revisiting text-to-image evaluation with gecko: on metrics, prompts, and human rating. In *International Conference on Learning Representations*, volume 2025, pp. 272–287, 2025.

xAI. Grok imagine image. xAI Developer Documentation, 2026. URL https://docs.x.ai/developers/models/grok-imagine-image. Accessed 19 September 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56704–56721. PMLR, 2024. URL https://proceedings.mlr.press/v235/yang24ai.html.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36:1601–1619, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3214–3225, 2021.

Wei Zhang and Judith Gelernter. Exploring metaphorical senses and word representations for identifying metonyms. *arXiv preprint arXiv:1508.04515*, 2015.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

A Appendix

A.1 Detailed VISTA Annotation Taxonomy

VISTA assigns each sample one label under *Figurative Form* and one under *Mapping Mechanism*. These views are complementary rather than hierarchical: the former describes how the tenor–vehicle relation is linguistically expressed, whereas the latter describes what makes the transfer interpretable (Lakoff & Johnson, 2008; Shutova, 2015; Group, 2007). The main text lists the category names for readability; this appendix provides the definitions and decision boundaries used during annotation.

A.1.1 Language Coverage

The selected languages cover distinct families, scripts, segmentation conventions, and writing directions. This design introduces typological variation while keeping the semantic problem of tenor–vehicle role assignment comparable across languages.

Language	Family	Writing system	Segmentation	Direction
Chinese	Sinitic	Logographic	No whitespace word segmentation	left-to-right
English	Germanic	Alphabetic	Whitespace-delimited words	left-to-right
Japanese	Japonic	Mixed scripts	No whitespace word segmentation	left-to-right
Arabic	Semitic	Abjad	Morphologically rich	right-to-left

A.1.2 Figurative Form

Figurative Form describes how the tenor–vehicle relation is expressed. The labels distinguish explicit comparison, implicit attribution, referential replacement, and lexicalized expression.

Label	Core distinction	Decision boundary
Simile	Explicit comparison	Tenor and vehicle are linked by a marker equivalent to "like" or "as."
Metaphor	Implicit attribution	Vehicle characterizes an explicit or recoverable tenor without an overt comparison marker.
Substitutive Metaphor	Referential replacement	Vehicle expression stands in for, or partially replaces, the tenor, which must be recovered from context (Zhang & Gelernter, 2015).
Idiom	Lexicalized form	Conventionalized multiword expression conveys a non-compositional meaning (Sag et al., 2002).

A.1.3 Mapping Mechanism

Mapping Mechanism describes what makes the transfer from vehicle to tenor interpretable. The labels distinguish conceptual movement, perceptual similarity, cross-modal transfer, and conventionalized meaning.

Label	Core distinction	Decision boundary
Semantic Shift	Conceptual transfer	Interpretation depends on abstraction, concretization, or related conceptual movement rather than perceptual similarity.
Perceptual Resemblance	Shared perceptual property	Mapping is licensed by visible or audible similarity, such as shape or sound (Perniss et al., 2010; Ćwiek et al., 2021).
Cross-modal Mapping	Sensory transfer	A property associated with one modality characterizes an experience in another, as in synesthetic expressions (Duffy & Simner, 2010).
Idiomaticity	Conventional meaning	Interpretation relies on lexicalized, non-compositional knowledge rather than a productive conceptual or perceptual mapping (Sag et al., 2002).

A.2 Prompts

This section collects the prompt templates used for image generation and evaluation.

A.2.1 Generation Prompts

We use source-language-matched generation templates for all TTI models. The English template below is shown for readability; Chinese, Japanese, and Arabic samples use corresponding translations. The placeholder `[CONTENT]` is replaced by the corresponding VISTA sentence.

Human-Interact Prompt

Please generate an image based on the following sentence:

`[CONTENT]`

Base Generation Prompt

Generate exactly one image that directly illustrates the following sentence.

Sentence: `[CONTENT]`

Use the sentence's language context, cultural meaning, and figurative meaning as visual guidance. Make the image a coherent visual scene rather than a diagram or poster. Return an image only.

Do not explain, analyze, translate, summarize, or answer in text. Do not include any readable text, subtitles, captions, labels, signs, handwriting, letters, numbers, logos, or watermarks anywhere in the image.

Generic Anti-Hallucination Prompt

Generate exactly one image that directly illustrates the following sentence.

Sentence: `[CONTENT]`

Use the sentence's language context, cultural meaning, and figurative meaning as visual guidance. Make the image a coherent visual scene rather than a diagram or poster. Return an image only. Do not explain, analyze, translate, summarize, or answer in text. Do not include any readable text, subtitles, captions, labels, signs, handwriting, letters, numbers, logos, or watermarks anywhere in the image.

General anti-hallucination constraints:
Only depict elements that are visually necessary for a coherent interpretation of the sentence. Do not add unrelated concrete objects, characters, props, symbols, decorations, or background elements. Avoid over-literalizing figurative or poetic language into extra objects unless those objects are required by the scene itself. If a phrase is ambiguous or nonliteral, prefer a natural visual scene that preserves the intended mood and meaning without inventing additional entities. Keep the composition focused, realistic in internal logic, and visually self-consistent.

A.2.2 Evaluator Prompt

We use a fixed MLLM evaluator prompt that asks the model to inspect only visible image evidence, answer the four multiple-choice questions, and return structured option selections with rationales. The following English template is shown for readability; source-language-matched translations are used for Chinese, Japanese, and Arabic samples.

MLLM Evaluator Prompt

You are a strict visual evaluator for a benchmark.

Your task is NOT to creatively interpret the image. Inspect only what is visible in the image and answer the multiple-choice questions. Do not assume something is present just because it appears in a question. If visual evidence is ambiguous, choose the more conservative option.
For questions about whether a literal comparison object appears, count recognizable real objects, icons, written marks, or symbolic motifs as visible; do not count only mood, color, lighting, or composition as the object itself.

Evaluation steps:
1. Read the four questions.
2. Inspect the image carefully.
3. For each question, compare all four options and choose the best supported answer.
4. Return only minified JSON with A/B/C/D answers.

Questions: `[QUESTIONS]`

Return ONLY minified JSON in this format:

```
{
  "answers":    {"Q1": "A", "Q2": "B", "Q3": "A", "Q4": "C"},
  "reasoning":  {"Q1": "...", "Q2": "...",
                 "Q3": "...", "Q4": "..."}
}
```

A.3 VISTA-Guard Skill Specification

This section provides the reusable VISTA-Guard skill specification used in our implementation.

VISTA-Guard.md

Name: VISTA-Guard.
Description: Identify figurative prompts whose vehicle should guide the depiction of a tenor without becoming an independent literal object; infer the figurative expression, tenor, and vehicle; and construct a guarded generation context that suppresses literal vehicle intrusion while preserving the intended figurative effect.

Use when: A raw image-generation prompt contains a figurative expression in which the vehicle should influence the tenor's visual qualities but should not appear as a separate scene entity.

Input: The original image-generation prompt.

Workflow:

1. Inspect the raw prompt and infer its source language.
2. Identify whether the prompt contains a figurative expression relevant to visual grounding.
3. If invoked, infer the tenor and vehicle from the prompt alone.
4. Preserve the tenor as the main visual subject.
5. Suppress the vehicle as an independent literal entity.
6. Convey vehicle-derived qualities indirectly through visual attributes such as shape, color, lighting, texture, composition, posture, atmosphere, or emotional tone.
7. Compose a model-ready guarded context for image generation.

Output: A guarded generation context that preserves the original prompt, states the inferred tenor and vehicle roles, discourages literal vehicle realization, and specifies how the intended figurative qualities should be conveyed indirectly.

Do not use when: The prompt is fully literal, both compared entities are intended to appear, or the literal coexistence of tenor and vehicle is an intentional visual pun.

A.4 Evaluation Subset Composition

The main experiments use a 193-sample subset selected from the 1,930-sample VISTA pool by $1/10$ stratified sampling. Table 5 reports the resulting composition, which preserves the language and annotation-view coverage of the full benchmark.

Table 5: Composition of the 193-sample evaluation subset.

Annotation	**zh**	**en**	**ja**	**ar**	**Total**
Simile	13	22	19	12	66
Metaphor	12	11	23	12	58
Substitutive Metaphor	12	9	2	16	39
Idiom	7	7	6	10	30
Semantic Shift	18	32	27	20	97
Perceptual Resemblance	10	6	11	10	37
Cross-modal Mapping	9	4	6	10	29
Idiomaticity	7	7	6	10	30
Language Total	**44**	**49**	**50**	**50**	**193**

A.5 V-Score Weight Sensitivity

To test whether the main findings depend on the default V-Score weights, we recompute scores under several aggregation schemes: role-focused scoring using Q1+Q2, semantic-safeguard scoring using Q3+Q4, equal weighting over all four questions, and the default weighting used in the main paper. As shown in Table 6, equal-weight scores remain close to the default V-Score results, and the two variants are strongly correlated at the image level (Spearman's $\rho = 0.988$). The Q3+Q4 scores are

consistently higher and resemble conventional evaluation criteria that emphasize scene plausibility and semantic recoverability, confirming that such criteria can obscure vehicle-intrusion failures.

Table 6: V-Score sensitivity to different aggregation schemes on Base generation.

Model	Q1+Q2	Q3+Q4	Equal	Default
Nano Banana Pro	71.24	90.93	81.09	79.12
GPT-Image-2	70.64	91.19	80.92	78.86
DALL-E-2	68.84	90.19	79.51	77.38
DALL-E-3	70.83	90.97	80.90	78.89
Grok Imagine	72.88	88.95	80.92	79.31

A.6 Cross-Evaluator Robustness

To examine whether V-Score is sensitive to the choice of MLLM evaluator, we additionally evaluate the same 40-prompt, 200-image subset used in the metric-comparison experiment with `qwen3.8-max`. As shown in Table 7, the second evaluator produces slightly lower absolute scores, but largely preserves the ranking pattern and the conclusion that FVI persists across generators.

Table 7: Cross-evaluator robustness of V-Score on the 40-prompt subset.

Model	Main Evaluator		Qwen Evaluator	
	Score	Rank	Score	Rank
Nano Banana Pro	77.17	2	75.72	2
GPT-Image-2	76.25	4	74.32	3
DALL-E-2	73.83	5	71.15	5
DALL-E-3	76.50	3	74.10	4
Grok Imagine	78.50	1	76.35	1

A.7 Supplementary Metric Comparison

Figure 8 provides the comparison between the TIFA-like protocol and V-Score. Although many images receive saturated TIFA-like scores, their V-Scores remain broadly distributed, illustrating that prompt-coverage evaluation and role-aware figurative evaluation capture different failure modes.

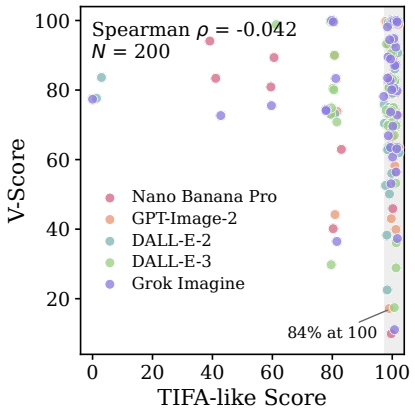


Figure 8: Image-level relationship between TIFA-like score and V-Score on 200 images. Points are slightly jittered for visibility.