DEPICT: SCORING TEXT-TO-IMAGE ALIGNMENT BY ANSWER AGREEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-text alignment is a core problem in computer vision with applications in caption evaluation, hallucination detection, data curation, and the benchmarking of text-to-image (T2I) generators. As T2I models improve, benchmarking has become demanding, requiring metrics capable of finding a series of issues like missing objects, swapped attributes, miscounts, and ignored negations. Recent work addresses this by fine-tuning evaluators on preference data or by prompting a vision-language model, either holistically with the caption or with decomposed verification questions. However, existing approaches fall short: fine-tuned metrics remain bound to one backbone and training distribution; holistic metrics miss fine-grained details; and decomposed metrics rely on a fixed-YES assumption that penalizes faithful images whenever that assumption fails. In contrast, we propose DEPICT, a training-free metric that replaces fixed reference answers with expected agreement between image-based and caption-only answers, weighting questions by how decisively the caption determines them. By replacing fixed references, our agreement rule increases negation accuracy from 19% to 88%. To recover the context lost during decomposition, DEPICT merges this agreement score with a holistic score. We evaluate DEPICT on five benchmarks and eleven backbones from three model families and find that it surpasses all training-free metrics and exceeds fine-tuned evaluators on two out of three human-correlation benchmarks.[1]

1 INTRODUCTION

Image-text alignment evaluation is fundamental across vision-language tasks, including caption scoring (Hessel et al., 2021), vision-language models (VLM) hallucination detection (Kogilathota et al., 2026), data filtering (Schuhmann et al., 2021; Gadre et al., 2023), and, increasingly, evaluating text-to-image (T2I) generators (Kamath et al., 2025; Ramos et al., 2026) and their reward signals (Xu et al., 2023; Black et al., 2024). As T2I models improve, the question has shifted from whether they can generate an image that matches a prompt to whether they can render every aspect of it faithfully, without dropping an object, swapping an attribute, miscounting, or ignoring a negation (Labs et al., 2025; Chefer et al., 2026; Zhao et al., 2026). An alignment metric must therefore correlate with human judgments on open-ended text, including these fine-grained errors (Ku et al., 2024).

Two families of evaluators have emerged. Fine-tuned metrics learn a scoring function from preference data (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023; Marjit et al., 2026), while training-free ones prompt an off-the-shelf VLM (Lin et al., 2024; Hu et al., 2023; Cho et al., 2024). We focus on the training-free family due to a fundamental structural advantage: while fine-tuned evaluators remain bound to a specific architecture and the static distribution of their training data, training-free metrics naturally scale alongside foundation models, inheriting backbone advancements without task-specific re-training. To evaluate how effectively these metrics leverage underlying backbone improvements, we conduct what is, to our knowledge, the first study assessing them across eleven backbones from three model families and five benchmarks.

Within the training-free family, two approaches have developed separately. *Holistic* scoring started with global embedding similarity between separately encoded text and image (Hessel et al., 2021) and newer approaches ask VLMs a single question about the whole caption and read the probability

[1] Code will be available for research purposes.

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.

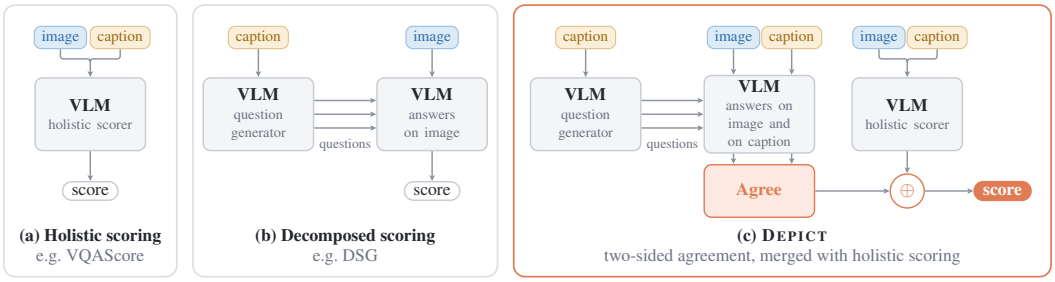


Figure 1: **Text-to-image scoring paradigms.** (a) *Holistic*: directly scores the image-caption pair. (b) *Decomposed*: answers caption-generated questions on the image. (c) DEPICT: checks agreement between image- and caption-based answers, then merges it with a holistic score.

of YES (Lin et al. (2024); Figure 1a). *Decomposed* scoring methods generate a set of verification questions from the caption, answer each against the image, and aggregate the results to obtain a final score (Hu et al. (2023); Cho et al. (2024); Kamath et al. (2025); Figure 1b). Holistic scoring preserves macro context and decomposed metrics offer fine-grained interpretability, but neither approach dominates across benchmarks. Decomposed metrics also share a flaw in how each answer is scored. The original TIFA scheme graded each question by string-matching the VLM's answer against a label derived from the caption, such as objects, counts, or actions (Hu et al., 2023). Later methods simplified this by assuming a constant YES target, scored either by hard string matches (Cho et al., 2024) or their probabilities (Kamath et al., 2025). We call this the *fixed-*YES *assumption*: the expected answer is fixed before looking at the image, so whenever it is wrong, the metric fails with it.

We propose DEPICT, a training-free alignment metric (Figure 1c). Given a caption, DEPICT decomposes the description into a set of visual assertions without requiring ground-truth labels. A VLM evaluates each assertion twice, once against the image and once against the caption alone, assigning a verdict to each. The score reflects the agreement between the two channels, which is high whenever both favor the same truth value, regardless of polarity. We further find that the holistic and decomposed paradigms are complementary rather than competing and DEPICT is the first metric to merge them. Across every human-judgment benchmark we test, the merged score correlates better with human ratings than either approach alone.

Our contributions are:

- **Diagnosis of the fixed-YES flaw.** We uncover a fundamental structural flaw in current decomposed metrics, the fixed-YES assumption, which causes their accuracy on negated prompts to drop below chance.
- **Agreement-based scoring.** We score each question by the agreement between a text-only and a visual answer, which removes the assumption and improves both negation accuracy and correlation with human judgements.
- **Systematic benchmark study.** We present, to our knowledge, the first comparison of training-free T2I alignment metrics across backbones from several model families and multiple benchmarks, with paired confidence intervals for every difference.
- **DEPICT.** We merge agreement-based decomposed scoring with holistic scoring in a training-free metric. It exceeds the strongest fine-tuned evaluator on its own backbone on most human-judgment benchmarks and consistently matches or beats training-free baselines across backbones and benchmarks.

2 RELATED WORK

Embedding-based metrics. Early alignment metrics like CLIPScore (Hessel et al., 2021) and BLIP2Score (Li et al., 2023) rely on global image-text similarity. Because these models are pre-trained for coarse caption matching rather than fine-grained relational reasoning, they behave as bag-of-words matchers (Yüksekgönül et al., 2023; Koishigarina et al., 2026) that struggle with compositionality and correlate weakly with human judgments (Hu et al., 2023; Cho et al., 2024).

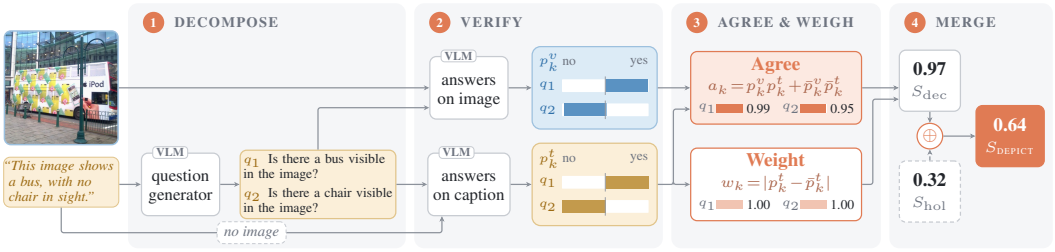


Figure 2: Overview of DEPICT. **(1)** A VLM converts the caption into yes/no questions q_k. **(2)** A frozen VLM answers each question on the image (p_k^v) and on the caption alone (p_k^t). **(3)** Questions are scored by answer agreement a_k (rewarding matching *no* answers) and weighted by commitment w_k. **(4)** The weighted decomposed score S_{dec} is averaged with holistic score S_{hol}.

Fine-tuned metrics. The initial approaches take the encoders above and align their outputs with human preference data (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023), while newer evaluators start from a VLM and learn from human ratings or from a larger teacher model (Tu et al., 2025; Wang et al., 2026; Marjit et al., 2026). In both cases, the learned score is bound to its backbone and its training data. DynEval (Marjit et al., 2026), the strongest of these, fine-tunes a decomposition pipeline on Qwen3-VL (Yang et al., 2025) and serves as our matched-backbone reference.

Training-free metrics. These metrics query an off-the-shelf VLM (Yang et al., 2025; Team et al., 2026), so any stronger backbone can be swapped in at no cost. Early approaches decomposed the caption into a question per assertion (Yarom et al., 2023; Singh & Zheng, 2023) and checked the image's answer against a reference (a caption-derived label in TIFA (Hu et al., 2023) or a fixed YES in DSG (Cho et al., 2024)). Soft scoring then replaced exact matching. VQAScore (Lin et al., 2024) evaluated the entire caption by reading the probability of a single YES, capturing overall context but remaining blind to fine-grained details. Meanwhile, Gecko (Wiles et al., 2025) and Soft-TIFA (Kamath et al., 2025) read per-question answer probabilities, making decomposed scores continuous while retaining the fixed reference. DEPICT replaces this static reference with the agreement between a text and a visual answer and is the first to merge the decomposed and holistic approaches.

3 MEASURING ALIGNMENT BETWEEN TEXT AND VISION

We present DEPICT, a training-free metric that scores an image I against a caption t. Rather than scoring each question against a fixed answer as prior decomposed metrics do (Hu et al., 2023; Cho et al., 2024), DEPICT treats the answer $y \in \{\text{YES}, \text{NO}\}$ as a latent variable. It generates questions directly from the caption, answers each from the image and from the caption alone, and scores a question by the probability that the two answers coincide, marginalized over both outcomes. Questions are then weighted by how decisively the caption settles them, and the resulting decomposed score is merged with a holistic one (Figure 2).

Generating questions from the caption. DSG (Cho et al., 2024) first parses t into semantic tuples and then converts each tuple into a question. We skip the intermediate parsing step by prompting a language model to produce N yes/no questions $\{q_k\}_{k=1}^N$ directly from t , each targeting a single verifiable detail and answerable from the image alone. Unlike DSG, the generator sees the whole caption, so questions do not inherit possible parsing errors and retain the context that tuple extraction discards, making post-hoc filtering and a dependency graph unnecessary. We ablate this in Section 5.5.

Continuous soft scoring. Decomposed metrics such as TIFA (Hu et al., 2023) and DSG (Cho et al., 2024) score each question q_k against a reference y_k^* using a hard indicator function $a_k^{\text{hard}} = \mathbb{I}(\text{VQA}(I, q_k) = y_k^*)$. Following soft approaches (Lin et al., 2024; Kamath et al., 2025), we read the softmax probability of the tokens, $p_k^v = p(\text{YES} \mid I, q_k)$, rather than decoding a discrete answer. Probabilities are not renormalized over the YES/NO pair, so $p(\text{YES}) + p(\text{NO}) \leq 1$ and mass placed elsewhere lowers the score, penalizing off-target outputs without forcing a binary choice.

From a fixed reference to expected agreement. Most decomposed evaluators score under a fixed-reference assumption, setting $y_k^* = \text{YES}$ for every question (Cho et al., 2024; Kamath et al., 2025). We relax this assumption progressively with the following scoring rules a_k:

$$\text{Faith:} \quad a_k = p_k^v \tag{1}$$

$$\text{Confirm:} \quad a_k = p_k^v \, p_k^t \tag{2}$$

$$\text{Agree:} \quad a_k = p_k^v \, p_k^t + \bar{p}_k^v \, \bar{p}_k^t. \tag{3}$$

Current soft metrics (Lin et al., 2024; Kamath et al., 2025) rely on *Faith*, measuring only the visual channel's probability of answering YES as in equation 1. *Confirm* replaces the fixed reference with a soft text channel, $p_k^t = p(\text{YES} \mid t, q_k)$, in which the model answers the same question from the caption instead of the image (Eq. 2). This is the continuous version of a text-derived reference y_k^* where the visual score is weighted by how strongly the caption entails YES (Hu et al., 2023). However, because *Confirm* only rewards positive agreement, it fails whenever the correct answer is NO. *Agree* resolves this by incorporating NO probabilities ($\bar{p}_k^v, \bar{p}_k^t$) and marginalizing over the latent answer as $a_k = \sum_y p(y \mid t, q_k)\, p(y \mid I, q_k) = p_k^t p_k^v + \bar{p}_k^t \bar{p}_k^v$ (Eq. 3). It is symmetric across modalities and generalizes prior rules, reducing to hard discrete matching for one-hot outputs, to *Faith* when p_k^t=1, and to *Confirm* when omitting the NO term.

Commitment Weighting. The text channel's answers p_k^t and $\bar{p}_k^t$ show how confidently the model can answer q_k from the caption alone. When they are close, the model is uncertain and the question says little about alignment, whether because it is ambiguous or off-target. We therefore weight each question by the text channel's commitment:

$$w_k = \left| p_k^t - \bar{p}_k^t \right|, \tag{4}$$

so that questions answered decisively with the caption context dominate the score while uncertain ones vanish. This serves as a soft quality filter, eliminating the need for post-hoc pruning (we ablate this mechanism in Section 5.5).

Decomposed Score Aggregation. The final decomposed alignment score is the commitment-weighted mean of the per-question scores a_k across all N questions:

$$S_{\text{dec}}(I, t) = \frac{\sum_{k=1}^{N} w_k \, a_k}{\sum_{k=1}^{N} w_k + \epsilon}, \tag{5}$$

where a_k is the per-question score of Eqs. 1–3 and ϵ prevents division by zero.

Holistic Answering and Merge. Decomposed questions capture fine-grained visual compositional logic but occasionally lose global context, whereas holistic evaluation captures macro-level visual alignment but misses subtle compositional failures. Following VQAScore (Lin et al., 2024), we compute a holistic alignment score $S_{\text{hol}}(I, t)$ by querying the VLM with the entire caption t directly:

$$S_{\text{hol}}(I, t) = p(\text{YES} \mid I, t). \tag{6}$$

The DEPICT score is the convex combination of the decomposed $S_{\text{dec}}(I, t)$ and the holistic $S_{\text{hol}}(I, t)$:

$$S_{\text{DEPICT}}(I, t) = (1 - \lambda)\, S_{\text{dec}}(I, t) + \lambda \, S_{\text{hol}}(I, t), \tag{7}$$

where $\lambda \in [0, 1]$ balances fine-grained compositional detail with macro-level relational context.

4 EXPERIMENTAL SETUP

Datasets, metrics, and statistics. To measure the correlation with human judgements, we use GenAI-Bench (Li et al., 2024), TIFA160 (Hu et al., 2023), and RichHF-18K (Liang et al., 2024) *misalignment_score*, reporting Spearman Rank Correlation Coefficient (SRCC) on raw scores and Pearson Linear Correlation Coefficient (PLCC) following Han et al. (2024). For diagnostics, we use Winoground (Thrush et al., 2022), reporting text, image, and group accuracy, and the COCO-MCQ split of NegBench (Alhamoud et al., 2025). Score differences (Δ) between metrics is paired, with BCa 95% intervals over $B = 10{,}000$ bootstrap replicates resampled by item. Dataset details are presented in Appendix A.1.

Baselines. We compare against three groups of metrics. *1) Training-free, embedding-similarity* metrics score the cosine similarity of caption and image embeddings from a contrastive encoder: CLIPScore (Hessel et al., 2021) and BLIPv2Score (Li et al., 2023). *2) Training-free, VLM-based* metrics prompt an off-the-shelf VLM and run on any backbone; these are the focus of our analysis, and we take one representative per cell of the decomposition $\times$ granularity grid of Section 5.2: DSG (Cho et al., 2024) (decomposed, discrete), Soft-TIFA (Kamath et al., 2025) (decomposed, continuous; the Faith rule of Eq. 1), and VQAScore (Lin et al., 2024) (holistic, continuous). *3) Fine-tuned evaluators* learn a scoring function from preference data or teacher annotations and are tied to their training backbone: ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), UniGenBench++ (Wang et al., 2026), LongT2IBench (Yang et al., 2026), T2I-Eval (Tu et al., 2025), FGA-BLIP2 (Han et al., 2024), and DynEval (Marjit et al., 2026).

Backbones. To decouple the scoring procedure from visual reasoning capability, all training-free methods are run on eleven backbones varying along four axes: 1) *family*, via Qwen3.5 (Yang et al., 2025), InternVL3.5 (Wang et al., 2025), and Gemma-4 (Team et al., 2026); *2) scale*, via two size ladders (Qwen3.5 4B/9B/27B, InternVL3.5 4B/8B/14B); *3) generation*, via Qwen3.5/3.6/3.8 at a fixed 27B; and *4) architecture*, via the Gemma-4 variants (per-layer embedding E4B, unified multimodal 12B, dense 31B). Qwen3-VL-4B provides a matched comparison against DynEval.

Implementation Details. Question generation and scoring use the same backbone, so every row of Tables 1 and 2 uses a single model; prompts are in Appendix A.2. The generator has no question budget and produces as many questions as the caption requires, so N varies per caption (statistics in Appendix A.1). For each question we obtain the visual answer p^v from the image-question pair and the textual answer p^t from the caption-question pair with the image removed; answer probabilities are full-vocabulary softmax masses summed over the variants of each answer token (e.g. `yes`, `Yes`, `␣yes`). We use commitment weighting, no pruning, and arithmetic aggregation, with $\epsilon = 10^{-3}$ in equation 5 and $\lambda = 0.5$ in equation 7 fixed a priori and not tuned on any benchmark (sensitivity in Appendix B.4). All training-free methods see identical images, prompts, and decoding settings: VQAScore uses its original prompt, DSG runs its published pipeline with only the VLM swapped, and Soft-TIFA-AM, which specifies only a scoring rule (an unweighted mean of p^v against a fixed YES), is applied to our question sets, where it coincides with the unweighted Faith row of Table 3. Computational pass counts and latency analysis are detailed in Appendix A.3.

5 Results and Discussion

We evaluate DEPICT against fine-tuned and training-free baselines across human alignment, compositional discrimination, and explicit negation. Across five benchmarks and eleven backbones, we show that expected agreement and term merging match or exceed fine-tuned evaluators, hold across backbones, and avoid failures of fixed-YES metrics. Qualitative evaluation is provided in Appendix E.

5.1 Does DEPICT compete with fine-tuned evaluators?

Table 1 compares DEPICT with fine-tuned evaluators, and with training-free metrics on both their original and current backbones, including Qwen3-VL-4B, the backbone of the strongest fine-tuned evaluator, which allows a matched comparison.

Competitive with fine-tuned evaluators. DEPICT is best or second best across all datasets and metrics of Table 1. When evaluated on the same Qwen3-VL-4B backbone as DynEval-4B, DEPICT surpasses the strongest fine-tuned baseline on GenAI-Bench and RichHF, trailing only on TIFA160. Furthermore, on this shared backbone, DEPICT matches or exceeds VQAScore (the second-best training-free approach) in SRCC across all three benchmarks while strictly outperforming it in PLCC.

Training-free metrics track the backbone. Swapping VQAScore's 2024 backbone for Qwen3-VL-4B, with no other change, raises its SRCC by 0.075 on GenAI-Bench and 0.124 on RichHF, enough to surpass every fine-tuned evaluator in the table on those two benchmarks, including one trained on a 72B model. Moving DEPICT from Qwen3-VL-4B to Qwen3.5-27B adds a further 0.06–0.10 in every column and yields the best score on GenAI-Bench and RichHF.

Method	Backbone	GenAI-Bench		TIFA160		RichHF	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Fine-tuned evaluators							
ImageReward†	BLIP	0.340	0.379	0.621	0.634	0.275	0.329
PickScore†	CLIP-H	0.354	0.363	0.428	0.434	0.392	0.413
UniGenBench++†	Qwen2.5-VL-72B	0.461	0.444	0.608	0.590	0.533	0.517
LongT2IBench†	Qwen2.5-VL-7B	0.375	0.376	0.678	0.692	0.517	0.516
T2I-Eval†	MiniCPM-V	0.500	0.477	0.601	0.625	0.503	0.499
FGA-BLIP2 (os)†	BLIP-2	0.564	0.567	0.760	0.744	0.512	0.546
DynEval-4B†	Qwen3-VL-4B	0.595	0.582	**0.802**	**0.803**	0.546	0.569
Training-free, embedding similarity							
CLIPScore†	CLIP	0.168	0.203	0.300	0.309	0.057	0.302
BLIPv2Score†	BLIP-2	0.273	0.298	0.429	0.454	0.143	0.310
Training-free, VLM-based, original backbone							
VQAScore	CLIP-FlanT5	0.553	0.518	0.695	0.659	0.483	0.409
Soft-TIFA-AM†	[Qwen3-VL-8B]	0.482	0.457	0.682	0.673	0.631	0.602
Training-free, VLM-based, current backbone							
DSG	Qwen3-VL-4B	0.437	0.441	0.610	0.613	0.489	0.472
Soft-TIFA-AM	Qwen3-VL-4B	0.418	0.412	0.667	0.666	0.528	0.534
VQAScore	Qwen3-VL-4B	0.628	0.594	0.658	0.591	0.607	0.548
DEPICT	Qwen3-VL-4B	0.630	0.627	0.691	0.675	0.607	0.597
DEPICT	Qwen3.5-27B	**0.720**	**0.725**	0.780	0.767	**0.667**	**0.664**

Table 1: **Human correlation** on GenAI-Bench, TIFA160, and RichHF-18K. **Bold** marks best value per column and underline second best. †: cited from original publication under matching protocol; others re-run. Qwen3-VL-4B matches top baseline backbone; final row uses our largest backbone.

5.2 WHERE DO THE GAINS COME FROM?

To separate procedure from backbone capability, Table 2 fixes the backbone at Qwen3.5-27B, so training-free methods differ only in how they score. The rows map onto the components of Section 3: DSG uses its own tuple-based questions with hard scoring, Soft-TIFA-AM scores our questions with *Faith*, VQAScore is the holistic term, and DEPICT's decomposed term scores the same questions with *Agree* before being merged with holistic term.

The baselines split by benchmark. Comparing training-free baselines on a shared backbone reveals neither paradigm dominates. Soft-TIFA leads on TIFA160, whereas VQAScore leads on GenAI-Bench and RichHF (Table 2). TIFA160 favors decomposition in general, with decomposed metrics scoring 0.09-0.22 higher on TIFA160 than on GenAI-Bench (against 0.04 for VQAScore) which is a pattern consistent across every decomposed evaluator in Table 1, fine-tuned or not.

The decomposed term closes the gap to holistic scoring. Switching the scoring rule from *Faith* (Soft-TIFA-AM) to *Agree* narrows the gap to VQAScore on GenAI-Bench from 0.15 to just 0.03 SRCC (Table 2). This shows that expected agreement alone drives the improvement, while also raising NegBench accuracy from below chance to 88.3% without hurting performance on TIFA160.

Merging takes the best of both. On all three correlation benchmarks DEPICT matches whichever term suits the benchmark and improves on it (Table 2). The gain is significant on GenAI-Bench and positive but within the interval on TIFA160 and RichHF. The improvement comes from the low end: the merged metric's lowest-scored items are rated worse by humans than VQAScore's on every benchmark and backbone, while the top tails coincide (Appendix D.3).

The merge carries no penalty where one term suffices. On Winoground's short compositional captions, the decomposed term adds little beyond what the holistic pass captures, keeping DEPICT within the confidence interval of VQAScore across all three accuracies (Table 2). Furthermore, soft

Method	Human corr. (SRCC)			Winoground (acc. %)			NegBench
	GenAI-B	TIFA160	RichHF	Text	Image	Group	acc. %
DSG	0.549	0.673	0.594	52.0	46.5	44.5	34.7
Soft-TIFA-AM (*Faith*)	0.546	0.762	0.636	75.8	71.8	64.8	19.1
VQAScore (*Holistic*)	0.698	0.737	0.656	**83.2**	**82.8**	**75.5**	84.7
Agree	0.671	0.763	0.612	74.0	74.0	65.8	**88.3**
DEPICT	**0.720**	**0.780**	**0.667**	82.2	**82.8**	74.5	86.9
Δ vs. best baseline	+.022	+.018	+.011	−1.0	+0.0	−1.0	+2.3
[95% CI]	[+.016, +.028]	[−.002, +.040]	[−.013, +.031]	[−3.5, +1.0]	[−3.0, +2.5]	[−3.8, +1.3]	[+1.8, +2.8]

Table 2: **Fixed-backbone comparison on Qwen3.5-27B.** All methods share images, prompts, and decoding. *Agree* differs from *Faith* only in scoring rule; DEPICT combines it with holistic. Δ gives the difference of DEPICT over the strongest baseline in each column, with its paired BCa 95% CI.

logit scoring across both terms (*Agree*) drives performance far more than decomposition itself, leaving discrete one-hot approaches like DSG lagging by 20 points.

5.3 WHY DO AGREEMENT AND MERGING WORK?

In this section, we analyze how *Agreement* overcomes the structural flaws of fixed-YES rules on negation, while merging recovers performance on contested samples where terms disagree.

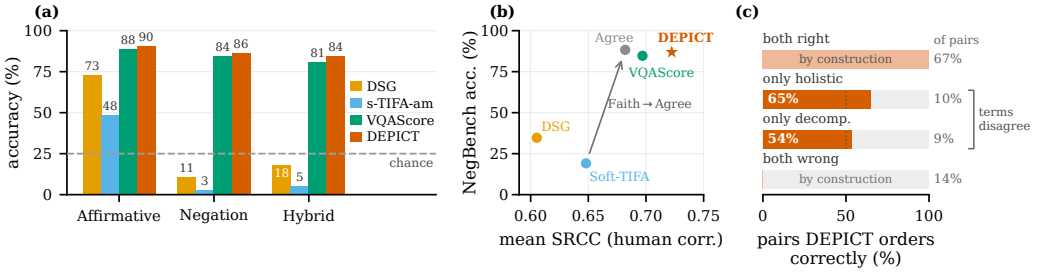
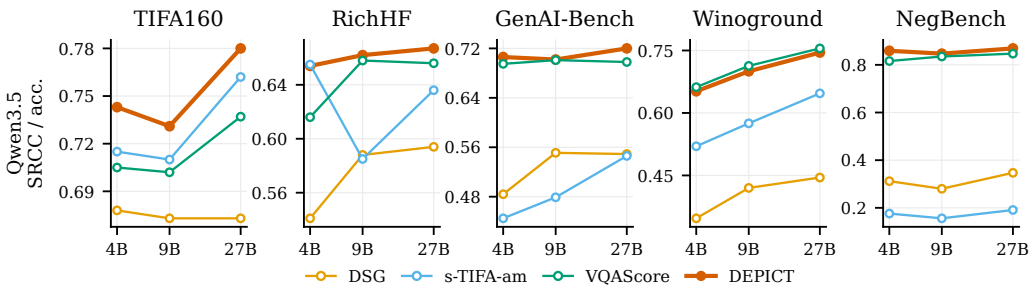


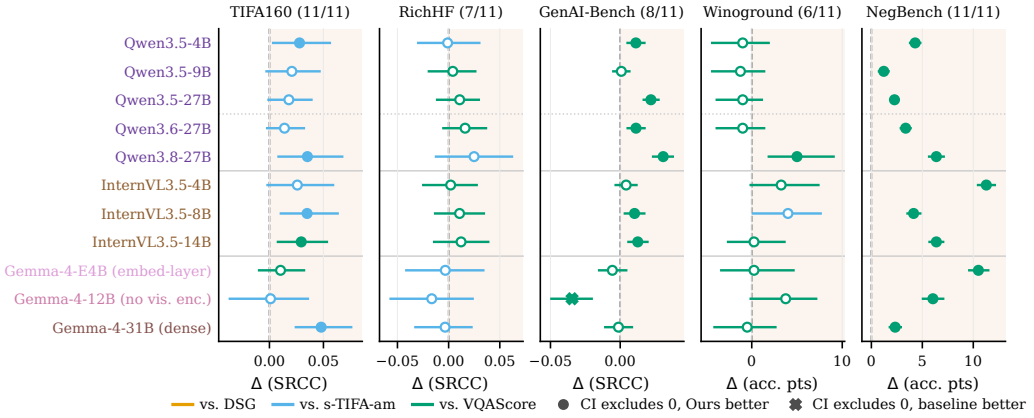
Figure 3: **Why agreement and merging work (Qwen3.5-27B). (a)** NegBench accuracy by template; VQAScore is unaffected because its single question contains the negation. **(b)** NegBench accuracy vs. mean SRCC; *Agree* improves on *Faith*, and DEPICT boosts correlation further. **(c)** GenAI-Bench item pairs grouped by term alignment.

The fixed-YES assumption fails on negation. By reading the expected answer from the caption, DEPICT is the only decomposed metric accurate across all three templates (Fig. 3a), while DSG and Soft-TIFA lose 85–95% of their accuracy on negated items and fall below chance. This occurs because 87% of questions for negated captions and 47% for hybrid ones expect NO vs 4% elsewhere (Appendix D.1), so a fixed-YES rule gives a low score to every correct NO before looking at the image, penalizing faithful images (Appendix B.2-B.3; text reference reliability in Appendix D.2).

The terms are complementary. *Agreement* alone lifts both NegBench accuracy and human correlation over *Faith* but trails VQAScore on correlation; merging both terms places DEPICT ahead on both (Fig. 3b). The merge of both terms (Eq. 7) works because the terms fail on different samples: on GenAI-Bench they disagree on 19% of item pairs. Fig. 3c shows that DEPICT tracks the results where both terms agree and it follows whichever term separates the two images more strongly where they disagree. DEPICT keeps 65% of the pairs only the holistic term orders correctly and 54% of those only the decomposed term does. Whereas individual terms fail on each other's exclusive wins, DEPICT achieves 60% accuracy across contested pairs, easily outperforming the holistic (53%) and decomposed (47%) terms alone. The merge relies on agreement: combining VQAScore with Soft-TIFA instead leaves GenAI-Bench unchanged and lowers NegBench to 62.2, below VQAScore alone, since the fixed-YES term drags down the holistic score on negated items (Appendix C.3).



(a) Absolute scores along the Qwen3.5 size ladder.



(b) Difference to the strongest baseline on eleven backbones.

Figure 4: **Training-free metrics across backbones.** Evaluated on five benchmarks (SRCC for correlation benchmarks; accuracy on Winoground and NegBench). (a) Other families in Appendix B.1. (b) Score differences (Δ) between DEPICT and the strongest baseline per backbone (paired bootstrap 95% CI). Values > 0 favor DEPICT; solid markers exclude zero, and crosses indicate significant losses. Colors denote the baseline and titles wins count.

5.4 How do metrics behave across backbones?

The backbone of a training-free metric is a free choice, so we ask whether DEPICT's advantage depends on it. Using the eleven-backbone sweep described in Sec. 4, Fig. 4a traces Qwen3.5 size scaling, while Fig. 4b reports baseline differences across family, architecture, and generation axes.

Size helps where the backbone is the bottleneck. Along the Qwen3.5 ladder, DEPICT leads or ties at every size on all three correlation benchmarks, where larger backbones mainly narrow the gap for the weaker decomposed baselines. On NegBench its advantage does not depend on size, as accuracy holds at 85–87% at every size while fixed-YES baselines stay at or below chance, showing that how the score is computed matters more than backbone size. In contrast, on Winoground all methods gain similarly, suggesting that the backbone is the bottleneck.

The advantage is consistent across backbones. Across all eleven backbones (Fig. 4b), DEPICT matches or beats the strongest baseline in 54 out of 55 backbone–benchmark pairs, winning in 43 cases with 23 being statistically significant. Across families, its NegBench gains are larger on InternVL3.5 than on Qwen3.5 (4–11 points against 1–4), even though InternVL3.5 is far weaker on Winoground at matched size, so agreement scoring compensates for a weaker backbone. Across architectures, DEPICT's only loss occurs on GenAI-Bench with the Gemma-4-12B variant without a vision encoder, the weakest backbone on every benchmark (Appendix B.1). Across generations, its margins are largest on the newest backbone, Qwen3.8-27B, significant on four out of five benchmarks, even though that backbone scores below its predecessors for nearly every method.

Comp.	Setting	Wino.	GenAI	Neg.
Questions	via tuples	44.8	0.573	64.4
	no w_k	65.0	0.668	**88.3**
Filter	pruning, no w_k	62.7	0.642	88.0
	pruning + w_k	63.2	0.644	88.0
Default		**65.8**	**0.671**	**88.3**

(a) Pipeline components.

Rule	Wino.	GenAI	Neg.
Faith, hard‡	48.5	0.513	21.9
Faith†	64.8	0.546	19.1
Confirm	64.0	0.624	66.4
Agree	65.0	0.668	88.3
+ w_k (default)	**65.8**	**0.671**	**88.3**

(b) Scoring rule on identical questions.

Table 3: **Component ablations of the decomposed term (Qwen3.5-27B).** Wino.: Winoground group accuracy; GenAI: GenAI-Bench SRCC; Neg.: NegBench accuracy; higher is better. (a) Default: questions generated directly from the caption, *Agree* scoring, commitment weighting w_k equation 4; each row changes one component. (b) Rules without w_k, then the default. †Identical to Soft-TIFA-AM. ‡DSG's hard fixed-YES rule on our questions, without its dependency graph.

5.5 WHICH COMPONENTS MATTER?

Table 3 ablates the design choices of the decomposed term on Qwen3.5-27B, replacing one component at a time while holding the rest at the final configuration and testing multiple scoring-rules. We ablate reasoning in the answering step in Appendix C.4.

Tuple Extraction. Replicating DSG's tuple extraction hurts all benchmarks (Table 3a, top), 47% of negated NegBench captions yield no tuples and Negation-subset accuracy falls to 4.1, six times below chance, showing it cannot express absence. On Winoground pairs, which share entities but differ in structure, it generates overlapping questions (Jaccard 0.44 vs. 0.27 without tuples; 13% identical), shrinking the paired score margin by ∼40% and costing 21 group-accuracy points (Appendix C.1). Generating questions directly from the caption preserves both.

Question pruning. Post-hoc pruning lowers performance on every benchmark. Earlier pipelines filtered questions to remove noise (Cho et al., 2024). Agreement scoring handles this per question: a question the two channels cannot settle scores near the midpoint rather than being rewarded or silenced, and w_k down-weights it (Appendix C.2), so pruning only removes questions carrying signal.

Scoring. The scoring rule is the largest single factor (Table 3b), discretizing Faith as DSG does costs 16.3 Winoground points and 0.033 SRCC, while Faith stays below chance on NegBench. Agree raises it 4.6× and GenAI-Bench SRCC by 23%, moving Winoground by at most one point.

6 CONCLUSIONS

Fine-tuned, holistic, and decomposed metrics each capture distinct part of caption-image alignment: fine-tuning aligns with human judgments but ties the evaluator to one backbone; holistic VQA preserves global context but misses fine-grained errors; and standard decomposition uncovers compositional logic but relies on a fixed-YES target that breaks on negated captions. DEPICT resolves these trade-offs without training by evaluating two-sided answer agreement weighted by caption commitment, merged with a holistic VQA pass. Agreement lifts NegBench accuracy from 19% to 88%, and the merged score matches or exceeds top evaluators across five benchmarks and eleven backbones. On DynEval's matched Qwen3-VL-4B backbone, DEPICT beats DynEval-4B on GenAI-Bench (0.630 vs. 0.595 SRCC) and RichHF (0.607 vs. 0.546), though fine-tuning retains a lead on TIFA160 (0.802 vs. 0.691).

Limitations remain: decomposition brings no gain on short prompts that hinge on compositional word order, such as Winoground, although enabling reasoning on answers improves results there; and because both channels share one backbone, a prior shared by both can make the same hallucination look like agreement. Answering each channel with a different backbone would decouple these priors, at the cost of hosting two models.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, create or modify scientific figures or images, create or edit software code, and edit a research paper to improve readability.

We have not used generative AI tools to design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, or interpret results. Furthermore, generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, and clean and reformat dataset are not applicable to this work.

We have reviewed all AI-assisted work. LLM-generated code was verified and tested during the process of development, process of modifying the figures was overseen by the authors to ensure changes matched the intended output and text edits were reviewed by multiple authors to guarantee improved readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

While DEPICT adds no generative capability itself, improved alignment metrics could indirectly assist downstream pipelines that produce deceptive imagery, either by measuring similarity during inference or by using its scores as a signal for improving models.

REPRODUCIBILITY STATEMENT

DEPICT is training-free: it uses no fine-tuning, no learned parameters, and no reference images, so reproducing it reduces to fixing the model, the prompts, and the read-out, each of which we specify. The metric (the holistic and decomposed arms and their equal-weight combination) is defined in Section 3. The four prompts used for every reported number (question generation, image answering, text-only answering, and the holistic template) are reproduced in Appendix A.2 and shipped with the code (code will be available for research purposes). Models are served with vLLM using greedy decoding (temperature 0) with reasoning disabled. Each benchmark, its split (including the COCO-MCQ split of NegBench), and its processing are described in Appendix A.1.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29612–29622. IEEE, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Jaemin Cho, Yushi Hu, Jason M. 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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ITq4ZRUT4a.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search

of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=dVaWCDMBof.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. EvalMuse-40K: A Reliable and Fine-Grained Benchmark with Comprehensive Human Annotations for Text-to-Image Generation Model Evaluation, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 20349–20360. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01866. URL https://doi.org/10.1109/ICCV51070.2023.01866.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing Benchmark Drift in Text-to-Image Evaluation, 2025.

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. In *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2305.01569.

Sai Akhil Kogilathota, Sripadha Vallabha EG, Luzhe Sun, and Jiawei Zhou. Halp: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085, 2026.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DldwXCCP25.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *International Conference on Learning Representations*, volume 2024, pp. 46689–46722, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. GenAI-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop @ CVPR 2024*, 2024. URL https://openreview.net/forum?id=hJm7qnW3ym.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX*, volume 15067 of *Lecture Notes in Computer Science*, pp. 366–384.

Springer, 2024. doi: 10.1007/978-3-031-72673-6_20. URL https://doi.org/10.1007/978-3-031-72673-6_20.

Shyam Marjit, Dheeraj Baiju, Anuj Shikarkhane, Akhil Sakthieswaran, Sayak Paul, and Anirban Chakraborty. Dyneval: Holistic evaluations of t2i generative models in the wild. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 417–435, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37010-5.

Vasco Ramos, Regev Cohen, Idan Szpektor, and João Magalhães. Early estimation of language to latent alignment in diffusion models. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part XX*, volume 17020 of *Lecture Notes in Computer Science*, pp. 625–643. Springer, 2026. doi: 10.1007/978-3-032-37553-7_36. URL https://doi.org/10.1007/978-3-032-37553-7_36.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *Advances in Neural Information Processing Systems*, 36: 70799–70811, 2023.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022.

Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22340–22361, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1088. URL https://aclanthology.org/2025.acl-long.1088/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. UniGenBench++: A Unified Semantic Evaluation Benchmark for Text-to-Image Generation, 2026.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajic, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. Revisiting text-to-image evaluation with gecko: on metrics, prompts, and human rating. In *ICLR*. OpenReview.net, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis, 2023.

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. In *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2304.05977.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhichao Yang, Tianjiao Gu, Jianjie Wang, Feiyu Lin, Xiangfei Sheng, Pengfei Chen, and Leida Li. Longt2ibench: A benchmark for evaluating long text-to-image generation with graph-structured annotations. In *AAAI*, pp. 11820–11828. AAAI Press, 2026.

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. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/056e8e9c8ca9929cb6cf198952bf1dbb-Abstract-Conference.html.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

A EXPERIMENTAL DETAILS

A.1 BENCHMARKS

Table 4 summarizes the five benchmarks. GenAI-Bench (Li et al., 2024) pairs compositional prompts with images from several generators, each rated for alignment on a 1-5 scale. TIFA160 (Hu et al., 2023) contains 160 prompts with five generated images each, rated on the same scale. For RichHF-18K (Liang et al., 2024) we use the test split and its per-image misalignment score. On all three we report SRCC on raw scores and PLCC after a four-parameter logistic fit, pooled over image-caption pairs.

Benchmark	Source	Evaluated units	Human signal	Metric	$\bar{N}$ (max)
Correlation with human ratings					
GenAI-Bench	GenAI-Bench prompts and generated images	9,600 pairs, 1,600 prompts	1–5 alignment	SRCC, PLCC	5.47 (13)
TIFA160	TIFA160 prompts, 5 images each	800 pairs, 160 prompts	1–5 alignment	SRCC, PLCC	4.89 (11)
RichHF	RichHF-18K test split	955 pairs, 493 captions	misalignment score	SRCC, PLCC	5.52 (32)
Accuracy					
Winoground	Winoground	400 examples, 800 matched pairs	correct pairing	text / image / group acc.	4.16 (10)
NegBench	COCO-MCQ split (COCO val2017)	5,914 images $\times$ 4 captions	correct caption	4-way MCQ acc.	1.92 (6)

Table 4: **Benchmarks.** PLCC is computed after a four-parameter logistic fit. NegBench captions follow three templates (affirmative, negation, hybrid); each image has one correct caption and three distractors. $\bar{N}$: mean number of generated questions per unique caption, with the maximum in parentheses.

Winoground (Thrush et al., 2022) contains 400 examples of two images and two captions that use the same words in a different order; a metric is correct on the text task if it scores each image higher with its own caption, on the image task if it scores each caption higher with its own image, and on the group task if both hold. NegBench (Alhamoud et al., 2025) (COCO-MCQ split) pairs each of 5,914 COCO val2017 images with four candidate captions built from three templates: affirmative (*"a kitchen with a table"*), negation (*"a kitchen with no people"*), and hybrid, which combines both. One caption is correct and three are distractors; we score each image–caption pair independently and predict the highest-scoring caption, so chance is 25%.

Question counts. Pooled over benchmarks, the generator produces a mean of 2.71 questions per unique caption (median 2). Descriptive captions yield about five on average, from 4.16 on Winoground to 5.52 on RichHF (Table 4), and the count grows with caption length, up to 32 for the longest RichHF caption. NegBench's templated captions each state one or two claims and yield 1.92 on average; a few yield none, the case the ϵ in Eq. equation 5 guards against.

A.2 PROMPTS

DEPICT uses three prompts of its own, one for question generation (Fig. 7) and one for each answering channel (Fig. 5- 6). The holistic term uses VQAScore's original prompt (Lin et al., 2024), `Does this figure show "{caption}"? Please answer Yes or No.`, and reads S_{hol} as the probability of YES. All prompts run on the same backbone, and answer probabilities are read from the token distribution as described in Section 4. The `focus`, `difficulty` and `used_tuples` fields are logged for analysis and not used in scoring.

VISUAL ANSWER PROMPT

```
You are answering a yes/no question about an image.

Rules:
1) Use only the visual evidence in the image.
2) Answer "yes" only if the image clearly satisfies the question; answer
"no" if it clearly does not.

Question: {question}
Answer:
```

Figure 5: **Visual answer prompt**, given the image and one question; produces p_k^v.

TEXTUAL ANSWER PROMPT

```
You are answering a yes/no question about what a faithful image of a
text-to-image generation prompt should contain.

The generation prompt is the specification: it describes what a correct
image must depict. Answer what that correct image should show, not just
what the prompt literally spells out.

Rules:
1) Treat the prompt as the specification of a faithful image. For
questions about the existence, completeness, count, or spatial relation
of objects named in the prompt, answer what a correct image should show:
- If a named object should be present (and, for completeness questions,
fully and properly rendered), answer "yes".
- If the prompt says or implies the object should NOT be present, answer
"no".
2) Respect negations ("without", "no", "not") and explicit numbers - e.g.
for "a clown without a balloon", a balloon should NOT be present, so a
balloon question is "no".
3) Answer just yes or no.
Generation prompt: {prompt}
Question: {question}
Answer:
```

Figure 6: **Text-only answer prompt**, given the caption and one question without the image; produces the reference p_k^t.

QUESTION GENERATION PROMPT

```
You are checking whether an image matches a description of it.  Generate
yes/no questions that can each be answered by looking at the image.

Description:
{prompt}

Every question must pass the gate below.  A question that fails it is
noise:  it will be answered unreliably and it will not tell a matching
image from a mismatched one.

THE GATE - DECIDABLE BY LOOKING. The answer must be settled by what is
visibly shown, not by what someone knows, feels, or infers.  Never ask
about intent, mood, atmosphere, style, aesthetic quality, what happened
before or after, or how well something is depicted.  Never ask anything
that needs knowledge of the world beyond the picture itself.

If a question fails the gate, do not keep it and do not weaken it -
replace it with a different question about the same claim that passes.

HOW MANY QUESTIONS: as many as the description needs.  There is no target
number.
- Every distinct thing the description names MUST have at least one
question.  This is a hard minimum, not a suggestion:  a description
naming six things gets at least six questions.
- Beyond that minimum, add a question for every further claim the
description makes about those things - what they look like, how many
there are, what they are doing, and how they are arranged relative to
each other.
- A longer, more detailed description therefore produces more questions
than a short, simple one.  Never stop at a round number.

FLOOR: never return an empty list, and never return only one or two
questions because the rest felt uncertain.  Every description asserts
something checkable.  When you are unsure whether a question passes the
gate, KEEP IT - too few questions is a worse failure than one doubtful
question.
Task:
- Output JSON only.  Do not add markdown, prose, or any text outside the
JSON.
- The JSON must contain a "questions" array of objects.  Each object must
include:
- "question":  a short, specific, yes/no question answerable by
inspecting the image
- "focus":  one of ["object","count","attribute","text","spatial","action"]
- "difficulty":  one of ["easy","medium","hard"]
- "used_tuples":  a list containing the EXACT substring(s) of the
description above that this question checks.  Copy the words verbatim;
do not paraphrase them.
- Phrase every question POSITIVELY: ask whether something IS the case.

Return the JSON only.
```

Figure 7: **Question-generation prompt.** `{prompt}` is replaced by the caption; generation runs at temperature 0.

A.3 COST

Pass Count vs. Measured Latency. As shown in Table 5, DEPICT requires $2N+2$ forward passes per pair ($\sim$ 12.9 passes on average), which is roughly double that of Soft-TIFA ($N+1$, $\sim$ 6.5). However, measured wall-clock latency (Table 6) shows a much smaller gap: under continuous batching, DEPICT takes 0.492 s per pair on GenAI-Bench, only 1.36$\times$ the time of Soft-TIFA (0.362 s) and 1.30$\times$ that of DSG (0.379 s). VQAScore remains the fastest (0.099 s) as it requires just one pass.

Why Overhead is Minimal. The gap between pass count and actual runtime exists because pass costs vary widely: **Question Generation Bottleneck:** Multi-token text generation for question synthesis dominates the overall runtime. This fixed overhead is shared by DEPICT, Soft-TIFA, and DSG. **Single-Token Passes:** DEPICT's additional text-only answers and holistic visual pass only decode a single token each, introducing minimal computational overhead.

Amortization Across Multiple Images. When evaluating m candidate images against a single prompt (e.g., best-of-N selection), question generation and text-only answers are computed once and cached. In this multi-image setting, each additional image costs DEPICT only $N+1$ visual passes, making its marginal cost virtually identical to Soft-TIFA (N passes).

Method	Gen.	Visual	Text	Per pair		Per extra image
				Total	@ $\bar{N}$	(cached)
VQAScore	0	1	0	1	1.0	1
Soft-TIFA-AM	1	N	0	$N+1$	6.5	N
DSG	3	N_{DSG}	0	$N_{\text{DSG}}+3$	9.8	N_{DSG}
DEPICT-dec	1	N	N	$2N+1$	11.9	N
DEPICT	1	$\boldsymbol{N+1}$	$\boldsymbol{N}$	$\boldsymbol{2N+2}$	**12.9**	$\boldsymbol{N+1}$

Table 5: **Forward passes.** *Per pair*: one image scored against one caption, $\bar{N}$=5.47 on GenAI-Bench ($\bar{N}$ per benchmark in Table 4). *Per extra image*: marginal cost when question generation and text-only answers are cached per caption. DSG writes its own questions ($\bar{N}_{\text{DSG}}=6.85$ on GenAI-Bench).

Method	GenAI-Bench	5-bench. mix
VQAScore	0.099	0.083
Soft-TIFA-AM	0.362	0.316
DSG	0.379	0.338
DEPICT-dec	0.443	0.370
DEPICT	**0.492**	**0.411**

Table 6: **Measured time.** Wall-clock seconds per pair on GenAI-Bench and on a mix of all five benchmarks, Qwen3.5-27B, vLLM with batching and prefix caching, one image per prompt, on a single NVIDIA B300.

B ADDITIONAL RESULTS

B.1 FULL BACKBONE SWEEP

Tables 7 and 8 report every method on all eleven backbones, grouped by the axes of Section 4. Fig. 8 repeats the paired comparison of Fig. 4b against each baseline separately.

Size. Table 7 shows that both ladders behave alike. Winoground rises with size for every method, the correlation benchmarks rise modestly and not always monotonically, and DEPICT's NegBench accuracy stays flat. Its NegBench margin is largest at the smallest size in both families (+4.3 on Qwen3.5-4B, +11.3 on InternVL3.5-4B), where the holistic baseline is weakest.

Architecture. Within Gemma-4 (Table 8, top rows), the per-layer-embedding E4B tracks the weaker dense 4B model (InternVL3.5-4B) on GenAI-Bench and Winoground, and the dense 31B comes close to Qwen3.5-27B on TIFA160 and GenAI-Bench but not on RichHF. The 12B variant without a vision encoder is the weakest backbone on every benchmark, below its smaller sibling, and the only one on which DEPICT loses significantly to a baseline (VQAScore on GenAI-Bench).

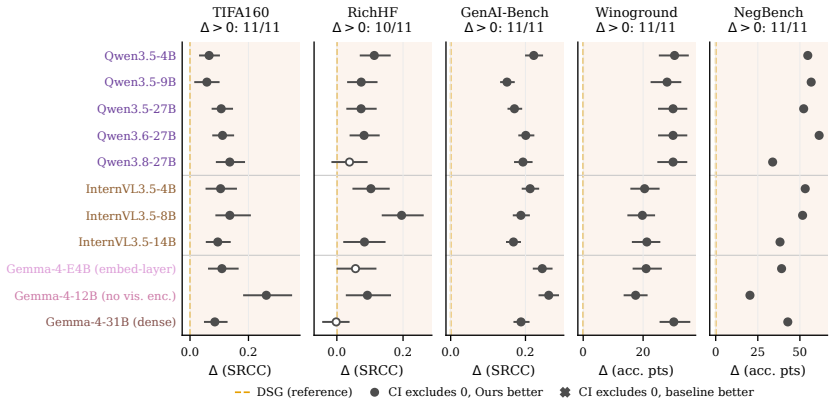
Generation. At a fixed 27B (Table 8, bottom rows), Qwen3.8 is a weaker answerer than Qwen3.5 and Qwen3.6 for nearly every method, yet it gives DEPICT its largest margins on all five benchmarks, including the only significant Winoground gain in the sweep (+5.0).

Backbone	DSG	Soft-TIFA-AM	VQAScore	DEPICT	Δ
TIFA160					
Qwen3.5-4B	0.678	0.715	0.705	**0.743**	+0.028*
Qwen3.5-9B	0.673	0.710	0.702	**0.731**	+0.021
Qwen3.5-27B	0.673	0.762	0.737	**0.780**	+0.018
InternVL3.5-4B	0.592	0.671	0.653	**0.696**	+0.026
InternVL3.5-8B	0.592	0.693	0.693	**0.728**	+0.035*
InternVL3.5-14B	0.616	0.678	0.681	**0.710**	+0.029*
RichHF					
Qwen3.5-4B	0.541	**0.655**	0.616	0.654	−0.001
Qwen3.5-9B	0.588	0.585	0.658	**0.662**	+0.004
Qwen3.5-27B	0.594	0.636	0.656	**0.667**	+0.011
InternVL3.5-4B	0.457	0.458	0.558	**0.560**	+0.002
InternVL3.5-8B	0.412	0.540	0.597	**0.607**	+0.011
InternVL3.5-14B	0.500	0.501	0.572	**0.584**	+0.012
GenAI-Bench					
Qwen3.5-4B	0.484	0.445	0.695	**0.706**	+0.011*
Qwen3.5-9B	0.551	0.479	0.701	**0.702**	+0.001
Qwen3.5-27B	0.549	0.546	0.698	**0.720**	+0.022*
InternVL3.5-4B	0.415	0.442	0.623	**0.627**	+0.004
InternVL3.5-8B	0.467	0.506	0.644	**0.655**	+0.010*
InternVL3.5-14B	0.503	0.450	0.658	**0.671**	+0.013*
Winoground					
Qwen3.5-4B	0.347	0.520	**0.662**	0.652	−0.010
Qwen3.5-9B	0.420	0.575	**0.713**	0.700	−0.013
Qwen3.5-27B	0.445	0.647	**0.755**	0.745	−0.010
InternVL3.5-4B	0.138	0.292	0.310	**0.343**	+0.033
InternVL3.5-8B	0.175	0.333	0.318	**0.372**	+0.040
InternVL3.5-14B	0.198	0.345	0.407	**0.410**	+0.003
NegBench					
Qwen3.5-4B	0.312	0.176	0.816	**0.859**	+0.043*
Qwen3.5-9B	0.280	0.156	0.835	**0.847**	+0.012*
Qwen3.5-27B	0.347	0.191	0.847	**0.869**	+0.023*
InternVL3.5-4B	0.265	0.113	0.684	**0.797**	+0.113*
InternVL3.5-8B	0.285	0.177	0.759	**0.801**	+0.041*
InternVL3.5-14B	0.421	0.073	0.739	**0.803**	+0.064*

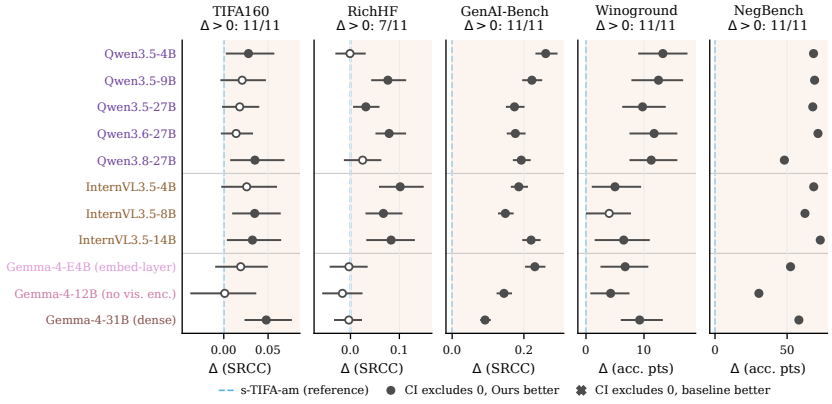
Table 7: **Size ladders**: Qwen3.5 (4B/9B/27B) and InternVL3.5 (4B/8B/14B). SRCC on the correlation benchmarks; Winoground group and NegBench accuracy as fractions. Δ: DEPICT minus the strongest baseline in the row; *: paired BCa 95% interval excludes zero (intervals in Fig. 8). **Bold**: best; underline: second.

Backbone	DSG	Soft-TIFA-AM	VQAScore	DEPICT	Δ
TIFA160					
Gemma-4-E4B	0.595	0.685	0.694	**0.704**	+0.010
Gemma-4-12B	0.285	0.545	0.496	**0.547**	+0.001
Gemma-4-31B	0.668	0.705	0.697	**0.753**	+0.048*
Qwen3.5-27B	0.673	0.762	0.737	**0.780**	+0.018
Qwen3.6-27B	0.677	0.774	0.752	**0.788**	+0.014
Qwen3.8-27B	0.600	0.701	0.680	**0.736**	+0.035*
RichHF					
Gemma-4-E4B	0.519	**0.579**	0.519	0.575	−0.003
Gemma-4-12B	0.331	**0.440**	0.369	0.423	−0.016
Gemma-4-31B	0.583	**0.584**	0.555	0.581	−0.003
Qwen3.5-27B	0.594	0.636	0.656	**0.667**	+0.011
Qwen3.6-27B	0.591	0.594	0.657	**0.673**	+0.016
Qwen3.8-27B	0.540	0.554	0.527	**0.579**	+0.025
GenAI-Bench					
Gemma-4-E4B	0.366	0.381	**0.617**	0.611	−0.006
Gemma-4-12B	0.165	0.283	**0.462**	0.428	−0.035*
Gemma-4-31B	0.512	0.608	**0.701**	0.700	−0.001
Qwen3.5-27B	0.549	0.546	0.698	**0.720**	+0.022*
Qwen3.6-27B	0.533	0.558	0.723	**0.734**	+0.011*
Qwen3.8-27B	0.497	0.497	0.659	**0.690**	+0.031*
Winoground					
Gemma-4-E4B	0.128	0.270	0.335	**0.338**	+0.003
Gemma-4-12B	0.060	0.193	0.198	**0.235**	+0.037
Gemma-4-31B	0.420	0.630	**0.728**	0.723	−0.005
Qwen3.5-27B	0.445	0.647	**0.755**	0.745	−0.010
Qwen3.6-27B	0.458	0.640	**0.767**	0.757	−0.010
Qwen3.8-27B	0.383	0.570	0.632	**0.682**	+0.050*
NegBench					
Gemma-4-E4B	0.358	0.226	0.645	**0.750**	+0.105*
Gemma-4-12B	0.363	0.262	0.506	**0.566**	+0.060*
Gemma-4-31B	0.407	0.253	0.812	**0.835**	+0.023*
Qwen3.5-27B	0.347	0.191	0.847	**0.869**	+0.023*
Qwen3.6-27B	0.262	0.163	0.843	**0.877**	+0.034*
Qwen3.8-27B	0.517	0.374	0.792	**0.855**	+0.064*

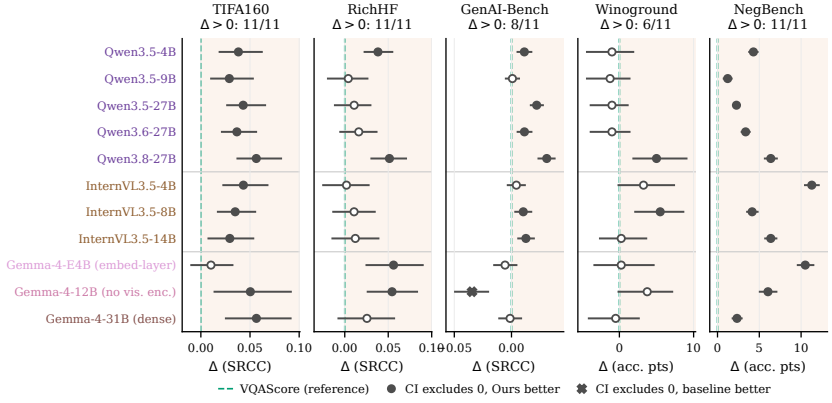
Table 8: **Architectures and generations.** Top of each block: Gemma-4 per-layer-embedding E4B, 12B without a vision encoder, and dense 31B. Bottom: Qwen3.5, 3.6 and 3.8 at a fixed 27B (Qwen3.5-27B repeated from Table 7 as the reference). Columns as in Table 7.



(a) DEPICT minus DSG



(b) DEPICT minus Soft-TIFA



(c) DEPICT minus VQAScore

Figure 8: Performance differences on every backbone and benchmark, with paired BCa 95% intervals. Filled points exclude zero.

B.2 NEGATION ACROSS BACKBONES

Table 9 repeats Fig. 3a on further backbones. The pattern holds on every one: the fixed-YES metrics fall below chance on negation-template items (DSG at most 0.23, Soft-TIFA at most 0.10) and stay near chance on hybrid ones, while DEPICT keeps at least 0.78 on negation and 0.72 on hybrid items.

Backbone	DSG	Soft-TIFA-AM	VQAScore	*Agree*	DEPICT
Affirmative					
Qwen3.5-4B	0.553	0.485	0.831	**0.888**	0.873
Qwen3.5-9B	0.467	0.427	0.874	**0.911**	0.902
Qwen3.5-27B	0.731	0.483	0.885	**0.912**	0.903
Qwen3-VL-4B	0.410	0.337	0.682	**0.819**	0.803
Gemma-4-31B	0.815	0.432	0.801	0.796	**0.816**
Gemma-4-E4B	0.540	0.454	0.623	0.716	**0.720**
Negation					
Qwen3.5-4B	0.191	0.018	0.830	**0.894**	0.872
Qwen3.5-9B	0.091	0.024	**0.835**	0.808	0.827
Qwen3.5-27B	0.107	0.026	0.844	**0.870**	0.862
Qwen3-VL-4B	0.067	0.039	0.499	**0.814**	0.781
Gemma-4-31B	0.072	0.034	0.880	0.891	**0.900**
Gemma-4-E4B	0.230	0.104	0.721	**0.830**	0.824
Hybrid					
Qwen3.5-4B	0.180	0.010	0.787	**0.862**	0.833
Qwen3.5-9B	0.266	0.007	0.795	0.803	**0.812**
Qwen3.5-27B	0.181	0.051	0.810	**0.867**	0.843
Qwen3-VL-4B	0.160	0.009	0.781	0.754	**0.818**
Gemma-4-31B	0.306	0.277	0.759	0.787	**0.794**
Gemma-4-E4B	0.294	0.086	0.595	**0.738**	0.722

Table 9: **NegBench accuracy by correct-answer template** on further backbones (n=5,914, chance 0.25). *Agree*: the decomposed term alone; DEPICT: merged with the holistic term. **Bold**: best per row. Backbones are a subset of those in Appendix B.1.

B.3 GENAI-BENCH BY SKILL

Table 10 breaks GenAI-Bench down by skill tags. DEPICT improves on VQAScore for every tag, significantly for all but *universal*, and the gain is three times larger on prompts with an advanced skill (+0.040) than on basic-only prompts (+0.013). The largest gains are on *differentiation* and *negation* (+0.067, +0.066). The negation tag also shows the fixed-YES failure on a standard benchmark: Soft-TIFA falls to 0.062 SRCC and DSG to 0.260 on these prompts, against 0.601 for DEPICT.

Skill tag	DSG	s-TIFA-AM	VQAScore	Ours	Δ
Basic vs advanced					
Basic only (no advanced tag)	0.591	0.661	0.685	**0.698**	+0.013*
Advanced (any advanced tag)	0.427	0.357	0.607	**0.646**	+0.040*
Basic skills					
Action Relation	0.584	0.621	0.709	**0.733**	+0.024*
Attribute	0.562	0.594	0.700	**0.726**	+0.026*
Part Relation	0.581	0.560	0.683	**0.715**	+0.032*
Scene	0.529	0.506	0.706	**0.722**	+0.016*
Spatial Relation	0.571	0.610	0.690	**0.720**	+0.030*
Advanced skills					
Comparison	0.428	0.521	0.534	**0.579**	+0.045*
Counting	0.472	0.534	0.596	**0.633**	+0.037*
Differentiation	0.423	0.500	0.498	**0.567**	+0.067*
Negation	0.260	0.062	0.534	**0.601**	+0.066*
Universal	0.501	0.583	0.685	**0.693**	+0.008

Table 10: **GenAI-Bench SRCC by skill tag** (Qwen3.5-27B). A prompt can carry several tags, so item counts overlap. DEPICT minus the strongest baseline; *: paired BCa 95% interval excludes zero.

B.4 MERGE-WEIGHT SENSITIVITY

Fig. 9 varies the weight λ of the holistic term in Eq. equation 7 from 0 (decomposed term only) to 1 (holistic term only). On all three correlation benchmarks the merged score exceeds both terms over a broad common range, $\lambda \in [0.32, 0.82]$, and the curves are flat near their optimum: the a-priori choice $\lambda = 0.5$ is within 0.003 SRCC of the best weight on every benchmark.

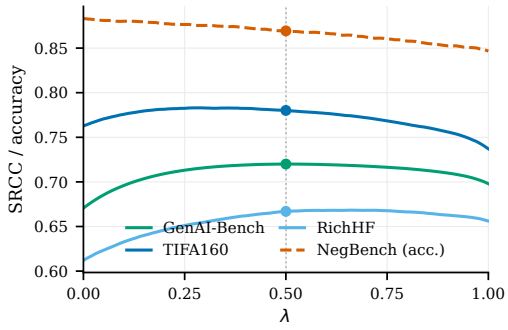


Figure 9: **Sensitivity to the merge weight** λ (Qwen3.5-27B). λ=0: decomposed term only; λ=1: holistic term only; the dotted line marks λ=0.5, fixed a priori and used throughout. SRCC for the correlation benchmarks, accuracy for NegBench (dashed).

C EXTENDED ABLATIONS

C.1 TUPLE EXTRACTION

Replacing direct question generation with DSG's tuple extraction is the largest loss among the pipeline components in Table 3. Table 11 shows the two mechanisms behind it. On NegBench's negation template the extractor returns no tuples for 47% of captions, the correct negated caption then scores 0, and accuracy falls to 4.1%, far below chance, showing tuples cannot express absence. On Winoground captions in a pair share their entities and differ only in how they relate, so tuple-derived question sets for the two captions overlap heavily (mean Jaccard 0.44 against 0.27 for direct generation, 13% identical), which shrinks the paired score margin and removes the contrast the benchmark tests showing tuples also cannot express binding.

(a) NegBench

Template	Zero-tuple	Acc.
Affirmative	22.7%	0.932
Negation	46.8%	0.041
Hybrid	23.1%	0.913
Overall	30.5%	0.644

(b) Winoground

	Direct QG	Tuples
Mean Jaccard (pair)	0.268	0.438
Identical question sets	0.5%	12.9%
Median pair margin	0.198	0.118

Table 11: **Why tuple extraction hurts** (Qwen3.5-27B). (a) Share of NegBench captions yielding no tuples, and accuracy with tuple-based questions, per template. (b) Overlap between the question sets generated for the two captions of a Winoground pair, and the median score margin between them, with direct generation vs. tuples.

C.2 WHEN COMMITMENT WEIGHTING ACTS

The decomposed term weights each question's agreement q_k by its commitment $w_k = |p_k^t - \bar{p}_k^t|$, which is near 1 when the caption settles the answer and near 0 when it leaves it open. The weighted and unweighted scores of an item differ exactly by $\mathrm{Cov}_k(w_k, q_k)/\bar{w}$, so weighting changes a score only when an item's questions differ in commitment, typically because one of them is left open.

Across 11 backbones and 5 benchmarks, weighting improves DEPICT in 36 of 55 evaluations, worsens it in 10 and leaves 9 unchanged ($|\Delta| \leq 0.01$), with a mean gain of 0.25 points. Grouping each benchmark's items into quartiles by their least committed question, the gain is 0.68 points on the quartile with the least committed questions (34 improved, 11 worse) and 0.02 on the most committed quartile, where 44 of 55 evaluations are unchanged. Weighting helps most on backbones whose text channel hedges more often: all three InternVL3.5 models and Qwen3.5-9B improve on every benchmark, while Qwen3.5-4B and Qwen3.5-27B lose slightly (Table 12).

Backbone	Mean Δ (all items)	Mean Δ (least committed)	Improved
InternVL3.5-14B	+1.70	+4.06	5/5
InternVL3.5-4B	+0.75	+1.16	5/5
InternVL3.5-8B	+0.30	+1.13	5/5
Qwen3.5-9B	+0.29	+0.99	5/5
Gemma-4-12B	+0.06	+0.45	3/5
Qwen3.8-27B	+0.04	+0.45	4/5
Qwen3.6-27B	+0.08	+0.26	3/5
Gemma-4-31B	−0.02	+0.13	1/5
Gemma-4-E4B	−0.01	−0.14	1/5
Qwen3.5-4B	−0.13	−0.22	2/5
Qwen3.5-27B	−0.27	−0.79	2/5

Table 12: **Gain of DEPICT from commitment weighting** per backbone, weighted minus unweighted, averaged over the five benchmarks, on all items and on the quartile of items with the least committed questions. Units: SRCC points ($\times 100$) for the correlation benchmarks, accuracy points otherwise. *Improved*: benchmarks with $\Delta > 0.01$ on all items.

C.3 MERGE ABLATION

In Table 13 we ask whether the benefit of merging depends on the decomposed term, by merging VQAScore with Soft-TIFA-AM instead of with *Agree*, using the same $0.5/0.5$ recipe. The two merges tie on TIFA160 and Winoground, and the Soft-TIFA merge is slightly higher on RichHF (-0.008, not significant). They separate where the scoring rule matters: on GenAI-Bench the Soft-TIFA merge does not improve on VQAScore alone (0.696 vs. 0.698), while DEPICT gains 0.024 over it, and on NegBench the Soft-TIFA merge drops to 62.2, well below VQAScore alone (84.7), because the fixed-YES term pulls the holistic score down on exactly the negated items it gets wrong.

Benchmark	Metric	VQAScore	*Agree*	Soft-TIFA-AM	DEPICT (*Agree* $\oplus$ VQA)	Soft-TIFA-AM $\oplus$ VQA
GenAI-Bench	SRCC	0.698	0.671	0.546	**0.720**	0.696
TIFA160	SRCC	0.737	0.763	0.762	**0.780**	**0.780**
RichHF	SRCC	0.656	0.612	0.636	0.667	**0.675**
Winoground	group acc	75.5	65.8	64.8	74.5	**75.2**
NegBench	MCQ acc	84.7	88.3	19.1	**86.9**	62.2

Table 13: **Is the merge specific to *Agree*?** Both merges combine a decomposed term with VQAScore using the same $0.5/0.5$ recipe; only the decomposed term differs. **Bold**: better of the two merges.

C.4 REASONING IN THE ANSWERING STEP

We evaluate enabling chain-of-thought reasoning in the answering step while holding the generated questions fixed (Table 14). Reasoning makes the model commit, pushing answer probabilities toward 0 or 1. On GenAI-Bench and NegBench, which rank many candidates by score, this produces ties, and every reasoning variant scores at or below the reasoning-off row. Winoground, which compares only two captions per image, instead gains up to +5.5 group points, for two reasons. With reasoning, the holistic term less often rejects both captions of an image (18.5% $\rightarrow$ 7.4% of images), a case where the two low scores are nearly tied and their comparison unreliable. And with reasoning in the decomposed term, the two terms fail on more different examples (the correlation ϕ between their

per-example errors falls from 0.52 to 0.43), which leaves the merge more errors to correct. We keep reasoning off to preserve graded probability estimates and avoid the extra token cost.

	Reasoning on	**Winoground** group acc.	**GenAI-Bench**† SRCC	**NegBench**† accuracy
Holistic term	none	75.5	**0.712**	**85.8**
	holistic query	**78.2**	0.664	84.4
Decomposed term	none	65.7	**0.659**	**89.7**
	image and text answers	**66.2**	0.607	85.2
Merged (final)	none	74.5	**0.727**	**88.6**
	holistic query	77.5	0.705	86.9
	image and text answers	77.5	0.711	85.8
	all	**80.0**	0.677	86.1

Table 14: **Chain-of-thought reasoning in the answering step** (Qwen3.5-27B), with questions held fixed so only the answering step changes. **Bold**: best per column within each block. The remaining combinations (image or text answers alone, or with the holistic query) lead to the same conclusion: none exceeds the reasoning-off row on GenAI-Bench or NegBench. †Run on a 100-prompt (600-image) GenAI-Bench subset and a 400-item NegBench subset; Winoground is the full set (n=400).

D Analyses behind Section 5.3

D.1 Answer distributions

In Table 15 we report the share of unique generated questions whose caption-implied answer is NO ($p^t < 0.5$) and the share the caption leaves undecided (commitment $w_k < 0.2$). Questions expecting NO make up 34.7% on NegBench but at most 4% elsewhere. Within NegBench the share is driven by the negation template (86.8%) and the hybrid template (47.1%), against 0.4% for affirmative captions; weighted by question count, the three reproduce the overall 34.7%. Undecided questions are rare ($\leq$ 1.4%), consistent with the small effect of commitment weighting (Table 3).

Benchmark	**Unique Q**	**NO share** % [95% CI]	**Undecided** $w_k < 0.2$, %
GenAI-Bench	8,753	3.0 [2.6, 3.5]	0.3
TIFA160	875	0.1 [0.0, 0.4]	0.0
RichHF	3,019	4.0 [2.7, 5.5]	1.4
Winoground	3,325	2.3 [1.6, 2.9]	1.1
NegBench	11,652	34.7 [33.8, 35.6]	0.2

Table 15: **Answer distributions per benchmark** (Qwen3.5-27B). Shares are over unique (caption, question) pairs: *NO share*, $p^t < 0.5$; *undecided*, commitment $w_k < 0.2$. Intervals from a caption-level bootstrap; NegBench is restricted to correct captions, Winoground to matched pairings.

D.2 Reference reliability

We check the polarity of the caption-only reference p^t on NegBench captions containing a negation, where the absent and present entities are known. Parsing these captions (covering 80.1% of them) and attributing each question to the negated or an affirmed entity, the text channel answers YES to all 3,042 affirmed-entity questions and NO to 92.1% of the 9,454 negated-entity ones. Of the rest, 443 are phrased negatively (*"Is there no person in the image?"*), where YES is correct, giving 96.8% for negated entities and 97.6% overall. The remaining 304 (3.2%) count as errors, though some may be parser misattributions.

D.3 SCORE DISTRIBUTIONS

Fig. 10 shows how the four metrics distribute their scores. VQAScore is near-binary on Qwen3-VL-4B, with most items in its two end bins, and the fixed-YES metrics pile up at the top: DSG places 31–41% of items at exactly 1.0 on both backbones, leaving them unrankable. DEPICT is the most evenly spread in every panel, with 60–64% of its mass between 0.2 and 0.8 on Qwen3.5-27B. Its gain over VQAScore comes from the low end: its lowest-scored 5% of items are rated worse by humans than VQAScore's on every benchmark and backbone, while the top tails coincide.

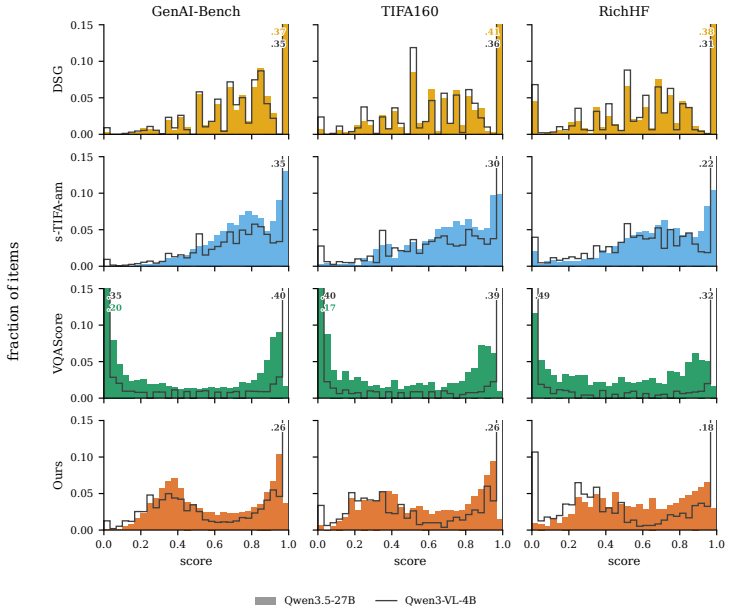


Figure 10: **Score distributions on the three human-correlation benchmarks.** Fraction of items per score bin for each metric on Qwen3.5-27B (filled) and Qwen3-VL-4B (outline); bars exceeding the axis are labelled with their height.

E QUALITATIVE EXAMPLES

Figs. 11–17 show qualitative results for DEPICT on Qwen3.5-27B. Figs. 11–12 trace the full pipeline on individual image–caption pairs. The remaining figures cover the two selection tasks, choosing the image that best matches a caption (Figs. 13–15) and the caption that best matches an image (Figs. 16–17), alongside VQAScore, Soft-TIFA, DSG, and the human judgments.

In these examples, DEPICT selects the human-preferred image or the correct caption where the baselines do not, most clearly on explicit negation (Fig. 17, where only DEPICT selects the correct caption) and on attribute and spatial binding (Figs. 14–15). Its scores are also graded, whereas DSG often assigns a tied score of 1.0 to several candidate images.

Fig. 18 shows two failures, one per selection task, and the per-question scores expose the cause of each. On NegBench #2034 the negation is handled correctly, since the text channel expects *no bottle* for the correct caption, but the image channel wrongly sees a bottle ($p^v \approx 0.7$). This supports the distractor *"This image features a bottle"* and contradicts the correct *"A bottle is not shown"*, and because the holistic term makes the same perceptual error (0.41 against 0.15), DEPICT ranks the correct caption third of four (0.22 against 0.53). On GenAI-Bench #247, all methods prefer the Midjourney image, rated 4.0 by humans, over the DALL-E 3 image rated 5.0. Both terms of DEPICT favour it: the holistic term scores it 0.95 against 0.69, and in the decomposed term the two images differ mainly on whether the scroll is lit by the crystals, which the image channel affirms for Midjourney (agreement 0.98) but largely rejects for DALL-E 3 (0.35), while neither image is judged to show exactly two crystals. The two images are close in faithfulness, and the one-point gap in human ratings may reflect rating noise as much as a metric error.

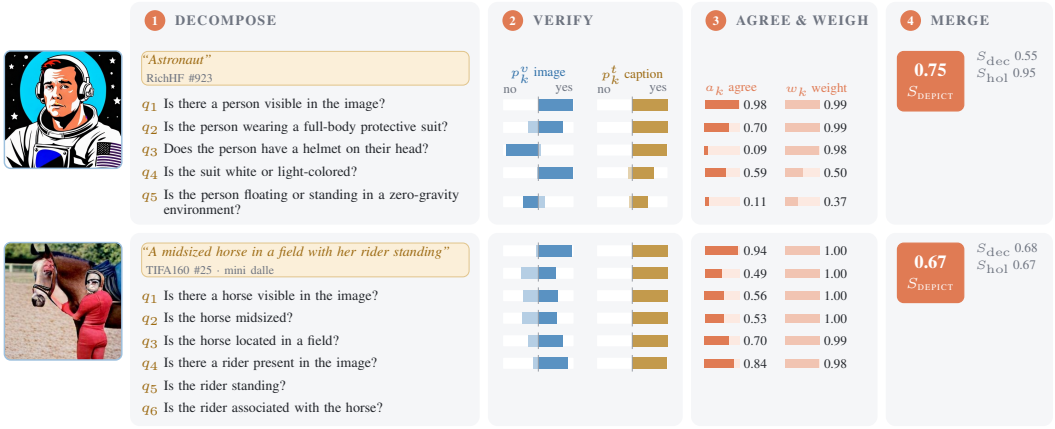


Figure 11: Qualitative Results with scores per questions.

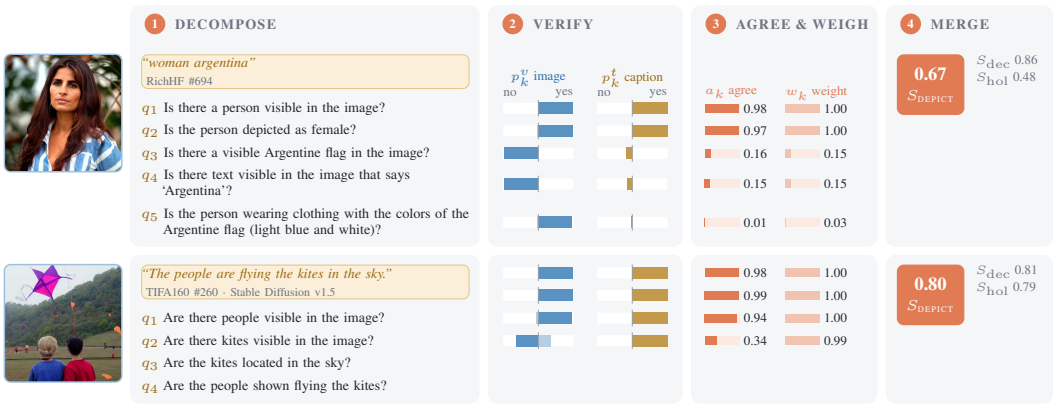


Figure 12: Qualitative Results with scores per questions.

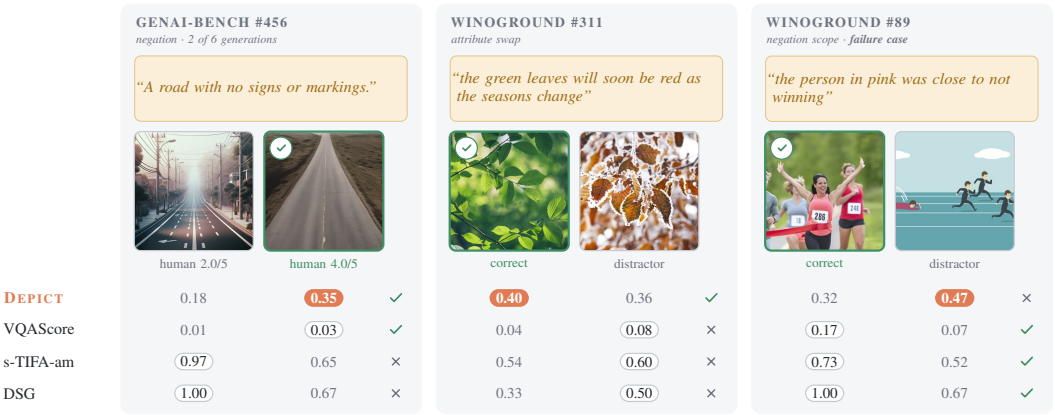


Figure 13: Winoground examples.

GENAI-BENCH #1248

"Three ceramic cups sit to the right of a wooden fork."



	DALL-E 3	DeepFloyd-IF	Midjourney 6	SDXL 2.1	SDXL Base	SDXL Turbo	
DEPICT	0.42	**0.49**	0.32	0.10	0.13	0.35	✓
VQAScore	0.05	0.18	0.01	0.01	0.02	0.20	×
s-TIFA-am	0.81	0.80	0.64	0.20	0.23	0.50	×
DSG	0.86	0.86	0.71	0.29	0.43	0.57	×
Human	3.00	**3.67**	2.00	2.00	2.00	2.67	/5

Figure 14: GenAI-Bench examples with 6 image, 4 scorers and human evaluation.

GENAI-BENCH #552

"A knight in shining armor reads a map by the light of a firefly lantern."



	DALL-E 3	DeepFloyd-IF	Midjourney 6	SDXL 2.1	SDXL Base	SDXL Turbo	
DEPICT	0.74	0.45	**0.77**	0.70	0.68	0.69	✓
VQAScore	0.58	0.31	0.63	0.66	0.48	0.52	×
s-TIFA-am	0.91	0.59	0.90	0.74	0.88	0.87	×
DSG	1.00	0.75	1.00	1.00	0.88	0.75	×
Human	3.33	2.00	**4.67**	2.33	3.00	3.00	/5

Figure 15: GenAI-Bench examples with 6 image, 4 scorers and human evaluation.

NEGBENCH #2102



candidate captions	DEPICT	VQAScore	s-TIFA-am	DSG
"This image shows a handbag."	0.33	0.20	0.46	0.50
"This image features a handbag, with no TV in sight."	0.17	0.01	0.63	0.33
"A tv is not included in this image."	0.13	0.04	0.77	0.50
"A handbag is not present in this image."	**0.35**	0.19	0.46	0.00
	✓	×	×	×

Figure 16: Negbench example with four candidate captions.

NEGBENCH #2123



candidate captions	DEPICT	VQAScore	s-TIFA-am	DSG
"This image shows a bottle"	0.33	0.23	0.42	0.50
"No pizza is present in this image."	0.02	0.03	0.98	0.67
"There is no bottle in this image."	**0.38**	0.23	0.42	0.00
"This image features a bottle, but there is no pizza."	0.13	0.03	0.70	0.33
	✓	×	×	×

Figure 17: Negbench example with four candidate captions.



Figure 18: Question scoring on the failed examples.