From Prompts to Question Graphs: Structure-Aware Evaluation and Selection of Text-to-Image Models

Anonymous authors
Paper under double-blind review

Abstract

Text-to-image evaluation usually compresses the satisfaction of multiple, interacting requirements into a single score for each prompt. This flat treatment creates three ambiguities: a missing entity can make all of its attributes and relations unanswerable, prompts containing more checks can receive disproportionate influence, and the same error may reflect either a basic generation failure or an inability to preserve a concept under composition. We introduce QC-T2I-Bench, a framework that represents each prompt as a graph of attributed atomic questions. Built on Davidsonian Scene Graphs, this representation explicitly connects entities, attributes, relations, and higher-level semantics through their prerequisite structure. We develop a dependency-aware aggregation rule that evaluates a question only when its prerequisites are satisfied and balances contributions across prompts of different complexity. The resulting graph-structured evidence supports analyses unavailable from conventional prompt scores: it separates initial concept realization from conditional preservation, measures the probability of jointly satisfying connected requirements, and provides capability profiles for model selection. Experiments across diverse text-to-image generators and bilingual English–Chinese prompts reveal a sharp compositional bottleneck: joint completion decreases from 80.7% for components involving two capabilities to 37.2% for those involving seven or more. Finally, without training a routing model, question-level profiles select generators under quality–cost constraints, matching ERNIE's 89.51-point estimate while reducing inference cost by 21.3% in GPU-s/MP. These results show that question graphs provide a common representation for reliable comparison, interpretable failure analysis, and efficient deployment of text-to-image models.

1 Introduction

Text-to-image (T2I) evaluation faces a representation problem. A prompt may simultaneously request entities, attributes, relations, rendered text, and world knowledge, yet most evaluation pipelines ultimately compress these interacting requirements into one prompt-level score. As models improve Esser et al. (2024); Black Forest Labs (2024); Wu et al. (2025), that compression becomes increasingly consequential: systems with similar totals may have sharply different capability profiles, and the scalar result does not explain where a model succeeds, why it fails, or which generator best fits a new request.

Conventional metrics provide global image- or prompt-level scores rather than evidence about individual requirements Hessel et al. (2021). Evaluation has therefore moved toward question-generation and visual-question-answering (QG/A) methods that decompose a prompt into local checks Hu et al. (2023); Cho et al. (2024); Li et al. (2024). This is an important step, but decomposition alone does not guarantee fine-grained evaluation. The central issue is how the resulting evidence is attributed, aggregated, and reused.

Many evaluation pipelines still organize this evidence around prompts. They collect prompts under predefined categories, generate local checks, average the checks into a prompt score, and assign that score back to the prompt's category Hu et al. (2023); Cho et al. (2024); Wei et al. (2025); Li et al. (2026). This design creates several recurring problems. A com-

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.

Benchmark	AA	OP	DA	CA	CE
Arena-T2I Hard	●	◐	●	◐	○
ConceptMix	●	◐	○	◐	○
CVTG-2K	◐	○	○	◐	○
DPG-Bench	●	○	●	○	○
DSG	●	●	●	○	○
EvalMuse-40K	●	○	○	○	○
GenAI-Bench	○	●	○	◐	○
GenEval	◐	○	○	○	○
GenEval 2	●	○	○	◐	○
LongTextBench	◐	○	○	◐	○
OneIG-Bench	◐	○	◐	◐	○
PhyBench	◐	○	○	○	○
PRISM-Bench	○	○	○	◐	○
Qwen-Image-Bench	●	○	○	○	○
T2I-CompBench++	◐	○	○	◐	○
T2I-CoReBench	●	○	◐	◐	○
TIFA	●	●	○	○	○
TIIF-Bench	●	○	○	◐	○
WISE	◐	○	○	○	○
QC-T2I-Bench	●	●	●	●	●

Table 1: Support for five structural properties across T2I benchmarks. ●, ◐, and ○ denote full, partial, and no support; AA: atomic attribution; OP: open prompts; DA: dependency awareness; CA: complexity awareness; CE: cross-prompt evidence.

plex prompt may contain requirements from multiple capabilities, making its category score hard to interpret. A new request may not fit the predefined categories. Simple and complex prompts may receive the same total weight even though they contain different numbers of requirements. A missing parent object may cause several dependent checks to fail, exaggerating one error. Finally, evidence about the same requirement is rarely compared across prompts, making it difficult to tell whether a model fails on the basic content or only after additional constraints are added.

Table 1 summarizes these gaps. Atomic attribution (AA), open-prompt evaluation (OP), complexity awareness (CA), and cross-prompt evidence (CE) are four desired properties. Dependency awareness (DA) prevents repeated penalties caused by one missing prerequisite. Existing benchmarks support different subsets of these properties, but none combines all five.

We introduce QC-T2I-Bench, which changes the basic unit of evaluation from the prompt to the atomic question. Each record preserves the question's capability label, visual answer, dependency validity, and prompt context. A two-level taxonomy organizes questions into 21 capabilities under Non-text Entities, Text Content, Attributes, Relations, and High-Level Semantics. Hierarchy-Constrained Question Aggregation (HCQ) removes dependency-invalid evidence and balances capabilities instead of giving every prompt the same total weight. Because prompts use a shared question schema rather than benchmark-specific prompt categories, the same protocol can also be applied to new requests.

We further use DSGs to organize the questions in each prompt according to their dependencies Cho et al. (2024). This lets us test whether a model can satisfy a group of related requirements together, rather than only checking them one by one. We then compare repeated entities across different prompts to distinguish a basic generation failure from a failure caused by additional attributes or relations. The same question records also form a capability profile for each model and can guide generator selection for new requests.

We evaluate QC-T2I-Bench on 13 T2I models with bilingual evidence. The framework supports three connected uses: reliable ranking, fine-grained diagnosis, and cost-aware routing. Bootstrap analysis separates most model pairs while preserving uncertainty for the closest systems. The DSG analysis shows where models fail as related requirements accumulate. Finally, a training-free cost-aware router matches the fixed ERNIE point estimate while reducing GPU-s/MP by 21.3%.

Our contributions are threefold. (1) We introduce QC-T2I-Bench, which converts open prompts into attributed atomic questions and uses validity-aware Hierarchy-Constrained Question Aggregation (HCQ) to avoid prompt-level normalization. (2) We develop a DSG-based compositional analysis that combines complete-component success, cross-prompt root controls, and topology-matched contrasts to localize failures. (3) We demonstrate reliable ranking, fine-grained diagnosis, and training-free quality–cost routing, including a 21.3% cost reduction at the matched ERNIE point estimate.

2 Related Work

Fine-grained evaluation benchmarks. T2I evaluation has progressed from controlled property tests to question-based verification. GenEval Ghosh et al. (2023) evaluates predefined object properties, while GenEval 2 Kamath et al. (2025) adds atom-level questions and atomicity analysis. TIFA Hu et al. (2023) decomposes prompts into VQA pairs, and DSG Cho et al. (2024) organizes atomic questions through valid dependencies. Other work improves supervision and calibration through dimension-specific concepts or element-level annotations Wei et al. (2025); Han et al. (2024), while VQAScore and GenAI-Bench study holistic alignment and its agreement with human preferences Lin et al. (2024); Li et al. (2024). More recent benchmarks use hierarchical capability rubrics or dependency-aware checklists Li et al. (2026); Ban et al. (2026). These developments make local verification more reliable, but their evidence is usually summarized within each prompt or a fixed reporting taxonomy.

Capability-specific and closed-set benchmarks. A complementary line of work deepens evaluation within selected capabilities. WISE and PhyBench focus on world knowledge and physical rules Niu et al. (2025); Meng et al. (2024); T2I-CompBench++, ConceptMix, and DPG-Bench stress compositional or dense instructions Huang et al. (2023); Wu et al. (2024); Hu et al. (2024); and PRISM-Bench, OneIG-Bench, and T2I-CoReBench expand coverage to bilingual, creative, and reasoning-heavy scenarios Fang et al. (2025); Chang et al. (2026); Li et al. (2025a). LongTextBench and CVTG-2K further examine long or multilingual instructions and visual text Geng et al. (2025); Du et al. (2025). Table 1 shows that atomic checking is common, but no prior benchmark connects all five structural properties in one question-centric framework.

Evidence reuse and generator routing. Generator-routing work addresses model selection directly. CATImage learns prompt-conditioned quality–cost decisions, while DiffAgent uses an LLM agent for API selection Li et al. (2025b); Zhao et al. (2024). OctoT2I, ImagePOSER, and GenArtist extend selection to stateful or multi-step orchestration through self-evolving memory, reinforcement learning, or tool planning Jiang et al. (2026); Mohebbi et al. (2025); Wang et al. (2024). These methods obtain routing signals by learning a policy, building agent state, or repeatedly evaluating intermediate outputs. Our router instead reuses question-level histories already collected for ranking and diagnosis, enabling training-free model selection from the same evaluation evidence.

3 Method

QC-T2I-Bench retains the identity, capability coordinates, and dependency context of every atomic judgment. These records support two complementary operations. The scoring path aggregates valid non-text and text evidence into capability coordinates and a hierarchy-constrained question aggregation (HCQ) score. The structural path reuses the recorded DSGs to measure component completion, construct cross-prompt root controls, and test topology-localized outcome coupling. The same records subsequently support model diagnosis and training-free routing.

Question construction and capability coordinates. We collect prompts from six public sources spanning knowledge, entities, attributes, relations, text rendering, reasoning, and long multilingual instructions; source counts are provided in the supplement. Each prompt p is converted into $c(p) = \{(q_i, t_i)\}_{i=1}^{n_p}$, where n_p is the number of atomic questions, q_i is

an independently judgeable question, and t_i is its secondary capability label. Construction follows four rules: Target Yes, Atomicity, Coverage, and Tag Validity. Together, they require a "yes" answer for a compliant image, one visual requirement per question, coverage of salient constraints, and attribution by the question's core predicate. We instantiate this fixed contract with Qwen3-235B-A22B Yang et al. (2025), iterative automated auditing, and a final cleanup pass; the supplement reports source composition and audit details.

Evaluation produces one record per question,

$$r = (p, q, m, \ell, g, t, y, v), \tag{1}$$

where p and q are the prompt and question, m is the evaluated model, ℓ is the language, g is the first-level reporting group, and t is the secondary capability. For non-text capabilities, $y \in \{0, 1\}$ is the native binary outcome and $v \in \{0, 1\}$ indicates dependency validity. Text Content retains the same capability coordinates but uses the transcription statistics defined below for official scoring. The taxonomy contains five first-level groups—Non-text Entities, Text Content, Attributes, Relations, and High-Level Semantics—and 21 secondary capabilities (visualized in the supplement). Each question receives exactly one secondary capability by its core predicate, while one prompt may contribute questions to several groups. Text Content is separated because it uses the dedicated transcription evaluator described below.

Dependency-aware capability scoring. Before adjudication, we organize prompt p into a Davidsonian Scene Graph $G_p = (V_p, E_p)$ Cho et al. (2024), where V_p is the set of question nodes and E_p is the set of directed prerequisite edges. An edge $u \to q \in E_p$ indicates that question u is a semantic prerequisite of q. During adjudication, q is scored only when its prerequisites succeed, preventing one missing entity from becoming repeated attribute and relation failures.

For model m, language scope Λ, and non-text capability t, let $\mathcal{Q}_{\ell,t}$ be the questions in language ℓ, $y_{m\ell q} \in \{0, 1\}$ their adjudicated outcomes, and $v_{m\ell q} \in \{0, 1\}$ their dependency validity. The valid count and score are

$$\begin{aligned} N_{m,\Lambda,t} &= \sum_{\ell \in \Lambda} \sum_{q \in \mathcal{Q}_{\ell,t}} v_{m\ell q}, \\ S_{m,\Lambda,t} &= \frac{1}{N_{m,\Lambda,t}} \sum_{\ell \in \Lambda} \sum_{q \in \mathcal{Q}_{\ell,t}} v_{m\ell q} y_{m\ell q}, \quad t \neq t_{\text{text}}. \end{aligned} \tag{2}$$

Thus $S_{m,\Lambda,t}$ is the mean over valid questions at one capability coordinate when $N_{m,\Lambda,t} > 0$ and is NA otherwise; invalid records enter neither sum.

Text-rendering branch. Rendered text remains a semantic capability coordinate, but its evidence type differs: a binary VQA judgment can accept semantically related text while missing character-level errors. We therefore use Qwen3-VL-Instruct-30B to extract text blocks and optimally match them to targets within each image i. Let $D_{m\ell i}$ be the target-conditioned UTF-16 edit distance and $L_{\ell i}$ the target length, and let $\mathcal{I}^{\text{text}}_{m\ell}$ contain images with an available transcription record. We micro-average over all available text records in language scope Λ:

$$S_{m,\Lambda,t_{\text{text}}} = \max\left(0, 1 - \frac{\sum_{\ell \in \Lambda} \sum_{i \in \mathcal{I}^{\text{text}}_{m\ell}} D_{m\ell i}}{\sum_{\ell \in \Lambda} \sum_{i \in \mathcal{I}^{\text{text}}_{m\ell}} L_{\ell i}}\right). \tag{3}$$

This micro-average gives every target character equal weight; empty extractions incur full deletion cost, while unavailable evaluator records are excluded. If the denominator is zero, the score is NA.

Hierarchy-constrained aggregation. We combine these secondary scores with HCQ, which micro-averages valid questions within each secondary capability, then macro-averages capabilities within each reporting group and across the five groups. With $\mathcal{T}_g$ the capabilities in group g, $T_g = |\mathcal{T}_g|$, and G the number of reporting groups,

$$S^{\text{HCQ}}_{m,\Lambda} = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{T_g} \sum_{t \in \mathcal{T}_g} S_{m,\Lambda,t}. \tag{4}$$

Here $G = 5$ and, in the order Non-text Entities, Text Content, Attributes, Relations, and High-Level Semantics, $(T_1, \ldots, T_5) = (5, 1, 6, 4, 5)$.

Compositional capability analysis. Beyond capability aggregation, we use the recorded DSG structure for three complementary diagnostics. Component exactness tests whether related requirements are jointly satisfied; cross-prompt root matching separates entity-generation difficulty from failures under attached constraints; and marginal-controlled coupling tests whether excess outcome dependence localizes to explicit DSG edges.

A prompt may contain several disconnected structures, so we define its DSG atlas as the maximal weakly connected components $\mathcal{A}_p = \{G_{p,1}, \ldots, G_{p,K_p}\}$. For a component G of a fixed prompt–language pair, let $\mathcal{Q}(G)$ be its questions and $\mathcal{T}(G) = \bigcup_{q \in \mathcal{Q}(G)} \{t(q)\}$ its component tag set. We score each maximal component once for its full tag set rather than enumerating its pairs or lower-order subsets. Components define structural context, not an assumption of statistical dependence; that distinction motivates the edge-controlled test below.

For model m and component G, $C_m(G)$ is the complete-component exactness indicator. It equals 1 only when every question is valid and succeeds, equals 0 when any valid question fails, and is undefined when the component cannot otherwise be scored. Unlike mean coverage, this criterion retains root failures even when their descendants are masked:

$$C_m(G) = \begin{cases} 0, & \exists q \in \mathcal{Q}(G) : \; v_{mq} = 1 \wedge y_{mq} = 0, \\ 1, & \forall q \in \mathcal{Q}(G) : \; v_{mq} = 1 \wedge y_{mq} = 1, \\ \mathrm{NA}, & \text{otherwise.} \end{cases} \tag{5}$$

Applying the same rule separately to roots and descendants gives root survival $R_m(G)$ and descendant completion conditional on root survival. To distinguish a composition-specific root failure from a generally difficult entity, we conservatively match entity roots across prompts without adding cross-prompt dependency edges. For entity identity e in prompt p, the leave-one-prompt-out baseline is

$$\bar{R}^{-p}_{m\ell e} = \frac{1}{|\mathcal{P}_{\ell e} \setminus \{p\}|} \sum_{p' \in \mathcal{P}_{\ell e} \setminus \{p\}} R_{m\ell p' e}. \tag{6}$$

Here, $\mathcal{P}_{\ell e}$ contains prompts in language ℓ with a matched root e, and $R_{m\ell p' e} \in \{0, 1\}$ is its prompt-level root-survival outcome for model m. Matches stay within one language and model and require at least three other prompts. Comparing the local outcome $R_{m\ell pe}$ with $\bar{R}^{-p}_{m\ell e}$ therefore controls for the model's baseline ability to realize the same entity.

Marginal-controlled coupling. Complete-component exactness decreases as more fallible requirements are conjoined even when their outcomes are independent. We therefore test whether joint success follows recorded DSG topology beyond this ordinary multiplication of marginal success rates. For model m, language ℓ, prompt p, matched pair stratum s, and context c, define

$$\begin{aligned} \widehat{P}^{c,11}_{m\ell ps} &= \frac{1}{|\mathcal{M}_{\ell psc}|} \sum_{(i,j) \in \mathcal{M}_{\ell psc}} Y_{m\ell pi} Y_{m\ell pj}, \\ \Delta^{c}_{m\ell ps} &= \widehat{P}^{c,11}_{m\ell ps} - \hat{\pi}^{-p,c}_{m\ell s,1} \hat{\pi}^{-p,c}_{m\ell s,2}. \end{aligned} \tag{7}$$

Here the nonempty set $\mathcal{M}_{\ell psc}$ contains ordered matched pairs (i, j), $Y_{m\ell pi}, Y_{m\ell pj} \in \{0, 1\}$ are their native binary outcomes, and positions 1 and 2 denote the two ordered endpoints. The stratum fixes their ordered capability tags, root/descendant roles, and DSG depths; c is either edge, a direct parent–child pair, or disc, a structurally matched pair from disconnected components in the same prompt. The first term is observed joint success, while the product of leave-one-prompt-out endpoint marginals is the success expected in the same model–language stratum.

To isolate dependence associated specifically with a recorded DSG edge, we compare the two residuals after matching:

$$\Gamma_{m\ell} = \frac{1}{|\mathcal{P}^*_{m\ell}|} \sum_{p \in \mathcal{P}^*_{m\ell}} \frac{1}{|\mathcal{S}^*_{m\ell p}|} \sum_{s \in \mathcal{S}^*_{m\ell p}} \left(\Delta^{\mathrm{edge}}_{m\ell ps} - \Delta^{\mathrm{disc}}_{m\ell ps} \right). \tag{8}$$

Model	English						Chinese					
	Total	Ent.	Text	Attr.	Rel.	Sem.	Total	Ent.	Text	Attr.	Rel.	Sem.
FLUX.2-dev	89.48	97.5	78.6	94.7	89.4	87.1	88.60	97.2	74.4	94.6	89.6	87.3
ERNIE-Image	89.43	97.2	79.7	95.0	89.5	85.7	89.58	97.5	79.7	95.4	90.5	84.8
Z-Image Base	88.11	95.5	79.1	93.6	86.9	85.3	89.15	96.7	79.8	94.3	88.5	86.4
Qwen-Image	87.72	95.7	79.4	93.4	87.0	83.1	88.55	96.3	82.4	93.6	87.9	82.6
Lens	87.34	97.3	72.8	94.6	89.1	82.9	87.37	97.5	73.0	94.4	89.6	82.3
FLUX.2-Klein-9B	87.08	95.8	72.3	94.0	87.9	85.3	85.35	96.7	61.0	94.4	89.1	85.5
HiDream-O1	86.80	95.3	73.1	92.7	87.3	85.5	87.00	95.1	76.1	92.2	87.0	84.6
Z-Image Turbo	85.67	93.9	78.5	91.6	84.0	80.4	85.98	95.1	77.5	91.4	84.8	81.1
FLUX.2-Klein-4B	84.33	96.0	61.9	93.5	86.8	83.4	82.29	95.3	52.0	93.6	88.3	82.3
GLM-Image	83.94	92.6	73.4	88.9	82.9	82.0	84.88	93.1	76.2	89.1	83.9	82.2
LongCat-Image	83.91	96.2	63.8	91.6	84.7	83.3	85.97	96.2	70.0	92.8	87.7	83.2
FLUX.1-dev	76.21	89.9	44.1	88.8	80.1	78.2	32.75	23.0	16.5	47.2	41.8	35.3
HiDream-I1	75.66	91.7	35.2	88.9	81.7	80.8	65.02	78.9	20.6	80.4	73.8	71.4

Table 2: Aligned English and Chinese HCQ results (%). Models are ordered by English total; each language is aggregated independently by equally averaging the five reporting views. Ent., Attr., Rel., and Sem. denote Non-text Entities, Attributes, Relations, and High-Level Semantics. Best values within each language are bold; full 21-dimension profiles are in the supplement.

Here $\mathcal{P}^*_{m\ell}$ contains prompts with matched edge and disconnected contexts, and $\mathcal{S}^*_{m\ell p}$ contains their eligible matched strata; the nested means give equal weight to strata within a prompt and then to prompts. We additionally macro-average these estimates across models. Thus $\Delta = 0$ is consistent with marginal multiplication, while $\Delta > 0$ indicates excess outcome coupling. The contrast Γ asks whether that excess is stronger on explicit DSG edges than on matched disconnected pairs; it is not a quality gain or a causal composition penalty. For this structural test only, Y is recovered before dependency masking, which would otherwise induce edge dependence mechanically. Matching thresholds and full estimators are reported in the supplement.

Together, the scoring and structural paths populate the evidence summarized by the bilingual views in Table 2; the following experiments reuse the same records for ranking, diagnosis, and routing.

4 Experiments

4.1 Experimental Setup

Models and generation. We evaluate 13 recent open-source T2I systems: ERNIE-Image Liu et al. (2026), Qwen-Image Wu et al. (2025), Z-Image Base and Turbo Z-Image Team et al. (2025), FLUX.1-dev and FLUX.2-dev Black Forest Labs (2024), FLUX.2-Klein 4B/9B, GLM-Image Z.ai (2026), HiDream-O1/I1 Cai et al. (2026; 2025), LongCat-Image Meituan LongCat Team et al. (2025), and Lens Chen et al. (2026). We use official defaults and disable configurable prompt enhancement so every generator receives the same prompt.

The fixed benchmark contains 6,573 conceptual prompts, 94,547 English questions, and 94,555 Chinese questions. Analyses are run separately by language: the main paper presents both leaderboards side by side and reports bilingual aggregation, dependency, and routing summaries, while complete language-specific profiles are in the supplement.

Evaluation protocol. Every model generates both language sets. We apply HCQ (Equation 4) independently to each language; the language-specific leaderboards do not pool English and Chinese evidence.

Aggregation robustness. Rank recovery alone confounds scoring noise with rule-induced changes in model gaps. Let $g(t)$ map capability t to its unique reporting group. We hold

	Var. increase	One short-prompt error (pp)		
Language	Prompt vs. Q	Q	Prompt	Increase
English	23.2%	0.0102	0.0151	48.4%
Chinese	21.2%	0.0110	0.0156	41.8%

Table 3: Adjudication-noise sensitivity at 1,000 prompts. "Question" gives equal weight to valid atomic judgments; "Prompt" normalizes prompts first. A short prompt contains at most three questions. Lower is more robust.

the HCQ capability weight $\alpha_t = 1/(GT_{g(t)})$ fixed and compare the noise introduced by two within-capability rules. Let $\mathcal{P}_t$ be the prompts contributing to capability t, $P_t = |\mathcal{P}_t|$, n_{pt} the valid question count from prompt p, and $N_t = \sum_{p\in\mathcal{P}_t} n_{pt}$. Question-micro weighting assigns each judgment α_t/N_t; prompt-first weighting assigns it $\alpha_t/(P_t n_{pt})$. Under independent, zero-mean, equal-variance adjudication errors, their noise-variance ratio is

$$\frac{V_t^{\text{prompt}}}{V_t^{\text{question}}} = \left(\frac{1}{P_t}\sum_{p\in\mathcal{P}_t} n_{pt}\right)\left(\frac{1}{P_t}\sum_{p\in\mathcal{P}_t}\frac{1}{n_{pt}}\right) \geq 1. \tag{9}$$

Here V_t^{prompt} and V_t^{question} are the variances of capability t's weighted adjudication-error contribution under the two rules. The arithmetic–harmonic mean inequality makes the ratio at least one, with equality only when every prompt contributes the same number of questions. Thus equal atomic weights minimize variance under the fixed capability construct, while prompt-first weighting amplifies errors in relatively sparse prompts. The full proof, unequal-variance extension, and rank-gap analysis are reported in the supplement.

4.2 Reliable Ranking

Leaderboard. Table 2 aligns the five capability views and HCQ total across languages; complete 21-dimensional profiles are in the supplement. FLUX.2-dev leads in English and ERNIE-Image in Chinese, but no model is best across every view, keeping each scalar order traceable to its diagnostic coordinates.

Bootstrap ranking reliability. We test whether the leaderboard survives prompt resampling using 2,000 paired bootstrap draws over the full set of retained conceptual-prompt clusters, resampling every attached question together. The 95% paired intervals exclude zero for 68 of the 78 model pairs. FLUX.2-dev has a 55.2% top-1 probability and a 95% rank interval of $[1, 2]$; ERNIE-Image has the complementary 44.8% top-1 probability and the same interval. Thus the broad ordering is resolved, but the 0.042-point gap between the first two models is not evidence of deterministic separation.

Adjudication-noise robustness. Prompt-first normalization increases both adjudication-noise variance and the influence of one error in a short prompt in English and Chinese (Table 3). The smaller full-benchmark effects would hide this difference at the roughly 1,000-prompt scale of recent benchmarks; complete diagnostics are in the supplement.

4.3 Fine-Grained Diagnosis

Capability boundaries. The five views in Table 2 separate strong entity and appearance generation from weaker text, relation, and semantic capabilities. The complete 21-dimensional matrices further localize persistent deficits in Text Content, Scientific and Cultural Knowledge, Temporal Setting, and Logical Constraints; cross-language shifts are capability- and model-specific rather than uniform.

Effect of dependency masking. The official metrics exclude dependency-invalid descendants. Counting them as failures leaves the broad ordering stable ($\rho = 0.978$) but changes two positions, indicating that masking removes cascading penalties without manufacturing the ranking; full shifts are in the supplement.

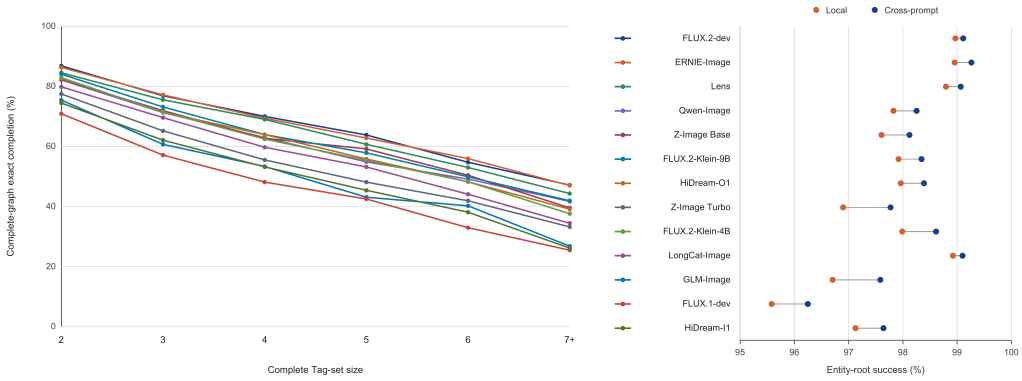

Figure 1: Scaling and root-control views of complete DSG components. Left: prompt-macro exact completion versus the number of distinct capability tags in the full graph. Right: local entity-root success (orange) and the same model's leave-one-prompt-out success on matched roots (blue). Line samples identify the corresponding model curves.

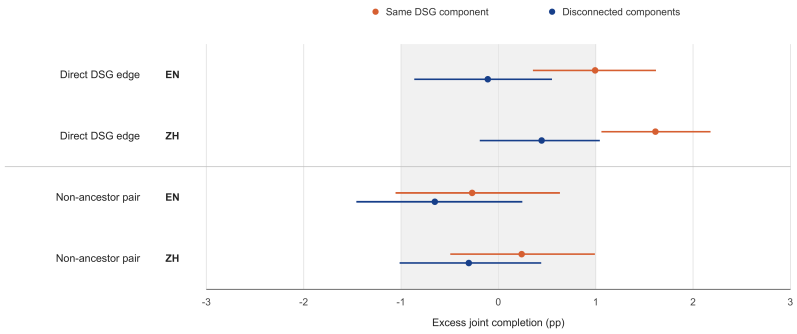

Figure 2: Outcome coupling after removing the joint success expected from ordinary marginal error multiplication. Each row compares question pairs in the same DSG component (orange) with matched pairs from disconnected components (blue). Points are model-macro residuals Δ from Equation 7; bars are 95% prompt-bootstrap intervals.

DSG-component composition diagnosis. The DSG atlas yields 10,584 maximal components from 5,792 conceptual prompts. In Figure 1, mean exact completion falls from 80.7% for two-tag components to 37.2% for components with seven or more tags. This describes joint difficulty, not interaction: exact success becomes stricter as requirements accumulate even under independent errors.

The cross-prompt control uses 4,319–4,347 eligible matched contexts per model. Local roots trail their matched baseline by only 0.14–0.88 points, with nine of thirteen intervals excluding zero. Root survival and conditional descendant completion can therefore locate entity versus attached-structure failure, but do not by themselves establish interaction.

Topology-localized coupling. After marginal control, the direct-edge contrast is +1.10 points (95% prompt-bootstrap interval $[+0.69, +1.50]$), whereas the non-ancestral contrast is smaller and unresolved at +0.38 points $[-0.17, +1.01]$ (Figure 2). Directly dependent requirements thus co-succeed and co-fail beyond their individual rates, and the excess localizes to explicit DSG topology rather than component membership alone. Chinese results reproduce the direct-edge pattern; full estimates are in the supplement.

The full composition fingerprints in the supplement add structure-specific resolution: for example, Qwen-Image is weak on a five-tag Man-Made–Color–Material–Spatial–Quantification component, whereas FLUX.2-Klein-4B is comparatively strong on the related Shape variant.

Method	EN	ZH	Bi.	GPU-s/MP	Save E/F
ERNIE-Image	89.43	89.58	89.51	28.15	0.0/+55.6%
FLUX.2-dev	89.48	88.60	89.04	63.34	-125.0/0.0%
Uniform random	85.08	82.62	83.85	19.20	+31.8/+69.7%
Q-Profile-Q	89.72	89.98	89.85	46.46	-65.0/+26.7%
Q-Profile-C	89.25	89.78	89.51	22.16	+21.3/+65.0%

Table 4: Out-of-fold routing under HCQ. Bi. is the fixed EN/ZH average; Save E/F is relative to fixed ERNIE-Image/FLUX2-dev. Q and C denote quality-first and cost-aware routing.

4.4 Scalable Cost-Aware Routing

Question histories reveal which generators satisfy which atomic requirements. We reuse them to select one T2I generator for each complete request; routing does not select the VQA evaluator.

Routers and evaluation protocol. Our training-free router, Q-Profile, estimates generator quality from the request's active views and reference-fold outcomes without learned parameters. Its quality-first variant (Q) selects the predicted maximum; its cost-aware variant (C) selects the least expensive generator within a validation-selected ϵ of that maximum.

We retain the original source-stratified nested five-fold choices and do not retune settings after the HCQ revision. English and Chinese scores are computed separately and combined with fixed 50/50 weights. Cost is latency times occupied GPUs per output megapixel (GPU-s/MP); paired intervals resample shared conceptual IDs across languages. Full protocol and retrieval ablations are in the supplement.

Quality-cost results. Q-Profile-Q gives the highest routed point estimate, but its paired gain over ERNIE is marginal at the lower endpoint and comes at higher cost. Q-Profile-C matches ERNIE's point estimate while reducing cost by 21.3% relative to ERNIE and 65.0% relative to FLUX2. Its interval against ERNIE crosses the preregistered 0.20-point non-inferiority margin, so the result supports a quality–cost tradeoff rather than lossless routing. Uniform random performs worse, excluding arbitrary model assignment as the explanation.

5 Conclusion

QC-T2I-Bench turns question decomposition from an intermediate step toward a prompt score into reusable evidence. Its attributed records combine five capability groups, dependency validity, and DSG context: HCQ yields complexity-aware rankings, while graph controls separate root realization from attached requirements. Across 13 models and two languages, exact completion falls from 80.7% for two-tag components to 37.2% for components with seven or more tags; after marginal control, excess coupling is resolved on direct DSG edges but not matched non-ancestral pairs, localizing rather than causally identifying composition-associated dependence. The same records support a training-free router that matches ERNIE's point estimate with 21.3% lower GPU-s/MP, although its interval precludes a lossless-routing claim. Thus one auditable evidence base supports ranking, diagnosis, and cost-aware selection without treating any scalar score as universal. Future work should calibrate question construction, DSG parsing, and adjudication across domains.

AI use statement

In this work, generative AI tools were used to assist benchmark-question and tag construction, methodology and experiment refinement, software implementation and debugging, data cleaning and formatting, result interpretation, and English–Chinese translation. They were also used for literature discovery and summarization and to draft or edit portions of the manuscript, including titles, abstracts, and presentation structure. Generated benchmark ar-

tifacts were subjected to the construction contract and audit pipeline described in the paper and supplement, and AI-assisted code and analyses were checked through executable reruns and consistency checks. The authors reviewed the AI-assisted work and take responsibility for the final text, claims, code, and artifacts.

Reproducibility statement

The method section specifies the question contract, scoring rules, structural estimators, and routing protocol. The experimental setup reports the evaluated models, dataset scale, language separation, and resampling design. The supplement provides source composition, audit details, complete capability profiles, estimator definitions, robustness analyses, and routing ablations, together with the artifacts needed to trace the reported results.

References

Yuanhao Ban, Tong Xie, Sohyun An, Yunqi Hong, Evan Frick, I-Hung Hsu, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Arena-t2i hard: Benchmarking and improving faithfulness with dependency-aware checklist, 2026. URL https://arxiv.org/abs/2606.31711.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Qi Cai et al. HiDream-I1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705, 2025. URL https://arxiv.org/abs/2505.22705.

Qi Cai et al. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061, 2026. URL https://arxiv.org/abs/2605.11061.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. Advances in Neural Information Processing Systems, 38, 2026.

Dong Chen et al. Lens: Rethinking training efficiency for foundational text-to-image models. arXiv preprint arXiv:2605.21573, 2026.

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.

Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. arXiv e-prints, pp. arXiv–2503, 2025.

Patrick Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first international conference on machine learning, 2024.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. arXiv preprint arXiv:2509.09680, 2025.

Zigang Geng et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. arXiv preprint arXiv:2507.22058, 2025.

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.

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. URL https://arxiv.org/abs/2412.18150.

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.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

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 Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20406–20417, 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.

Xu Jiang, Bin Chen, Gehui Li, Yule Duan, Ronggang Wang, and Jian Zhang. OctoT2I: A self-evolving agentic text-to-image router. arXiv preprint arXiv:2606.01803, 2026. URL https://arxiv.org/abs/2606.01803.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. arXiv preprint arXiv:2512.16853, 2025.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. arXiv preprint arXiv:2406.13743, 2024.

Niantong Li et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. arXiv preprint arXiv:2605.28091, 2026. doi: 10.48550/arXiv.2605.28091. URL https://arxiv.org/abs/2605.28091.

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? arXiv preprint arXiv:2509.03516, 2025a.

Qinchan Li, Kenneth Chen, Changyue Su, Wittawat Jitkrittum, Qi Sun, and Patsorn Sangkloy. Cost-aware routing for efficient text-to-image generation. arXiv preprint arXiv:2506.14753, 2025b. URL https://arxiv.org/abs/2506.14753.

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. arXiv preprint arXiv:2404.01291, 2024.

Jiaxiang Liu et al. Ernie-image technical report, 2026. URL https://arxiv.org/abs/2605.25347.

Meituan LongCat Team et al. LongCat-Image technical report. arXiv preprint arXiv:2512.07584, 2025. URL https://arxiv.org/abs/2512.07584.

Fanqing Meng et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. arXiv preprint arXiv:2406.11802, 2024.

Hossein Mohebbi, Mohammed Abdulrahman, Yanting Miao, Pascal Poupart, and Suraj Kothawade. Image-POSER: Reflective RL for multi-expert image generation and editing. arXiv preprint arXiv:2511.11780, 2025. URL https://arxiv.org/abs/2511.11780.

Yuwei Niu et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265, 2025.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. GenArtist: Multimodal LLM as an agent for unified image generation and editing. In Advances in Neural Information Processing Systems, volume 37, pp. 128374–128395, 2024.

Xinyu Wei et al. Tiif-bench: How does your t2i model follow your instructions? arXiv:2506.02161, 2025.

Chenfei Wu et al. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. Advances in Neural Information Processing Systems, 37:86004–86047, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Z-Image Team et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699, 2025. URL https://arxiv.org/abs/2511.22699.

Z.ai. GLM-Image model card. https://huggingface.co/zai-org/GLM-Image, 2026.

Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. DiffAgent: Fast and accurate text-to-image API selection with large language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6390–6399, 2024.