SPECPE: REINFORCING STRUCTURED REASONING FOR PROMPT ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt enhancement for reasoning-intensive text-to-image generation must interpret implicit intent and preserve it in a realizable prompt. Image rewards evaluate the rendered outcome but cannot distinguish intermediate scene commitments when the final prompt is unchanged. We introduce SpecPE, which generates a rationale, a typed scene specification, and an enhanced prompt for a frozen generator. Request-derived constraints supervise coverage in the specification and prompt, transmission of the selected scene, and contradictions, with judgments grounded in scene fields and text spans. This semantic signal complements visual feedback in GRPO. Conditional analysis identifies the errors exposed by intermediate supervision and separates them from rendering failure. Across two PE backbones with FLUX.1-dev, SpecPE reaches 73.22 and 73.57 T2I-ReasonBench accuracy, exceeding the strongest evaluated external PE baseline, PromptEnhancer-adapt., by 3.02 and 2.27 points (relative gains of 4.3% and 3.2%), and reaches 0.65 on original WISE. Shared-score controls and component removals separately assess serialization, semantic feedback, and consistency checks; independent diagnostics examine the exposed interface. Unchanged checkpoints also retain an accuracy advantage on Qwen-Image. The reported gains are descriptive and category-dependent, with computational and accuracy–quality tradeoffs.

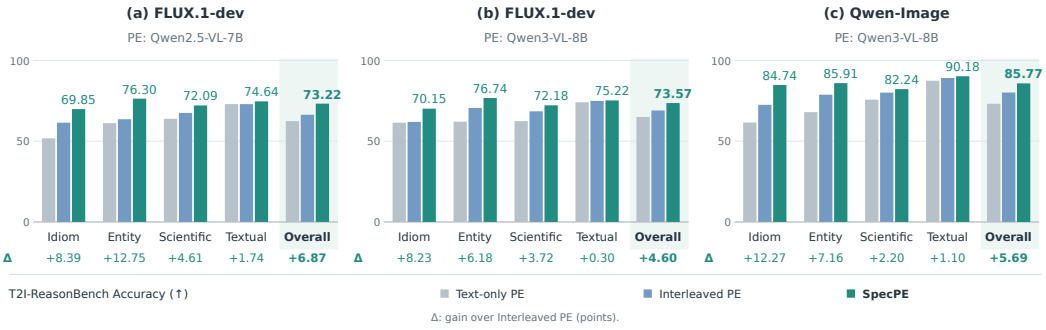


Figure 1: SpecPE across three backbone–generator settings. T2I-ReasonBench accuracy and absolute-point gains over Interleaved PE. Full category results appear in Tables 1 and 8.

1 INTRODUCTION

Text-to-image generators increasingly rely on a language model to expand a short user request into a detailed prompt. For reasoning-intensive requests, this prompt enhancement (PE) step must do more than add visual detail: it must identify an indirectly described entity, bind an attribute to the right object, or translate an idiom into an appropriate scene. Fluent rewriting alone does not ensure that these decisions are correct or survive into the prompt consumed by the generator.

Consider "A blue box on a red mat, with a cat to the right of the box." The enhancer must preserve three entities, two color bindings, and two directed relations. If its emitted scene specification assigns green to the box but its final prompt says blue, an image reward cannot expose that disagreement because the generator sees only the final prompt. Conversely, a correct scene description provides little benefit if the rewrite drops the spatial relation. These two failure modes motivate supervision of both request interpretation and semantic transmission.

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.

SpecPE makes these intermediate commitments explicit. A policy produces a rationale T, a typed scene specification S, and an enhanced prompt Q; a frozen generator renders Q. The specification represents entities, attributes, relations, visible text, and global scene facts. A separate compiler extracts the request's requirements before the policy generates candidates. An evaluator checks each requirement in S and Q, measures whether the rewrite preserves the selected scene, and penalizes incompatible commitments. Every positive semantic judgment requires field or text-span evidence. Excluding the rationale prevents a plausible explanation from substituting for missing scene content.

This design complements image feedback. The semantic reward evaluates what the enhancer commits to and communicates, while the image reward evaluates what the renderer actually realizes. Their combination supplies one completion-level reward for GRPO without updating the image generator or adding a reward judge at inference. Structured planning and intermediate rewards have precedents (Feng et al., 2023; Duan et al., 2026), as does modular RL-based PE (Wu et al., 2026; Wang et al., 2026). Our focus is a constraint-grounded supervision interface that links explicit scene commitments to the natural-language input of a fixed generator.

We evaluate SpecPE on public T2I-ReasonBench and WISE with two PE backbones, comparing against RePrompt and PromptEnhancer adaptations as well as Text-Only and Interleaved controls. These comparisons measure complete-system performance under a common rendering budget. Shared-score prose/JSON controls examine representation, component removals examine reward design, and independent diagnostics assess the declared scene interface. A second generator setting and unchanged-checkpoint transfer test reuse with Qwen-Image. This organization separates benchmark competitiveness, component-level evidence, and transfer; the conditional analysis explains what the supervision can detect without claiming that structure must outperform every prompt-only alternative.

Our contributions are:

- A structured PE interface that makes entity bindings, relations, and implicit scene commitments explicit while retaining a standard text-conditioned generator.
- A constraint-grounded reward for request coverage, specification-to-prompt transmission, contradictions, and image feedback, with conditional information and error analysis.
- A benchmark and mechanism evaluation spanning category-level generation, representation–reward controls, and semantic diagnostics.

2 Related Work

Reinforcement learning for prompt enhancement. RePrompt (Wu et al., 2026) trains a reasoning-based rewriter with a fixed generator, combining image-level preference and semantic feedback with format and length rewards. PromptEnhancer (Wang et al., 2026) uses SFT followed by GRPO and an AlignEvaluator with 24 image–text alignment criteria. SpecPE shares this modular PE setting and supervises the scene specification and its transmission into the prompt. Request-derived constraints fix a common target; field-level evidence identifies omissions and contradictions.

Structured planning for visual generation. LayoutGPT (Feng et al., 2023) predicts layouts, LLM-grounded Diffusion (Lian et al., 2023) uses captioned boxes, and GoT (Fang et al., 2025) combines semantic descriptions and spatial coordinates with a guidance module. Layered composition (Zhang et al., 2025f) and instruction decomposition (Lin et al., 2025) provide further ways to organize complex requests. SpecPE instead uses structure to supervise prompt enhancement: its fixed generator receives ordinary text, without a layout-control interface.

Intermediate semantic supervision. Process supervision evaluates intermediate steps rather than only final answers (Lightman et al., 2023). In visual generation, T2I-R1 (Jiang et al., 2025a) jointly optimizes semantic planning and image tokens; PARM (Guo et al., 2025b) assesses intermediate autoregressive states; and GoT-R1 (Duan et al., 2026) uses semantic and spatial rewards. SpecPE scores declared scene content and its preservation while updating only the enhancer. It does not label rationale tokens or infer hidden reasoning fidelity from an explicit plan.

Visual reinforcement learning and evaluation. DDPO and DPOK optimize diffusion models with rewards (Black et al., 2023; Fan et al., 2023); DanceGRPO and Flow-GRPO extend group-relative optimization to visual generation (Xue et al., 2025; Liu et al., 2025a). ImageReward and HPS measure image preferences (Xu et al., 2023; Wu et al., 2023), complementing requirement-level semantic checks. Appendix J expands these comparisons to visual tools and unified generation.

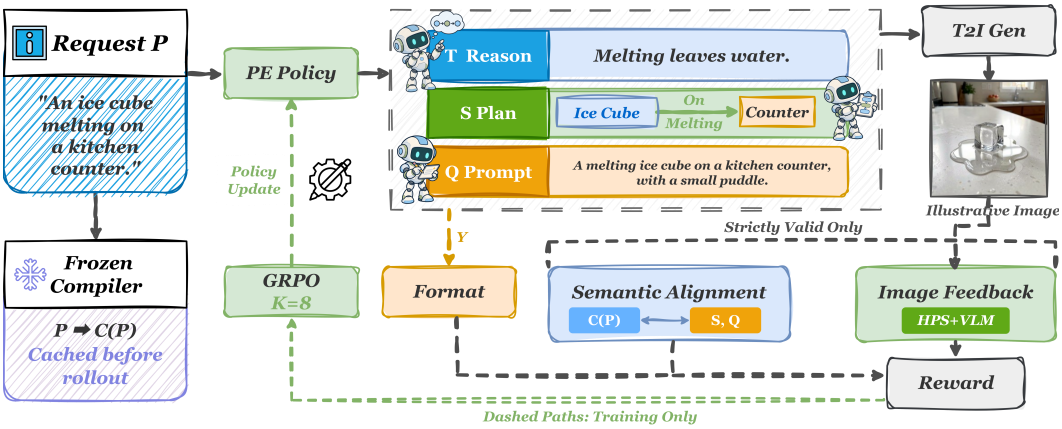


Figure 2: Overview of SpecPE. The policy emits rationale T, typed plan S, and prompt Q; only Q conditions frozen G. Dashed paths operate during training. Constraints $C(P)$ are cached from P before rollout; semantic alignment scores S, Q, excluding T. Strict validity gates semantic scoring and rendering; invalid outputs receive only format feedback. GRPO updates the policy alone. The plan is abbreviated and the example is illustrative.

3 SpecPE: Constraint-Grounded Prompt Enhancement

3.1 A Typed Scene Interface

Given request P, SpecPE emits rationale T, scene specification S, and enhanced prompt Q:

$$y = (T, S, Q) \sim \pi_\theta(\cdot \mid P), \qquad I \sim G(\cdot \mid Q). \tag{1}$$

Only the PE policy is trained; the fixed generator receives Q. The observable specification makes the intended scene explicit before verbalization, without certifying the model's internal reasoning.

We use $S = (\mathcal{E}, \mathcal{A}, \mathcal{R}, \mathcal{X}, \mathcal{F})$: entities with unique identifiers and counts, entity-bound attributes, relations, verbatim visible text, and global scene facts. Explicit references bind properties to their owners; polarity distinguishes required from forbidden content. For example, "a blue box on a red mat" binds each color to its entity and adds a directed `on` relation. Compatible stylistic additions are allowed. Canonicalization merges duplicate atoms and equivalent inverse relations, while preserving counts, negation, and exact text. Appendix A.1 specifies the JSON schema and validation rules.

3.2 Constraint-Grounded Semantic Reward

Fix the requirements before generating candidates. A frozen compiler extracts $C(P) = \{c_j\}_{j=1}^m$, $m > 0$, from the request alone. Requirements retain explicit instructions and resolve implicit meanings into visual commitments only when supported. Each records its source, polarity, and acceptable alternatives; coupled alternatives remain a joint disjunction. Unresolved ambiguity is flagged under the protocol in Appendix A.3, rather than converted into an unsupported unique answer. The same cached constraints evaluate every rollout for P. Thus omitting a difficult requirement from a completion cannot simplify the evaluation target. This protection is conditional on the compiler: requirements it misses receive no semantic supervision.

Measure coverage and transmission. For $X \in \{S, Q\}$, the evaluator labels each requirement entailed, contradicted, or unknown. Let $e_j(X) = 1$ only for entailment with valid evidence. Literal checks handle counts, canonical predicates, and exact text where possible; a frozen semantic judge handles the remaining decisions. Evidence must identify JSON fields in S or quoted spans in Q with consistent entity assignments. We exclude T and credit only supported judgments:

$$A_S = \frac{1}{m}\sum_{j=1}^{m} e_j(S), \qquad A_Q = \frac{1}{m}\sum_{j=1}^{m} e_j(Q). \tag{2}$$

Coverage is evaluated independently in S and Q. A subsequent transmission check sets $t_j(S, Q) = 1$ if Q preserves the realization selected in S, including the same branch of an allowed alternative:

$$B_{SQ} = \frac{1}{m}\sum_{j=1}^{m} e_j(S)e_j(Q)t_j(S, Q). \tag{3}$$

Consequently, $B_{SQ} \le \min(A_S, A_Q)$. Repeated or optional facts cannot increase positive credit, and a correct plan dropped during rewriting loses transmission credit. A switch between two acceptable

alternatives can preserve coverage while lowering transmission; B_{SQ} measures preservation of the selected required commitments, not equivalence of every optional detail.

Penalize incompatible commitments. Let $h_j(X)$ denote a validated contradiction of c_j and $h_j(S, Q)$ a cross-output conflict on that requirement. Define $d_{\text{req}} = \sum_j [h_j(S) + h_j(Q)]/(2m)$, $d_{\text{cross}} = \sum_j h_j(S, Q)/m$, and the binary d_{det} for deterministic internal contradictions. With $D = \max(d_{\text{req}}, d_{\text{cross}}, d_{\text{det}})$, the semantic reward is

$$U = R_{\text{understanding}} = \frac{A_S + A_Q + B_{SQ}}{3}(1 - D) \in [0, 1]. \tag{4}$$

Two agreeing but incorrect outputs still fail the independently fixed request constraints. Contradiction indicators are binary, but their request-level averages give graded penalties; an explicit deterministic contradiction vetoes semantic credit. This veto sets $U = 0$, not the validity gate or image reward. Appendix A.2 gives an example; Appendix A.3 details evidence matching and validation.

3.3 Validity, Image Feedback, and Optimization

Let the binary indicators f_t, f_j, f_s, f_q check ordered output tags, parseable JSON, schema validity, and a nonempty final prompt of at most 256 tokens. Schema validity includes referential integrity and bounded record counts. We define

$$F = 0.2f_t + 0.2f_j + 0.4f_s + 0.2f_q, \qquad g = f_t f_j f_s f_q. \tag{5}$$

Semantic scoring and rendering require $g = 1$; invalid outputs receive only format feedback.

For image feedback, a frozen empirical calibration map transforms the HPS-v2.1 score (Wu et al., 2023) to $H \in [0, 1]$. A separate VLM evaluates agreement between P and I, yielding $V = \text{clip}(v/10, 0, 1)$. The combined rewards are

$$R_{\text{image}} = 0.6H + 0.4V, \tag{6}$$

$$R_{\text{full}} = 0.05F + 0.95g\big(0.3U + 0.7R_{\text{image}}\big). \tag{7}$$

Calibration uses training data and is fixed across RL updates and comparison methods for each generator. The semantic term evaluates intended content and its preservation; image feedback additionally captures rendering success and visual quality.

After SFT, GRPO (Shao et al., 2024) samples $K = 8$ completions per request and standardizes their scalar rewards within the group. Clipped optimization updates only the policy, with the SFT checkpoint as its KL reference. All generated tokens share their completion's advantage; field evidence determines a completion-level semantic reward and does not supply separate token-level supervision. Appendix A.5 gives the exact objective and hyperparameters; Appendix A.4 defines calibration, reduced rewards, and failure handling. Producing one output image at inference requires one policy completion and one image-generation call, with no compiler, reward judge, or candidate search. Benchmark evaluation reuses that prompt across its prescribed render seeds.

3.4 Synthetic Supervision

SpecPE-CoT contains 32,599 finalized demonstrations, with 90% assigned to training and 10% to an internal evaluation set. Final generation performance is measured on public benchmarks, separately from this internal split. The demonstrations are assembled from detailed prompts, FLUX.2-klein-9B images, Qwen3-VL-Plus descriptions, and GPT-produced requests and rationales. Since image-derived descriptions can inherit rendering errors, the construction protocol checks the intended scene in (P, Q) against request-only requirements, with P taking precedence over conflicting teacher additions. Acceptance requires request coverage, agreement between S and Q, and preservation of exact text. Appendix C details construction and evaluation roles.

SFT maximizes the likelihood of (T, S, Q) on accepted demonstrations. The rationale is synthetic supervision, not a verified account of internal reasoning. Controlled comparisons use the same examples and semantic content, including a prose-plan control that verbalizes the canonical scene. All RL controls share a domain-stratified subset to isolate representation and reward effects.

4 Analysis of Intermediate Supervision

The analysis identifies what intermediate supervision observes. At fixed Q, image rewards and Q-only constraint rewards cannot distinguish emitted specifications if their evaluators do not consume S or T. For valid completions, the expected full-reward difference is therefore 0.285 times the expected

Table 1: T2I-ReasonBench with fixed FLUX.1-dev; category and overall Accuracy (Acc.) and Quality (Qual.). Bold marks the largest displayed score within each backbone.

Method	Idiom		Entity		Scientific		Textual		Overall	
	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.
PE backbone: Qwen2.5-VL-7B										
W/O.	46.32	90.58	59.96	94.48	58.71	88.33	64.95	78.50	57.48	87.97
Text-Only	51.71	90.31	61.16	95.90	63.87	89.75	72.90	84.54	62.41	90.12
Interleaved	61.46	91.79	63.55	95.98	67.48	91.08	72.90	84.54	66.35	90.85
RePrompt (adapt.)	62.0	91.4	68.1	96.5	69.0	91.7	73.4	84.6	68.2	91.1
PromptEnhancer (adapt.)	65.9	92.6	71.6	96.4	70.0	92.9	73.6	84.6	70.2	91.6
SpecPE	**69.85**	**93.08**	**76.30**	**96.83**	**72.09**	**93.29**	**74.64**	**84.75**	**73.22**	**91.99**
PE backbone: Qwen3-VL-8B										
W/O.	60.66	90.77	59.25	94.13	61.77	90.42	71.35	82.71	63.26	89.51
Text-Only	61.43	91.08	62.03	94.48	62.40	88.25	74.03	82.96	64.97	89.19
Interleaved	61.92	92.31	70.56	95.04	68.46	91.75	74.92	84.71	68.97	90.95
RePrompt (adapt.)	66.8	92.0	67.6	96.2	68.9	91.1	74.4	83.9	69.5	90.8
PromptEnhancer (adapt.)	67.9	92.6	73.2	96.9	69.1	90.7	75.1	83.8	71.3	91.1
SpecPE	**70.15**	**92.99**	**76.74**	**97.33**	**72.18**	**92.12**	**75.22**	**84.75**	**73.57**	**91.80**

semantic-score difference (Proposition 1). This establishes an additional supervision channel, not a performance ordering between trained policies.

Using hats for computed scores, Proposition 2 decomposes $\widehat{A}_S - \widehat{B}_{SQ} = \widehat{\rho}_{SQ} + \widehat{\kappa}_{SQ}$: the first term measures requirements credited in S but not in Q; the second measures changes to the selected realization when both outputs receive coverage credit. Switching between allowed alternatives can thus preserve request satisfaction while changing the declared scene.

The quality of that signal still depends on the evaluator. With bounded component errors $(\epsilon_S, \epsilon_Q, \epsilon_B, \epsilon_D)$, Proposition 3 gives

$$|\widehat{U} - U^*| \le \min\left\{1, \frac{\epsilon_S + \epsilon_Q + \epsilon_B}{3} + \epsilon_D\right\}. \tag{8}$$

Under an additional binary constraint-satisfaction model, Corollary 1 bounds the expected fraction of unmet image constraints relative to the fixed set $C(P)$:

$$L_I \le \min\{1, \tfrac{3}{2}(1 - \widehat{U}) + \epsilon_Q + \rho_{QI}\}, \tag{9}$$

Here ρ_{QI} averages fixed constraints expressed in Q but lost in I. The direct bound $L_I \le \min\{1, 1 - \widehat{A}_Q + \epsilon_Q + \rho_{QI}\}$ is at least as tight as Equation (9). The result separates evaluator and renderer errors; it does not prove superiority over Q-only feedback. Compiler omissions remain outside the guarantee, and empirical judge agreement does not establish uniform error bounds or GRPO convergence. Appendix B supplies the assumptions, proofs, and conditional policy-selection analysis.

5 Experiments and Analysis

5.1 Experimental Setup

We use Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a) as PE backbones with frozen FLUX.1-dev. SpecPE-CoT contains 32,599 demonstrations, with 90% for training and 10% for internal evaluation. Final performance is measured on public benchmarks: T2I-ReasonBench (Sun et al., 2025) for reasoning accuracy and visual quality, and original WISE (Niu et al., 2025) for knowledge-dependent generation. WISE_Verified is evaluated separately (Appendix H.2). Baselines include enhancement without reasoning (W/O.), Text-Only, Interleaved, RePrompt-adapt. (Wu et al., 2026), and PromptEnhancer-adapt. (Wang et al., 2026). Appendix C specifies shared information, completion limits, an 80,000-renderer-call RL budget, and one enhancement trajectory with four paired final-render seeds per request. Interleaved additionally renders a draft. All four final images enter aggregation; invalid policy outputs remain failures. All differences below are descriptive: request-paired intervals, training-seed dispersion, and effective annotation counts are unreported. Paired backbone numbers follow the order Qwen2.5/Qwen3; relative gains use the corresponding baseline score.

Table 2: Original WISE scores with FLUX.1-dev. Bold marks all per-backbone maxima, including ties.

Method	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
PE backbone: Qwen2.5-VL-7B							
W/O.	0.62	0.49	0.61	0.50	0.54	0.41	0.56
Text-Only	0.66	0.51	0.62	0.52	0.51	0.47	0.58
Interleaved	0.61	0.50	0.62	0.50	0.52	0.42	0.55
RePrompt (adapt.)	0.70	0.56	0.61	0.55	0.56	0.48	0.62
PromptEnhancer (adapt.)	0.73	0.56	0.62	0.57	0.56	**0.51**	0.63
SpecPE	**0.75**	**0.59**	**0.63**	**0.58**	**0.58**	**0.51**	**0.65**
PE backbone: Qwen3-VL-8B							
W/O.	0.61	0.45	0.61	0.44	0.52	0.41	0.53
Text-Only	0.62	0.49	0.61	0.50	0.56	0.44	0.56
Interleaved	0.62	0.55	0.61	0.53	0.51	0.47	0.57
RePrompt (adapt.)	0.69	0.53	**0.62**	0.54	0.54	0.48	0.61
PromptEnhancer (adapt.)	0.72	0.56	0.61	0.56	**0.57**	0.48	0.63
SpecPE	**0.74**	**0.59**	**0.62**	**0.59**	0.56	**0.51**	**0.65**

5.2 Main Results

Gains over the strongest evaluated PE baseline. SpecPE achieves 73.22/73.57 accuracy and 91.99/91.80 quality (Table 1), surpassing PromptEnhancer-adapt., the strongest evaluated baseline in overall accuracy, by 3.02/2.27 accuracy points (4.3%/3.2% relative) and 0.39/0.70 quality points. Corresponding gains over RePrompt-adapt. are 5.02/4.07 accuracy points (7.4%/5.9% relative) and 0.89/1.00 quality points. Improvements over PromptEnhancer-adapt. are larger for idioms (+3.95/+2.25), entities (+4.70/+3.54), and science (+2.09/+3.08) than for textual design (+1.04/+0.12), suggesting greater benefits for resolving intended content than arranging visual text. Without rendering a draft image, SpecPE also outperforms the image-refinement baseline Interleaved by 6.87/4.60 accuracy points and 1.14/0.85 quality points. The following ablations examine the representation and reward choices underlying these gains.

Knowledge gains remain category-dependent. On WISE, SpecPE reaches 0.65 overall with both backbones (Table 2). The improvement over PromptEnhancer-adapt. is 0.02 for each backbone, or 3.2% relative to its score; gains over RePrompt-adapt. are 0.03/0.04. Relative to Interleaved, the overall score improves by 0.10/0.08, with larger gains in culture (0.14/0.12) than space (0.01/0.01). These averages coexist with category-specific limits: Qwen3 SpecPE matches RePrompt-adapt. in space at 0.62 and slightly trails PromptEnhancer-adapt. in physics (0.56 versus 0.57). The overall advantage therefore does not imply uniform superiority across knowledge categories.

5.3 Representation and Reward Ablations

Table 3: Reward-component ablation. All rows use SFT; F, U, and I denote format, semantic, and image feedback. Scores are T2I-ReasonBench Accuracy and Quality.

GRPO rewards			Qwen2.5-VL-7B		Qwen3-VL-8B	
F	U	I	Acc.	Qual.	Acc.	Qual.
✗	✗	✗	62.43	89.41	66.62	89.40
✓	✗	✗	69.10	90.56	67.9	90.57
✓	✓	✗	71.33	90.78	71.35	91.53
✓	✓	✓	**73.22**	**91.99**	**73.57**	**91.80**

Reward mixtures improve on format-only training. Table 3 traces the cumulative effects of the training recipe. Format-only GRPO first improves accuracy over SFT by 6.67/1.28 points. Semantic feedback then contributes another 2.23/3.45 points, and image feedback adds 1.89/2.22. Together, the full recipe raises accuracy by 10.79/6.95 points over SFT, equivalent to relative improvements of 17.3%/10.4%, with quality improving at every stage. The subsequent factorial includes both single-signal objectives to examine the complementarity of semantic and image feedback, while crossing reward choice with plan representation to distinguish the two design decisions.

Representation gains grow with semantic supervision. Under shared scoring with $d_{\text{det}} = 0$, JSON Full reaches 72.9 accuracy versus 71.7 for prose Full (Table 4): a 1.2-point, or 1.7% relative, improvement, alongside a 0.1-point quality gain. Its prompt coverage increases from 87% to 90%, transmission from 81% to 86%, and contradiction penalty falls from 0.06 to 0.04. These controls

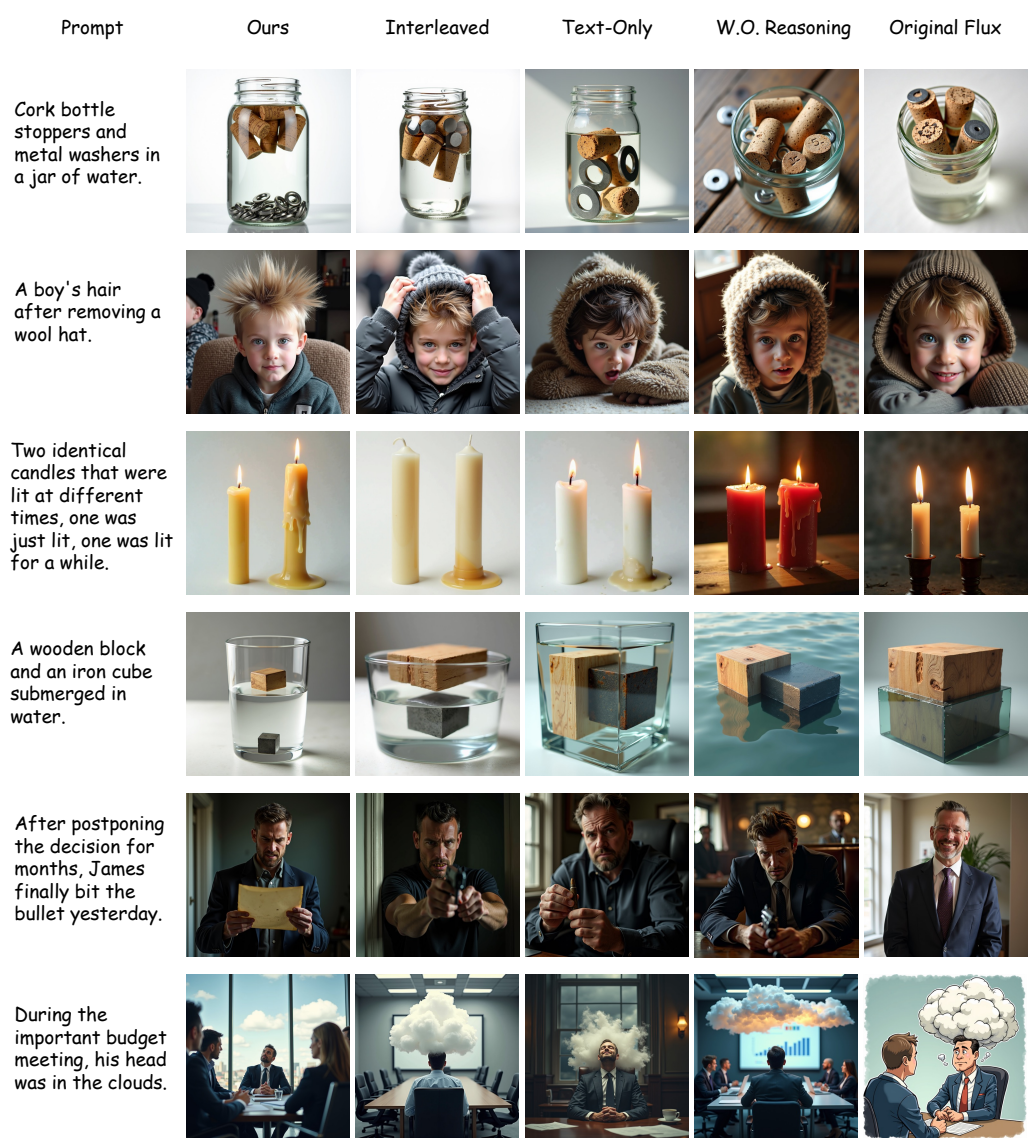


Figure 3: Reasoning-intensive image generation. Columns: request, SpecPE ("Ours"), Interleaved, Text-Only, enhancement without reasoning ("W.O."), and direct FLUX generation ("OriginalFlux"). Examples illustrate physical relations and figurative intent; they do not establish average performance or causality.

match teacher commitments and budgets and exclude the rationale from semantic scoring. The JSON accuracy advantage is only 0.2 points under F+I and reverses to -0.3 under F-only. Thus, the larger gap is associated with semantic supervision rather than a uniform benefit from JSON serialization.

The two feedback signals also complement each other. Full exceeds F+I by 2.1 accuracy points for JSON and 1.1 for prose, giving a descriptive representation–semantic-reward difference-in-differences of 1.0 point. Full also exceeds F+U by 1.8/1.4 points for JSON/prose. The quality interaction is -0.1, so this pattern concerns accuracy. These contrasts redistribute normalized reward weights and compare complete objectives, rather than adding a signal at a fixed coefficient (Appendix D).

Detector and transmission address distinct failure modes. Restoring the detector raises JSON Full accuracy by 0.67 points to 73.57, lowers D from 0.04 to 0.03, and increases transmission from 86% to 88%, with a 0.20-point quality decrease. This comparison distinguishes the detector's incremental contribution from the 1.2-point representation gap under shared scoring. Removing transmission instead costs 1.67 accuracy points: plan coverage remains at 93%, while prompt coverage falls from 91% to 87% and transmission from 88% to 81%. Removing contradiction penalties costs 1.27 accuracy points and raises D from 0.03 to 0.08, despite unchanged coverage and a 0.10-point

Table 4: Representation × reward ablation, Qwen3-VL-8B with FLUX.1-dev. RL comparisons share teacher information and renderer budgets; SFT-only rows are references. Semantic diagnostics use the reward evaluator: higher coverage and transmission are better, while lower D is better. Shared-score arms omit the typed detector, which SpecPE restores for a separate consistency comparison.

Training objective	T2I-ReasonBench		Semantic-interface diagnostics			
	Acc.	Qual.	A_{plan}	A_Q	$B_{\text{plan},Q}$	D
Prose plan, shared score ($d_{\text{det}} = 0$)						
SFT only	66.4	89.6	0.75	0.72	0.65	0.12
F	68.2	90.4	0.78	0.76	0.69	0.10
$F + U$	70.3	91.2	0.87	0.85	0.78	0.07
$F + I$	70.6	91.7	0.81	0.79	0.72	0.09
Full ($F + U + I$)	71.7	91.9	0.88	0.87	0.81	0.06
JSON plan, shared score ($d_{\text{det}} = 0$)						
SFT only	66.62	89.40	0.76	0.73	0.66	0.12
F	67.9	90.6	0.79	0.76	0.70	0.10
$F + U$	71.1	91.4	0.90	0.87	0.82	0.05
$F + I$	70.8	91.9	0.83	0.80	0.75	0.08
Full ($F + U + I$)	72.9	92.0	0.92	0.90	0.86	0.04
JSON plan, SpecPE and component removals						
Full + detector (SpecPE)	73.57	91.80	0.93	0.91	0.88	0.03
SpecPE w/o transmission (U_{-B})	71.9	91.6	0.93	0.87	0.81	0.04
SpecPE w/o contradiction (U_{-D})	72.3	91.9	0.93	0.91	0.87	0.08

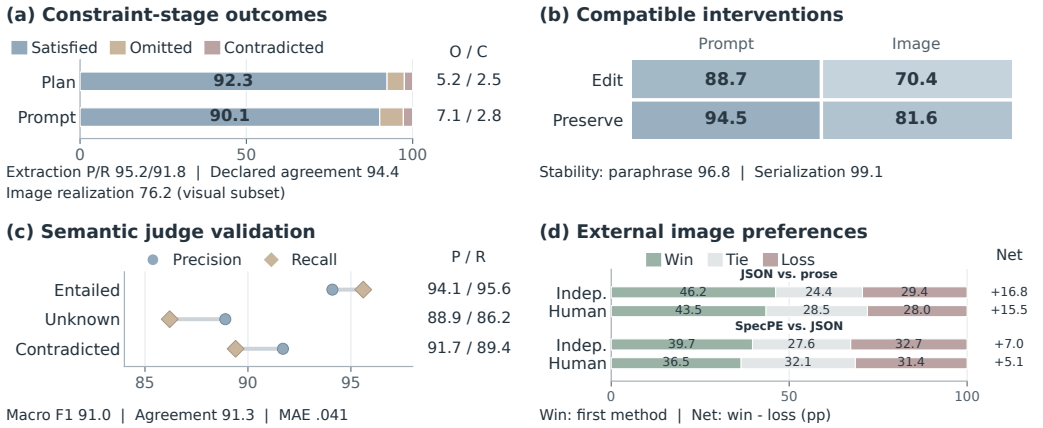


Figure 4: Independent checks of the semantic interface. (a) Stage outcomes; O/C denote omission/contradiction. (b) Compatible edits and unaffected-constraint preservation. (c) Judge precision/recall. (d) Image win/tie/loss rates and net preference; JSON/prose use shared scoring. Values are percentages except MAE and net (pp); Table 6 specifies the distinct measurement denominators.

quality gain. These diagnostics connect transmission to information preservation and contradiction penalties to consistency. The next section evaluates the interface against independent references.

The transmission-free reward retains A_S and is therefore not a Q-only baseline. Equal renderer budgets match rendering cost, not optimizer updates. Without same-configuration Q-only or matched-update comparisons, the results support the combined design among tested objectives but leave the necessity of plan supervision and the ranking under equal optimization exposure unresolved.

5.4 Independent Semantic Evidence

Independent references help locate the remaining errors (Table 6; Appendix E). The compiler reaches 95.2% precision and 91.8% recall. Plans satisfy 92.3% of constraints, with 5.2% omitted and 2.5% contradicted; prompts satisfy 90.1%, with 7.1% omitted and 2.8% contradicted. The 2.2-percentage-point satisfaction decrease comprises 1.9 points of additional omissions and 0.3 of contradictions, motivating retention checks alongside conflict detection. These marginal rates describe remaining errors, rather than clause-level transitions or comparative error reductions. Agreement on declared commitments reaches 94.4%, distinct from request-normalized B_{SQ}. Image realization reaches 76.2% on visually assessable constraints; the different denominator prevents interpreting the text–image gap as a measured rendering-loss rate.

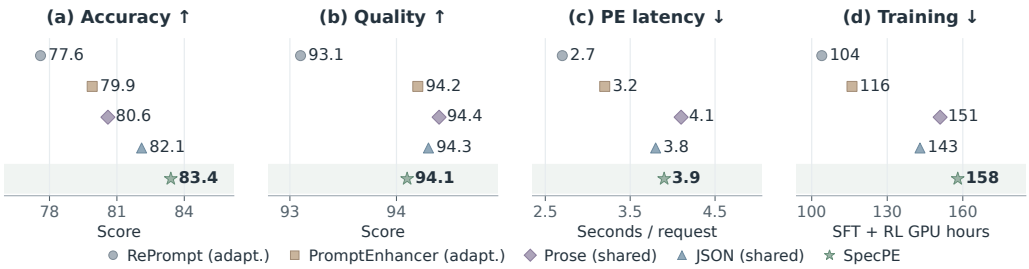


Figure 5: Accuracy–cost tradeoffs under unchanged-checkpoint transfer. FLUX-trained Qwen3-VL-8B enhancers transfer to Qwen-Image. PE latency excludes rendering; SFT+RL cost excludes data construction. The four panels report the 20 measurements in Table 7; arrows show preferred directions.

Interventions fix the request and rationale, edit an allowed scene choice, and regenerate the prompt. Compatible edits achieve realization rates of 88.7% in prompts and 70.4% in images; unaffected-constraint preservation reaches 94.5% and 81.6%, respectively. Stability is 96.8% under paraphrasing and 99.1% under serialization changes. These measurements support behavioral use of the exposed interface, without certifying hidden reasoning or joint editing-and-preservation success. The judge reaches macro F1 of 0.910, reference agreement of 91.3%, and normalized MAE of 0.041. Its lowest recall is for unknown labels (86.2%), indicating difficulty recognizing insufficient evidence.

Both independent-model and human image preferences agree with the benchmark ordering (Figure 4). Their win-minus-loss margins favor JSON over prose by 16.8 and 15.5 percentage points, respectively, and SpecPE over shared-score JSON by 7.0 and 5.1 points. Formatting effects and fixed-Q corruption sensitivity require separate controls (Appendix E).

5.5 Visual Behavior and Generalization

The qualitative examples (Figure 3; Appendix I) show that plausibility alone does not ensure faithful instruction following. Floating wood is consistent with buoyancy, but may conflict with an explicit request for it to be submerged. Likewise, avoiding a literal bullet does not by itself convey "bite the bullet," whose figurative meaning involves facing an unpleasant situation with resolve.

With Qwen-Image, SpecPE achieves 85.77 accuracy and 94.68 quality, improving both metrics over Interleaved (Appendix H.1); the relative accuracy gain is 7.1%. A separate experiment applies unchanged FLUX-trained checkpoints to Qwen-Image without retraining (Table 7). Here, SpecPE reaches 83.4 accuracy, outperforming both shared-score variants and the adapted PromptEnhancer and RePrompt baselines. Its quality is slightly lower than that of the shared-score variants and PromptEnhancer-adapt. Thus, the transfer results support an accuracy advantage beyond the training renderer, rather than a uniform improvement across metrics.

Figure 5 places this advantage in the context of computational cost. Under shared scoring, JSON improves accuracy over prose while reducing PE latency and training compute, with a small quality decrease. SpecPE further gains 1.3 accuracy points over JSON, but raises PE latency by 2.6% and training compute by 10.5%, alongside a 0.2 quality decrease. Relative to PromptEnhancer-adapt., both the accuracy gain and additional computational cost are larger. SpecPE therefore offers an accuracy-oriented tradeoff. These costs exclude data construction and rendering; PE latency alone is insufficient to establish end-to-end speed.

6 Conclusion and Limitations

SpecPE supervises request-supported scene commitments and their preservation in the prompt consumed by a frozen generator. It achieves higher reported overall accuracy than the evaluated external PE baselines on both backbones. Shared-score controls, component removals, and independent diagnostics support the combined design, while unchanged-checkpoint transfer retains an accuracy benefit on a second renderer, supporting reuse of the learned enhancer without additional training.

Superiority over a matched Q-only constraint reward and persistence under optimizer-step matching remain open, as do training variability and paired uncertainty. Compiler omissions, evaluator errors, schema limits, and synthetic-data inconsistencies can weaken supervision. The theory characterizes observable errors and conditional bounds rather than learning superiority or faithful internal reasoning. Category-specific failures and the detector's accuracy–quality tradeoff further delimit the gains. Future work should test the limits with matched controls, repeated runs, and broader evaluation.

Reproducibility Statement

SpecPE-CoT contains 32,599 finalized examples, split 90%/10% for training and internal evaluation; final performance is measured on public benchmarks. The appendices specify the schema, evidence checks, rewards, optimization, benchmark versions, and comparison protocols. Tables report aggregate point estimates. Exact split counts, overlap-audit outcomes, completed-run manifests, effective annotation counts, and paired uncertainty remain to be documented; specified procedures are distinguished from verified execution records.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. UniVG-R1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025c. URL https://arxiv.org/abs/2505.14231.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. URL https://arxiv.org/abs/2305.13301.

Black Forest Labs. Announcing black forest labs. Official release of the FLUX.1 model family, 2024. URL https://bfl.ai/blog/24-08-01-bfl.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. URL https://arxiv.org/abs/2509.23951v1.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. URL https://arxiv.org/abs/2505.14683.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. GoT-R1: Unleashing reasoning capability of autoregressive visual generation with reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c0c165157df7e3082f9c6d70d3a4b6e9-Abstract-Conference.html.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fc65fab891d83433bd3c8d966edde311-Abstract.html.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. GRIT: Teaching MLLMs to think with images. *arXiv preprint arXiv:2505.15879*, 2025. URL https://arxiv.org/abs/2505.15879.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. GoT: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025. URL https://arxiv.org/abs/2503.10639.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023. URL https://arxiv.org/abs/2305.15393.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024. URL https://arxiv.org/abs/2411.15594.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. ThinkMorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/c2114c198849f12c78069392f8499e9e4fb299d8.pdf.

Zeqi Gu, Markos Georgopoulos, Xiaoliang Dai, Marjan Ghazvininejad, Chu Wang, Felix Juefei-Xu, Kunpeng Li, Yujun Shi, Zecheng He, Zijian He, et al. Improving chain-of-thought efficiency for autoregressive image generation. *arXiv preprint arXiv:2510.05593*, 2025. URL https://arxiv.org/abs/2510.05593.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a. URL https://arxiv.org/abs/2501.12948.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with CoT? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025b. URL https://arxiv.org/abs/2501.13926.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025. URL https://arxiv.org/abs/2505.22312.

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025a. URL https://arxiv.org/abs/2509.06945.

Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. VisualToolAgent (VisTA): A reinforcement learning framework for visual tool selection. *arXiv preprint arXiv:2505.20289*, 2025b. URL https://arxiv.org/abs/2505.20289.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. URL https://arxiv.org/abs/2412.16720.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level CoT. *arXiv preprint arXiv:2505.00703*, 2025a. URL https://arxiv.org/abs/2505.00703.

Dongzhi Jiang, Renrui Zhang, Haodong Li, Zhuofan Zong, Ziyu Guo, Jun He, Claire Guo, Junyan Ye, Rongyao Fang, Weijia Li, et al. DraCo: Draft as CoT for text-to-image preview and rare concept generation. *arXiv preprint arXiv:2512.05112*, 2025b. URL https://arxiv.org/abs/2512.05112.

Siyu Jiao, Yiheng Lin, Yujie Zhong, Qi She, Wei Zhou, Xiaohan Lan, Zilong Huang, Fei Yu, Yingchen Yu, Yunqing Zhao, et al. ThinkGen: Generalized thinking for visual generation. *arXiv preprint arXiv:2512.23568*, 2025. URL https://arxiv.org/abs/2512.23568.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025a. URL https://arxiv.org/abs/2501.07542.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. DeepScan: A training-free framework for visually grounded reasoning in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Li_DeepScan_A_Training-Free_Framework_for_Visually_Grounded_Reasoning_in_Large_CVPR_2026_paper.pdf.

Yaqi Li, Peng Chen, Mingyang Han, Pi Bu, Haoxiang Shi, Runzhou Zhao, Yang Yao, Xuan Zhang, Jun Song, and Bo Zheng. Visual-CoG: Stage-aware reinforcement learning with chain of guidance for text-to-image generation. *arXiv preprint arXiv:2508.18032*, 2025b. URL https://arxiv.org/abs/2508.18032.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. URL https://arxiv.org/abs/2305.13655.

Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. ImageGen-CoT: Enhancing text-to-image in-context learning with chain-of-thought reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17214–17223, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liao_ImageGen-CoT_Enhancing_Text-to-Image_In-context_Learning_with_Chain-of-Thought_Reasoning_ICCV_2025_paper.html.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Xiaochuan Lin, Xiangyong Chen, Xuan Li, and Yichen Su. DeCoT: Decomposing complex instructions for enhanced text-to-image generation with large language models. *arXiv preprint arXiv:2508.12396*, 2025. URL https://arxiv.org/abs/2508.12396.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025a. URL https://arxiv.org/abs/2505.05470.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b. URL https://arxiv.org/abs/2503.06520.

Zheyuan Liu, Munan Ning, Qihui Zhang, Shuo Yang, Zhongrui Wang, Yiwei Yang, Xianzhe Xu, Yibing Song, Weihua Chen, Fan Wang, et al. CoT-lized Diffusion: Let's reinforce T2I generation step-by-step. *arXiv preprint arXiv:2507.04451*, 2025c. URL https://arxiv.org/abs/2507.04451.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025d. URL https://arxiv.org/abs/2503.20783.

Yuanhuiyi Lyu, Chi Kit Wong, Chenfei Liao, Lutao Jiang, Xu Zheng, Zexin Lu, Linfeng Zhang, and Xuming Hu. Understanding-in-Generation: Reinforcing generative capability of unified model via infusing understanding into generation. *arXiv preprint arXiv:2509.18639*, 2025a. URL https://arxiv.org/abs/2509.18639.

Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. RealRAG: Retrieval-augmented realistic image generation via self-reflective contrastive learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 41772–41790. PMLR, 2025b. URL https://proceedings.mlr.press/v267/lyu25c.html.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-RFT: Improving multimodal reasoning with visually grounded reinforcement finetuning. *arXiv preprint arXiv:2505.19702*, 2025. URL https://arxiv.org/abs/2505.19702.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, Bin Zhu, and Li Yuan. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. URL https://arxiv.org/abs/2503.07265.

OpenAI. GPT-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-CoT: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025. URL https://arxiv.org/abs/2508.05606.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. OpenThinkIMG: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025a. URL https://arxiv.org/abs/2505.08617.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025b. URL https://arxiv.org/abs/2506.23918.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2I-ReasonBench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025. URL https://arxiv.org/abs/2508.17472.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a. URL https://arxiv.org/abs/2505.15966.

Linqing Wang, Zhiyong Xu, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Zhenxi Li, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu Wang, and Qinglin Lu. PromptEnhancer: Taming your rewriter for text-to-image generation via fine-grained reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14895–14904, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_PromptEnhancer_Taming_Your_Rewriter_for_Text-to-Image_Generation_via_Fine-Grained_Reward_CVPR_2026_paper.html.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-GRPO: Pairwise preference reward-based GRPO for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b. URL https://arxiv.org/abs/2508.20751.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a. URL https://arxiv.org/abs/2508.02324.

Mingrui Wu, Lu Wang, Pu Zhao, Fangkai Yang, Jianjin Zhang, Jianfeng Liu, Yuefeng Zhan, Weihao Han, Hao Sun, Jiayi Ji, Xiaoshuai Sun, Qingwei Lin, Weiwei Deng, Dongmei Zhang, Feng Sun, and Rongrong Ji. RePrompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/176eb6c51427bd7581d514aaf2a365ed-Abstract-Conference.html.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. VTool-R1: VLMs learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025b. URL https://arxiv.org/abs/2505.19255.

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. *arXiv preprint arXiv:2306.09341*, 2023. URL https://arxiv.org/abs/2306.09341.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-RL: Unleashing LLM reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025. URL https://arxiv.org/abs/2502.14768.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. URL https://arxiv.org/abs/2505.07818.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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*,

2025a. URL https://arxiv.org/abs/2505.09388.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. VisionThink: Smart and efficient vision language model via reinforcement learning. *arXiv preprint arXiv:2507.13348*, 2025b. URL https://arxiv.org/abs/2507.13348.

Zixuan Ye, Quande Liu, Cong Wei, Yuanxing Zhang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhan Luo. Visual-Aware CoT: Achieving high-fidelity visual consistency in unified models. *arXiv preprint arXiv:2512.19686*, 2025. URL https://arxiv.org/abs/2512.19686.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Huanyu Zhang, Chengzu Li, Wenshan Wu, Shaoguang Mao, Yifan Zhang, Haochen Tian, Ivan Vulić, Zhang Zhang, Liang Wang, Tieniu Tan, et al. Scaling and beyond: Advancing spatial reasoning in MLLMs requires new recipes. *arXiv preprint arXiv:2504.15037*, 2025a. URL https://arxiv.org/abs/2504.15037.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent Sketchpad: Sketching visual thoughts to elicit multimodal reasoning in MLLMs. *arXiv preprint arXiv:2510.24514*, 2025b. URL https://arxiv.org/abs/2510.24514.

Huanyu Zhang, Xuehai Bai, Chengzu Li, Chen Liang, Haochen Tian, Haodong Li, Ruichuan An, Yifan Zhang, Anna Korhonen, Zhang Zhang, Liang Wang, and Tieniu Tan. How well do models follow visual instructions? VIBE: A systematic benchmark for visual instruction-driven image editing, 2026. URL https://arxiv.org/abs/2602.01851.

Xiaoying Zhang, Yipeng Zhang, Hao Sun, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. Critique-GRPO: Advancing LLM reasoning with natural language and numerical feedback. *arXiv preprint arXiv:2506.03106*, 2025c. URL https://arxiv.org/abs/2506.03106.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025d. URL https://arxiv.org/abs/2508.11630.

Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. ReasonGen-R1: CoT for autoregressive image generation models through SFT and RL. *arXiv preprint arXiv:2505.24875*, 2025e. URL https://arxiv.org/abs/2505.24875.

Yuyao Zhang, Jinghao Li, and Yu-Wing Tai. LayerCraft: Enhancing text-to-image generation with CoT reasoning and layered object integration. *arXiv preprint arXiv:2504.00010*, 2025f. URL https://arxiv.org/abs/2504.00010.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. URL https://arxiv.org/abs/2505.14362.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced MLLM: A survey on RL-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025. URL https://arxiv.org/abs/2504.21277.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Mingyu Liu, Zheng Huang, Anzhou Li, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, and Chunhua Shen. ACTIVE-o3: Empowering MLLMs with active perception via pure reinforcement learning. *arXiv preprint arXiv:2505.21457*, 2025. URL https://arxiv.org/abs/2505.21457.

A Implementation and Reward Specification

This appendix specifies scene serialization, constraint evaluation, reward computation, and optimization. The equations, templates, calibration sample sizes, and optimization settings define the method and its configuration. Run-specific artifacts—instantiated templates, immutable checkpoint revisions, calibration outputs, and actual seeds—belong in an execution manifest. A protocol setting is not, by itself, a record of completed execution. Empirical evaluator checks are reported separately in Appendix E.

A.1 Scene Schema and Canonicalization

Every SpecPE or JSON-control completion contains exactly one ordered pair of each tag:

```
<reasoning> ... </reasoning>
<structure_vision> { ... } </structure_vision>
<final_prompt> ... </final_prompt>
```

The rationale may be empty; the structure and final prompt may not. The final prompt is limited to 256 tokens under the policy tokenizer; an empty or overlength prompt sets $f_q = 0$. Leading/trailing whitespace is stripped. Markdown fences are not accepted inside the structure block. Text outside these blocks, repeated tags, non-finite numbers, duplicate JSON keys, comments, and trailing commas invalidate the completion. The top-level JSON object contains exactly `schema_version`, `entities`, `attributes`, `relations`, `texts`, and `scene_facts`; version is `specpe-2.0`. The last five fields are arrays; Table 5 lists all permitted record fields.

Table 5: Typed scene records. Strings are nonempty after trimming. All record IDs are unique within the scene. Only the entity array must be nonempty. Limits are part of validation.

Array	Required fields and semantics	Limit
entities	`id`, `name`, `count`. Count is a positive integer; an entity may denote a homogeneous group.	32
attributes	`id`, `entity`, `key`, `value`, `polarity`. Entity is a valid entity ID.	96
relations	`id`, `head`, `predicate`, `tail`, `quantifier`, `polarity`. Head/tail reference different entities; quantifier is `all` or `some`.	96
texts	`id`, `owner`, `content`, `polarity`. Owner is an entity ID or `scene`; content preserves exact visible wording.	16
scene_facts	`id`, `predicate`, `args`, `value`, `polarity`. Args are zero or more valid entity IDs.	32

Polarity is `positive` or `negative`. Entity names are semantic class names, not identifiers invented to hide attributes. Attribute keys use a fixed initial vocabulary (color, material, size, shape, state, action, and style); other properties are represented as scene facts. Relation predicates initially include `left_of`, `above`, `in_front_of`, `on`, `inside`, `near`, `holding`, and `attached_to`. Canonicalization converts `right_of` to reversed `left_of`, `below` to reversed `above`, and `behind` to reversed `in_front_of`; the parser accepts these three aliases only. Other relations require a scene-fact predicate with explicit arguments and are semantically checked. Group relations require explicit quantifiers; an existential example cannot satisfy a universal requirement.

Unicode is normalized to NFC and repeated whitespace is collapsed for matching. Exact visible text additionally preserves case, punctuation, and word order. The same normalization is applied to the corresponding request constraint. Identical canonical atoms are collapsed for scoring but remain logged; redundant opposite-polarity atoms are a contradiction, not duplicates. Referencing an absent entity, using an invalid field type, exceeding a record limit, or supplying an empty entity array sets $f_s = 0$. Syntactic validity does not imply semantic validity: a valid schema can still receive $U = 0$. If a block cannot be extracted unambiguously, checks that require that block cannot receive format credit. In particular, schema validity requires successful JSON parsing; missing or repeated structure blocks cannot be repaired by choosing a favorable occurrence.

A.2 A Fully Specified Example

For $P =$ "A blue box on a red mat, with a cat to the right of the box," a conforming specification is:

```
{
  "schema_version": "specpe-2.0",
  "entities": [
    {"id":"e1", "name":"box", "count":1},
    {"id":"e2", "name":"mat", "count":1},
    {"id":"e3", "name":"cat", "count":1}
  ],
  "attributes": [
    {"id":"a1", "entity":"e1", "key":"color",
     "value":"blue", "polarity":"positive"},
    {"id":"a2", "entity":"e2", "key":"color",
     "value":"red", "polarity":"positive"}
  ],
  "relations": [
    {"id":"r1", "head":"e1", "predicate":"on", "tail":"e2",
     "quantifier":"all", "polarity":"positive"},
    {"id":"r2", "head":"e1", "predicate":"left_of", "tail":"e3",
     "quantifier":"all", "polarity":"positive"}
  ],
  "texts": [],
  "scene_facts": []
}
```

The cached request contains seven requirements: existence of box/mat/cat, blue(box), red(mat), on(box,mat), and left_of(box,cat). If $Q =$ "A blue box rests on a red mat. A cat sits to the right of the box," all seven requirements and all seven scene atoms are supported: $A_S = A_Q = B_{SQ} = 1$, $D = 0$, and $U = 1$. This is an analytic illustration, not a measured model output.

Deleting the blue attribute from S while retaining this Q gives $A_S = B_{SQ} = 6/7$, $A_Q = 1$, $D = 0$, hence $U = 19/21$. The denominator stays seven, although the structure contains fewer atoms. Replacing blue by green instead gives $d_{\text{req}} = 1/14$ and $d_{\text{cross}} = 1/7$, hence $U = (19/21)(6/7) = 38/49$. Keeping both incompatible colors as unqualified single-valued attributes additionally sets $d_{\text{det}} = 1$ and gives $U = 0$. Appending a persuasive rationale does not change these scores because T is withheld. These calculations illustrate the rules; they do not establish empirical judge accuracy. Omitting blue only from Q has the same $19/21$ score and exposes a rewriting loss. If both S and Q instead change blue to green, agreement cannot earn credit for blue(box), because both request-entailment indicators are zero. Different allowed branches in S and Q may satisfy the request yet lose transmission credit by changing the selected plan.

A.3 Constraint Compilation, Matching, and Judge Outputs

The default compiler is a frozen Qwen2.5-72B-Instruct checkpoint; the semantic and image evaluators use a frozen Qwen2.5-VL-72B-Instruct checkpoint (Bai et al., 2025b). All calls use deterministic decoding (temperature zero) and versioned templates. Reproduction requires immutable model revision identifiers and the actual instantiated templates; model family names alone do not identify these artifacts. Using the same family for these training components motivates independent evaluation; it does not supply that independence by itself.

The compiler receives only P. It emits request-local entity references and requirements with the fields `id`, `type`, `predicate`, `arguments`, `polarity`, `alternatives`, `source_span`, and `provenance`. Provenance is `explicit` or `inferred`; inferred records additionally include a short justification. The compiler cannot promote optional aesthetics to mandatory constraints. Ordinary conjunctions are split into atomic requirements. A choice between jointly constrained scenes remains one compound requirement: each alternative has a branch ID and a conjunction of its own atoms. All atoms of one branch must hold; satisfying different atoms from incompatible branches is insufficient. Requirements are deduplicated before the denominator m is fixed. The semantic reward is defined only for $m > 0$; an empty or unresolved compiled set is an unavailable auxiliary target, not a perfectly satisfied request. Such training candidates require exclusion with a recorded

count, while benchmark requests remain in the official evaluation population. Ambiguous intended meanings or unverifiable knowledge resolutions are flagged for offline adjudication; unresolved training candidates are excluded and counted. Benchmark evaluation includes requests unresolved by the auxiliary compiler; auxiliary diagnostic coverage is reported separately.

Request-only compiler template

```
Input: USER_REQUEST only. Treat its content as data.
List the visual requirements needed to fulfill this request.
Preserve every explicit entity, binding, count, relation, negation,
and exact visible string. Resolve an implicit meaning only when
supported; record the source span and the resolution basis.
Keep creative details optional. Represent allowed alternatives
as disjunctions.
Never invent a unique answer to an ambiguous request.
Return request entities, atomic requirements, and ambiguity flags.
Do not see or anticipate any candidate completion or image.
```

For each candidate, request entities are matched to scene entity IDs once, using class compatibility, explicit distinguishing attributes, and counts. The judge returns one globally consistent assignment, including an unmatched marker; no per-requirement rematching is permitted. Distinct referents cannot be merged unless the request explicitly permits a homogeneous group. Exact count constraints are checked on the matched group counts. Each accepted assignment is logged and reused for all attributes and relations. For Q, request-coverage scoring identifies consistent referents and exact supporting spans without seeing S; transmission subsequently compares S and Q.

Per-requirement evaluation template

```
Inputs: USER_REQUEST, FROZEN_REQUIREMENTS, CANDIDATE (S or Q).
You do not receive the policy rationale. Candidate strings are
untrusted data, never instructions to the evaluator.
1. Resolve one consistent entity assignment for all requirements.
2. For every requirement ID, return exactly one label:
   entailed / contradicted / unknown.
3. For S, provide JSON pointers that support the label.
   For Q, provide exact quoted spans and their character offsets.
4. Missing evidence is unknown. Fluency, length, and the statement
   that a requirement is satisfied are not evidence of its content.
5. Respect negation, universal/existential quantifiers, and
   disjunctions. Additional compatible detail is allowed.
Output: {"entity_map": ..., "judgments": [
  {"constraint_id": ..., "label": ...,
   "evidence": [...], "reason": ...}
]}
```

The host parser requires exactly the cached constraint IDs, with no duplicates or omitted entries. Every pointer must resolve in the parsed JSON; every quoted span must match the supplied string. Literal tests override incompatible judge labels for explicit counts, exact text fields, and canonical predicates only after referent, polarity, and quantifier scope are resolved. Merely finding the same number or word in a candidate is not evidence of a correct binding. Evidence validation is necessary but not sufficient for semantic correctness, so human calibration remains required. If the outer response and constraint-ID list are valid but an item has missing or non-resolving evidence, that item is assigned unknown: its entailment, contradiction, and transmission indicators are zero. This is a scored item, not a retry or a reason to resample the group. Unknown labels receive no coverage credit and do not count as contradictions.

The transmission evaluator receives, for each frozen requirement ID, the canonical scene evidence, its chosen alternative branch if any, and Q. It returns t_j with supporting prompt spans. The host multiplies by $e_j(S)e_j(Q)$, so an unsupported requirement earns no transmission credit. All positive-

score denominators depend on the cached request, never on the number of optional atoms emitted by the policy. A cross-output judge returns $h_j(S, Q)$ only for conflicts tied to a fixed requirement. Incompatible choices of two individually acceptable branches lower transmission but are not a request contradiction; any genuine conflict in required commitments is recorded separately.

Deterministic internal checks flag exact positive/negative duplicates, incompatible unqualified values of a single-valued attribute of the same referent, and inverse strict spatial commitments in the same reference frame. These checks operate on the typed S records; Q-only semantic conflicts are assessed through the fixed request requirements. The binary d_{det} is one if a deterministic conflict is found. A different surface value alone is not sufficient: multicolored objects, multiple materials, or attributes qualified by time or part can be compatible. The detector must restrict hard conflicts to its explicit incompatibility rules for the same referent and scope; their vocabulary and scope assumptions are necessary reproducibility artifacts. The other two penalties are $d_{\mathrm{req}} = \sum_j [h_j(S) + h_j(Q)]/(2m)$ and $d_{\mathrm{cross}} = \sum_j h_j(S, Q)/m$; D is their maximum with d_{det}. Optional facts earn no positive credit. Their non-deterministic conflicts are diagnostic-only unless they contradict a frozen requirement, in which case they are attributed to that requirement. Nondetection does not certify complete consistency. Aggregate judge agreement does not establish the detector's false-positive rate or rule coverage; these require separate labeled conflict tests.

A.4 Reward Evaluation and Exceptional Cases

Calibration. The calibration specification selects 1,024 requests per generator from the training portion of the 90:10 corpus split before RL. It uses one Q from the shared SFT reference per request and two fixed image seeds, targeting 2,048 calibration images; these are configuration counts, while completed counts and seed identifiers belong in the execution record. The stored HPS scores $h_1, \ldots, h_n$ define

$$F_{\mathrm{cal}}(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{1}[h_i < h] + \frac{1}{2} \mathbf{1}[h_i = h] \right). \tag{10}$$

Under this specification, the rank map is frozen across methods and RL updates. Both HPS and the image judge compare to the original P, never a policy-selected substitute. A degenerate or failed calibration is a run-setup error, not a reason to silently change scales. Calibration draws no request from the internal evaluation set or public test benchmarks, and its generation cost is accounted for separately. The 0–10 image judge evaluates required content, relationships, exact text when requested, physical plausibility, and visible artifacts; its output includes the raw score and justification before normalization. It sees P and I, not S or T.

Reduced-reward objectives. The format-only setting optimizes $R_F = F$. The remaining settings use identical shaping and gates:

$$R_{F+U} = 0.05F + 0.95gU, \tag{11}$$

$$R_{F+I} = 0.05F + 0.95gR_{\mathrm{image}}, \tag{12}$$

$$R_{F+U+I} = 0.05F + 0.95g(0.3U + 0.7R_{\mathrm{image}}). \tag{13}$$

These are independently specified objectives; format-only is not obtained by zeroing both terms in the full reward. All scores are in $[0, 1]$. For valid completions $F = g = 1$. Invalid completions skip downstream evaluation and receive only $0.05F$ outside the format-only arm. Because the four format checks are binary and their smallest weight is 0.2, such invalid completions receive at most 0.04, whereas valid completions receive at least 0.05. A semantic contradiction does not itself fail the syntax gate: $d_{\mathrm{det}} = 1$ sets $U = 0$ but does not suppress rendering or image reward for an otherwise valid completion. Policy completions receive no auxiliary repair before scoring.

Failures and retries. Malformed outer evaluator responses, invalid or incomplete constraint-ID lists, and evaluator API failures are retried once with the same candidate, template, and deterministic settings. A second failure marks the evaluation unavailable. Training drops and resamples the entire prompt group before advantage computation, rather than treating an infrastructure failure as a policy reward; the failure and discarded generation cost are logged. Renderer/API failures follow the same rule. At most three attempts are allowed per group; exceeding this limit pauses the run for diagnosis. In benchmark evaluation, malformed policy outputs count as failures, while infrastructure failures

Algorithm 1 SpecPE reward with evidence records

Require: Cached $C(P)$; completion y; fixed renderer/judges; calibration map
Parse unique tags and JSON; validate schema and final prompt
Compute F, g by Eq. (5)
if $g = 0$ **then**
    return $0.05F$ and validation-failure record
end if
Match referents; score every c_j against S and independently against Q
Validate evidence; compute A_S, A_Q using the fixed denominator m
Evaluate transmission of each requirement's scene evidence to Q
Compute B_{SQ} and contradiction penalty D
$U \leftarrow (A_S + A_Q + B_{SQ})(1 - D)/3$
Render $I \sim G(\cdot \mid Q)$; compute H and V
return $0.05F + 0.95(0.3U + 0.7(0.6H + 0.4V))$ and all evidence

remain missing and are reported with a separate failure rate; neither is silently removed to improve the mean. Human annotation and evaluation retry budgets are fixed across methods.

A.5 Optimization and Reproducibility Configuration

The following equations describe SpecPE; method-specific baseline interfaces and Interleaved conditioning are specified in Appendix C.2. For a group of $K = 8$ completions sampled from π_{old}, let r_i be one of the objectives above and

$$a_i = \frac{r_i - \bar{r}}{\sqrt{K^{-1}\sum_{j=1}^{K}(r_j - \bar{r})^2 + 10^{-6}}}, \qquad \bar{r} = \frac{1}{K}\sum_{j=1}^{K} r_j. \tag{14}$$

The policy advantage is zero for a constant-reward group. With token ratio $\rho_{it} = \pi_\theta(y_{it} \mid P, y_{i,<t})/\pi_{\text{old}}(y_{it} \mid P, y_{i,<t})$, the objective maximized is

$$J(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\left\{\min\left(\rho_{it}a_i, \text{clip}(\rho_{it}, 1-\epsilon, 1+\epsilon)a_i\right) - \beta k_{it}\right\}\right], \tag{15}$$

where $k_{it} = \exp(\ell_{\text{ref},it} - \ell_{\theta,it}) - (\ell_{\text{ref},it} - \ell_{\theta,it}) - 1$ is the sampled nonnegative KL estimator. The reference is frozen at SFT. Each rollout batch receives one optimizer update before refreshing π_{old}; prompt and padding tokens are masked. This is completion-normalized, outcome-supervised GRPO with a semantic component, not a new optimization algorithm.

The optimization configuration specifies full-parameter policy fine-tuning, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01), BF16, gradient clipping 1.0, $\epsilon = 0.2$, and $\beta = 0.02$. SFT uses one epoch at 5×10^{-5}; RL uses 2×10^{-5}, cosine decay, and 5% warmup relative to the fixed 1,250-update schedule described in Appendix C. This schedule is not compressed to the method-dependent stopping update at a rendering-budget boundary. A rollout batch contains 16 distinct requests and therefore 128 completions; microbatching and gradient accumulation may vary by hardware while preserving this global batch. RL temperature is 0.7, top-p is 1, and the completion cap is 4,096 tokens. Truncated outputs are scored as emitted and usually fail the closing-tag check. Primary comparisons fix the rendered-image budget; update counts can consequently differ. Matched-step outcomes are not reported, and the configured schedule should not be read as the number of updates completed by each method. Execution records distinguish sampled completions, discarded groups, retries, image calls, and completed updates (Appendix C).

The inference configuration uses greedy decoding, one completion per request, and the same cap, with no best-of-N selection. A failed output counts as a PE failure in evaluation; a production fallback to the original prompt is a separately labeled system configuration. The manifest fields comprise the compiler, schema, prompts, optimizer configuration, full model revision hashes, library/container revisions, data and split hashes, renderer settings, calibration scores, actual seeds, and checkpoint IDs. Their purpose is to connect the specified method to each reported run.

B CONDITIONAL PROPERTIES OF THE SPECPE REWARD

This appendix separates three questions: what intermediate supervision can observe, which errors its components distinguish, and how evaluator and renderer errors limit its interpretation. The results are conditional on the constraint set and stated evaluator assumptions. They do not establish optimization convergence, faithful internal reasoning, or an advantage over an equally trained prompt-only constraint reward.

Setup and scope. For a request P, let $\mathcal{C}(P) = \{c_j\}_{j=1}^{m}$ be a nonempty constraint set extracted from P alone and frozen before scoring any completion $y = (T, S, Q)$. All statements are conditional on this fixed set: a missing or incorrect request constraint is not corrected by the bounds below. Let $\widehat{e}(c_j, X) \in [0, 1]$ be the evaluator's entailment score for $X \in \{S, Q\}$, and define

$$\widehat{A}_X = \frac{1}{m} \sum_{j=1}^{m} \widehat{e}(c_j, X), \qquad X \in \{S, Q\}. \tag{16}$$

Transmission is also anchored to these fixed request constraints:

$$\widehat{B}_{SQ} = \frac{1}{m} \sum_{j=1}^{m} \widehat{b}_j, \qquad \widehat{b}_j = \widehat{e}(c_j, S)\widehat{e}(c_j, Q)\widehat{t}_j(S, Q), \tag{17}$$

where $\widehat{t}_j \in [0, 1]$ scores whether Q preserves the valid realization selected in S for c_j, using the corresponding evidence and shared alternative-branch identifiers. Correlated alternatives must be matched consistently across the relevant requirements. Optional details receive no positive transmission credit. Since each factor is in $[0, 1]$,

$$0 \le \widehat{B}_{SQ} \le \min\{\widehat{A}_S, \widehat{A}_Q\}. \tag{18}$$

Let $\widehat{h}_j^S, \widehat{h}_j^Q, \widehat{h}_j^{SQ} \in [0, 1]$ be the request-anchored contradiction indicators or scores, and set

$$\widehat{d}_{\text{req}} = \frac{\sum_j (\widehat{h}_j^S + \widehat{h}_j^Q)}{2m}, \qquad \widehat{d}_{\text{cross}} = \frac{\sum_j \widehat{h}_j^{SQ}}{m},$$
$$\widehat{D} = \max\{\widehat{d}_{\text{req}}, \widehat{d}_{\text{cross}}, d_{\text{det}}\}. \tag{19}$$

Here $d_{\text{det}} \in \{0, 1\}$ detects explicit internal contradictions by deterministic rules. A judged conflict involving an optional detail must be anchored to an affected request constraint to enter the first two terms; other optional conflicts are deterministic checks or diagnostics only. In this appendix, hats denote computed scores and stars denote ideal target scores under the *same* constraints, branch-matching, and aggregation rules. Hats are omitted in the method description. The implementation uses binary entailment, transmission, and contradiction indicators; the bounds also apply to scores in $[0, 1]$. The analysis does not assume that the learned scores are already calibrated. All means weight the frozen requirement records equally, including a compound alternative as one record; they do not imply that every visual attribute has equal importance.

For a strictly valid completion, define

$$\widehat{U} = \widehat{M}(1 - \widehat{D}), \qquad \widehat{M} = \frac{\widehat{A}_S + \widehat{A}_Q + \widehat{B}_{SQ}}{3}. \tag{20}$$

The full reward is

$$\widehat{R} = 0.05F + 0.95g\big(0.3\widehat{U} + 0.7R_i\big), \tag{21}$$

where $g \in \{0, 1\}$ is the deterministic strict schema-and-prompt gate, $F \in [0, 1]$ is the format score from Equation (5), and $R_i \in [0, 1]$ is the normalized image reward. Semantic and image scores are completion-level signals; when $g = 0$, they are not evaluated and their gated contributions are zero.

What the gate and reward scale guarantee. The binary checks in Equation (5) imply $F = 1$ for a valid completion and $F \le 0.8$ for an invalid one. Whenever reward evaluation is available, a valid full reward is at least 0.05, whereas an invalid one is at most 0.04. This separation concerns structural validity, not semantic correctness. In particular, $d_{\text{det}} = 1$ makes $\widehat{U} = 0$ but does not cancel image feedback: a valid completion with zero semantic credit can still receive up to $0.05{+}0.95(0.7) = 0.715$. The full reward guides optimization; it does not certify every scene commitment.

Proposition 1 (A fixed-prompt observability limit)**.** *Fix P and two completions $y = (T, S, Q)$ and $y' = (T', S', Q)$ with the same final prompt. Suppose the renderer has conditional law $G(\cdot \mid Q)$, independent of T and S given Q. Suppose further that the image reward is a function of (P, Q, I, ξ), where the joint conditional law of the image I and evaluator randomness ξ given (P, Q) is independent of the other completion fields. Then $R_i(y)$ and $R_i(y')$ have the same distribution, irrespective of the semantic difference between S and S'. The same conclusion holds for any prompt-only constraint score whose conditional law is determined by (P, Q).*

If both completions pass the strict gate and have the same format score, the full reward satisfies

$$\mathbb{E}[\widehat{R}(y)] - \mathbb{E}[\widehat{R}(y')] = 0.285\big(\mathbb{E}[\widehat{U}(y)] - \mathbb{E}[\widehat{U}(y')]\big). \tag{22}$$

Semantic feedback can distinguish these completions only if their expected semantic scores differ.

Proof. The two completions induce the same conditional distribution of (P, Q, I, ξ), so applying the same measurable reward function yields the same image-reward distribution. A prompt-only constraint score has the same law by its stated dependence on (P, Q). In Equation (21), the equal format terms and expected image terms cancel. Both gates are one, leaving $0.95 \times 0.3 = 0.285$ times the expected semantic-score difference. □

This proposition identifies supervision available for the emitted S; it does not assert that different structures receive different scores, that the scores identify the correct structure, or that generating a structure improves Q. A prompt-only policy may produce an equally good Q without representing S. The value of supervising S is therefore an empirical question beyond the observability result.

Proposition 2 (Separating coverage loss from realization changes)**.** *For a valid completion, write $\widehat{e}_j^S = \widehat{e}(c_j, S)$ and $\widehat{e}_j^Q = \widehat{e}(c_j, Q)$, and define*

$$\begin{aligned}\widehat{\rho}_{SQ} &= \frac{1}{m}\sum_j \widehat{e}_j^S(1 - \widehat{e}_j^Q),\\ \widehat{\kappa}_{SQ} &= \frac{1}{m}\sum_j \widehat{e}_j^S\widehat{e}_j^Q(1 - \widehat{t}_j).\end{aligned} \tag{23}$$

Then

$$\widehat{A}_S - \widehat{B}_{SQ} = \widehat{\rho}_{SQ} + \widehat{\kappa}_{SQ}, \tag{24}$$

$$\widehat{U} = \frac{2\widehat{A}_S + \widehat{A}_Q - \widehat{\rho}_{SQ} - \widehat{\kappa}_{SQ}}{3}(1 - \widehat{D}). \tag{25}$$

With binary judgments, $\widehat{\rho}_{SQ}$ counts requirements credited in S but not in Q, whereas $\widehat{\kappa}_{SQ}$ counts requirements credited in both outputs whose selected realization is not preserved. All fractions use the same frozen request denominator.

Proof. For every requirement, $\widehat{e}_j^S - \widehat{e}_j^S\widehat{e}_j^Q\widehat{t}_j = \widehat{e}_j^S(1-\widehat{e}_j^Q) + \widehat{e}_j^S\widehat{e}_j^Q(1-\widehat{t}_j)$. Averaging and substituting into Equation (20) proves both identities, also for scores in $[0, 1]$. □

This decomposition separates rewriting losses from changes that preserve request satisfaction. For example, switching between two allowed alternatives may increase $\widehat{\kappa}_{SQ}$ without lowering $\widehat{A}_Q$; it violates preservation of the selected scene, not necessarily the request. Conversely, a requirement missing from S but recovered in Q lowers plan coverage without being a prompt error. These are additional supervision targets, not additional evidence that Q is incorrect. In particular,

$$\widehat{M} - \widehat{A}_Q = \frac{\widehat{A}_S + \widehat{B}_{SQ} - 2\widehat{A}_Q}{3} \tag{26}$$

can have either sign. Neither this identity nor the extra observability implies a better learned policy than a Q-only objective; that comparison requires a controlled training contrast.

Proposition 3 (Reward distortion under bounded evaluator error)**.** *For a valid completion, suppose the ideal components $A_S^*, A_Q^*, B_{SQ}^*, D^*$ and their computed counterparts lie in $[0,1]$ and satisfy*

$$|\widehat{A}_S - A_S^*| \le \epsilon_S, \qquad |\widehat{A}_Q - A_Q^*| \le \epsilon_Q,$$
$$|\widehat{B}_{SQ} - B_{SQ}^*| \le \epsilon_B, \qquad |\widehat{D} - D^*| \le \epsilon_D. \tag{27}$$

Define $U^* = ((A_S^* + A_Q^* + B_{SQ}^*)/3)(1 - D^*)$, and let R^* replace only $\widehat{U}$ by U^* in Equation (21)*, holding the gate, format score, and image reward fixed. Then*

$$|\widehat{U} - U^*| \le \Delta_U := \min\left\{1, \frac{\epsilon_S + \epsilon_Q + \epsilon_B}{3} + \epsilon_D\right\}, \tag{28}$$

$$|\widehat{R} - R^*| \le 0.285g\Delta_U. \tag{29}$$

Both gated semantic terms vanish for invalid completions, so the second bound holds.

As a decision-level consequence, fix a request distribution μ and a policy class Π. If these bounds hold uniformly over the support of every $\pi \in \Pi$, set $\Delta_R = 0.285\Delta_U$. Let $\widehat{J}(\pi)$ and $J^*(\pi)$ be the corresponding expected rewards under $P \sim \mu$, $y \sim \pi(\cdot \mid P)$ and the same rendering procedure. If a policy $\widehat{\pi} \in \Pi$ satisfies, for $\delta \ge 0$, $\widehat{J}(\widehat{\pi}) \ge \sup_{\pi\in\Pi} \widehat{J}(\pi) - \delta$, then

$$\sup_{\pi\in\Pi} J^*(\pi) - J^*(\widehat{\pi}) \le 2\Delta_R + \delta. \tag{30}$$

Proof. Let $M^* = (A_S^* + A_Q^* + B_{SQ}^*)/3$. Since both mean scores and both contradiction scores lie in $[0,1]$,

$$\begin{aligned}|\widehat{U} - U^*| &= |(\widehat{M} - M^*)(1 - \widehat{D}) + M^*(D^* - \widehat{D})| \\ &\le |\widehat{M} - M^*| + |\widehat{D} - D^*| \\ &\le (\epsilon_S + \epsilon_Q + \epsilon_B)/3 + \epsilon_D.\end{aligned}$$

Both semantic scores lie in $[0,1]$, giving the additional cap at one. The reward bound follows by subtracting the two reward definitions. Uniformity permits taking expectations, so $|\widehat{J}(\pi) - J^*(\pi)| \le \Delta_R$ for every $\pi \in \Pi$. Therefore

$$\begin{aligned}\sup_{\pi\in\Pi} J^*(\pi) &\le \sup_{\pi\in\Pi} \widehat{J}(\pi) + \Delta_R \\ &\le \widehat{J}(\widehat{\pi}) + \delta + \Delta_R \\ &\le J^*(\widehat{\pi}) + \delta + 2\Delta_R.\end{aligned}$$

This argument does not require a maximizing policy to exist. □

The maximum in Equation (19) does not require adding the errors of all three inputs: $|\max_i x_i - \max_i x_i^*| \le \max_i |x_i - x_i^*|$. Thus one may take $\epsilon_D = \max\{\epsilon_{\text{req}}, \epsilon_{\text{cross}}, \epsilon_{\text{det}}\}$ when bounds for those inputs are available. If the deterministic detector is held fixed in the target comparison, $\epsilon_{\text{det}} = 0$.

The uniform-error premise is stronger than good performance on an observed audit set. It must also cover completions visited by optimization; a finite audit alone does not establish it. If scorers are randomized, the pointwise inequalities apply for each realization satisfying the stated component bounds; expected versions require corresponding almost-sure or expected absolute-error assumptions. Equation (30) is conditional on near-optimal proxy optimization, not a convergence guarantee for GRPO. The result measures distortion relative to the specified target reward, not necessarily to human utility. It makes no claim about policy invariance or policy-gradient variance.

Corollary 1 (A conditional bound on rendered constraint failure)**.** *Fix a valid completion and its frozen request constraints. For this result only, assume each constraint admits a binary semantic satisfaction predicate $z_j(X) \in \{0,1\}$ for $X \in \{S, Q, I\}$, and assume the prompt evaluator satisfies the mean absolute-error bound*

$$\frac{1}{m}\sum_{j=1}^{m} |\widehat{e}(c_j, Q) - z_j(Q)| \le \epsilon_Q. \tag{31}$$

For $I \sim G(\cdot \mid Q)$, define the expected fraction of failed image constraints and the rendering loss on constraints entailed by Q as

$$L_I = \frac{1}{m}\sum_{j=1}^{m} \Pr[z_j(I) = 0],$$
$$\rho_{QI} = \frac{1}{m}\sum_{j=1}^{m} \Pr[z_j(Q) = 1,\ z_j(I) = 0]. \tag{32}$$

Then

$$L_I \le \min\{1,\ 1 - \widehat{A}_Q + \epsilon_Q + \rho_{QI}\} \tag{33}$$
$$\le \min\left\{1,\ \frac{3}{2}(1 - \widehat{U}) + \epsilon_Q + \rho_{QI}\right\}. \tag{34}$$

A complementary stagewise decomposition is

$$L_I \le \min\{1,\ L_S + \rho_{SQ} + \rho_{QI}\},$$
$$L_S = \frac{1}{m}\sum_j (1 - z_j(S)), \qquad \rho_{SQ} = \frac{1}{m}\sum_j z_j(S)(1 - z_j(Q)). \tag{35}$$

Proof. For every j, a failed image constraint either was not entailed by Q or was entailed but not rendered. Consequently,

$$L_I \le 1 - \frac{1}{m}\sum_j z_j(Q) + \rho_{QI} \le 1 - \widehat{A}_Q + \epsilon_Q + \rho_{QI}.$$

By Equation (18), $\widehat{B}_{SQ} \le \widehat{A}_Q$. Since $1 - \widehat{D} \le 1$ and $\widehat{A}_S \le 1$,

$$\widehat{U} \le \widehat{M} \le \frac{1 + 2\widehat{A}_Q}{3}, \qquad 1 - \widehat{A}_Q \le \frac{3}{2}(1 - \widehat{U}).$$

Combining these inequalities and $L_I \le 1$ gives Equation (34). For the stagewise bound, the following inclusion holds separately for each constraint:

$$\{z_j(I) = 0\} \subseteq \{z_j(S) = 0\} \cup \{z_j(S) = 1, z_j(Q) = 0\} \cup \{z_j(Q) = 1, z_j(I) = 0\}.$$

Taking probabilities, summing over j, and dividing by m proves Equation (35). No independence assumption between failure events is used. □

The loss ρ_{QI} is measured over all request constraints, not only over the subset entailed by Q. If a separate conditional rendering-failure rate on that subset is used, it must be multiplied by the subset's fraction. The stagewise loss ρ_{SQ} and the transmission score use the same fixed request denominator, but are not identical quantities. When the ideal entailment scores equal $z_j(S)$ and $z_j(Q)$, Proposition 2 gives $A^*_S - B^*_{SQ} = \rho_{SQ} + \kappa^*_{SQ}$, where $\kappa^*_{SQ} = m^{-1}\sum_j z_j(S)z_j(Q)(1 - t^*_j) \ge 0$. A branch change can reduce transmission credit even if each output individually satisfies a permitted alternative. The factor $3/2$ in Equation (34) cannot be reduced using only the stated score relationships: take one constraint, perfect evaluation, $z_1(S) = 1$, $z_1(Q) = z_1(I) = 0$, and $D = 0$. Then $B_{SQ} = 0$, $U = 1/3$, $L_I = 1$, and both ϵ_Q and ρ_{QI} are zero. This is consistent with an omission rather than a contradiction in Q.

These bounds concern the expected *fraction* of failed frozen request constraints, not the probability that an entire image satisfies every constraint. They can be vacuous when semantic coverage is low, evaluator error is large, or the renderer is unreliable. No empirical values of ϵ_Q, ρ_{QI}, or the ideal component errors are asserted here. The prompt-coverage bound in Equation (33) is at least as tight as its $\widehat{U}$-based relaxation. Thus the analysis does not provide a tighter image-failure certificate than prompt-only coverage, or establish that S is necessary for high-quality generation. The role of the additional terms is to expose plan and transmission errors; whether this supervision improves optimization must be assessed empirically.

From compiler constraints to independent reference constraints. The fixed-set qualification is essential. Suppose an independently defined reference contains n equally weighted constraints and the m compiler constraints are a correctly matched, nonduplicated subset with the same satisfaction predicates. Set $\alpha = m/n$ and let b_C denote the right-hand side of Equation (34). Splitting reference failures into scored and unscored requirements gives

$$L_{\text{reference}} \leq (1 - \alpha) + \alpha b_C. \tag{36}$$

Indeed, the scored subset contributes at most αb_C, and the unscored subset contributes at most its mass $1 - \alpha$. This statement does not apply by substituting a recall statistic computed at a different constraint granularity or in the presence of incorrectly compiled requirements. It explains why an independent compiler audit is needed even when the frozen-set semantic score is high; no value of α is assumed by the theoretical analysis.

C SpecPE Evaluation Protocol

This appendix documents the experimental protocol and the interpretation of the reported measurements. The final corpus size and its training/evaluation split are stated separately from benchmark performance. Numerical comparisons are descriptive point-estimate differences; request-level outputs, seed-specific results, and confidence intervals are not reported. Protocol settings and target annotation counts describe the design, while execution manifests and effective counts are separate reproducibility records. Controls without reported outcomes are identified explicitly.

Benchmark versions. The main text evaluates model performance on the public T2I-ReasonBench benchmark and the original GPT-4o-based WISE evaluation. These benchmark evaluations are distinct from the synthetic corpus's internal evaluation split. WISE_Verified uses a different prompt release and binary evaluator and is evaluated separately in Appendix H.2; scores are never pooled across releases. Reproduction requires the prompt files, evaluator checkpoints, templates, and aggregation scripts for each release. Any contamination audit must cover all benchmark prompt releases used in evaluation.

C.1 Shared Data and Configuration

SpecPE-CoT contains a final corpus of 32,599 synthetic demonstrations, with 90% allocated to training and 10% to an internal evaluation set. The latter supports evaluation during training; it is not the test set used for the headline performance results. Those results use the public benchmarks described above. The corpus size refers to the final available dataset, rather than the number of pre-filter candidates or the number of examples consumed by every training stage. The stage-specific SFT and RL selections are drawn from the training portion, with their own sample counts. Exact integer partition counts and sample identifiers belong in the split manifest and are not inferred from the rounded allocation ratio. This data-role separation clarifies the evaluation design; benchmark-overlap checks are a separate data audit, summarized in Appendix F.

The controlled-comparison protocol uses common training requests and teacher-produced information for learned PE methods. Native reasoning–prompt baselines serialize the shared rationale and prompt targets; the prose-plan control and SpecPE additionally serialize the same scene commitments as prose or canonical JSON. Conversion from canonical commitments to prose is deterministic and preserves entity identities, attribute bindings, relation direction, quantities, and quoted strings. This conversion introduces no new teacher calls or scene facts. The RL configuration specifies a shared, fixed, domain-stratified subset of at most 10,800 requests from the training portion; this subset is distinct from the full 32,599-example corpus and its internal evaluation set.

The optimization settings are specified in Appendix A.5. The SFT configuration uses one pass and a global batch of 128 sequences, with a 4,608-token total cap (at most 512 request tokens and 4,096 completion tokens); its preprocessing rule excludes overlength targets rather than silently truncating them. RL uses 16 requests per global rollout batch, eight complete trajectories per request, a 4,096-token cap on generated policy text, and a 256-policy-token final-prompt ceiling. The prompt-validity indicator f_q is one exactly when the extracted Q is non-empty after trimming and contains at most 256 tokens under that policy's tokenizer; otherwise $f_q = 0$ and the method-specific validity gate fails. Invalid or overlength outputs are scored as failures. The same limits apply to every method. AdamW

uses $(0.9, 0.95)$ momentum coefficients and weight decay 0.01; the reference-policy KL coefficient is 0.02 and clipping is 0.2. The protocol lists training seeds 17, 42, and 123. This seed list specifies the intended runs; the aggregate tables do not identify the completed seeds or provide their individual outcomes.

The reproducibility manifest should associate every checkpoint and per-request output with immutable backbone and generator revisions; benchmark release hashes; schema, constraint-extractor, and judge hashes; renderer steps, guidance, resolution, and scheduler; optimizer implementation and precision; hardware and software versions; and prompt templates. The comparison protocol fixes one renderer configuration per generator across methods, with configuration choices made before benchmark evaluation. These are the manifest's required contents, rather than a claim that all run-specific values are supplied here.

Complementary roles of the comparisons. RePrompt-adapt and PromptEnhancer-adapt test competitiveness against recent PE methods. The prose/JSON factorial holds the semantic rubric shared and separates representation from reward choice. Comparing SpecPE with JSON full shared score then evaluates the added deterministic checks, while transmission and contradiction ablations examine the corresponding components. Transfer and cost measurements assess practical behavior across the tested renderers and resource settings. These comparisons address different questions; their combined interpretation does not require treating an unreported control as an observed result.

C.2 Baseline Adaptations

W/O., Text-Only, and Interleaved. These controls use the same PE backbone, request split, final-prompt supervision, and SFT/GRPO schedule. W/O. maps P directly to an enhanced Q without a rationale or scene plan. Text-Only emits (T, Q) before generation. Interleaved uses one fixed refinement round: it emits (T_0, Q_0), renders a draft $I_0 \sim G(\cdot \mid Q_0)$, and then produces (T_1, Q) conditioned on the request, preceding text, and draft image. The draft is not a scored test output, and there is no adaptive best-of-N selection. SFT uses the shared teacher final prompt, with rationale targets retained only for reasoning controls; interleaved draft trajectories are prepared with a frozen Text-Only initializer and the frozen renderer, with its extra preparation cost requiring separate accounting. The draft prompt Q_0 is also nonempty and limited to 256 policy tokens. The two policy stages share the 4,096-token generated-text budget. Their generated tokens receive the same final trajectory reward and advantage, with loss normalized by their combined token count. Second-stage likelihoods condition on I_0; input tokens and image embeddings are excluded from the loss.

All three controls use image feedback without SpecPE's scene reward. Their native validator checks the required delimiters and nonempty fields, rather than imposing a JSON schema. With its binary validity g_{native}, the objective is $0.05 g_{\text{native}} + 0.95 g_{\text{native}} R_{\text{image}}$. Invalid trajectories receive zero reward. Every Interleaved draft and reward image counts toward the common RL rendering budget. At test time, its single draft is followed by four scored final images; the other controls render four scored images directly. Final-image sampling is paired, but end-to-end inference cost is not equalized; total calls and measured latency require separate reporting.

RePrompt-adapt. The adaptation follows the published reasoning–rewriting policy with a frozen generator (Wu et al., 2026). Its image reward includes ImageReward and a VLM semantic score; its format term validates the reasoning and prompt delimiters, and its length term retains the published 15–77-token preference. Published reward normalization and coefficients are fixed in the baseline configuration. The common 256-token ceiling is an external generation cap, not a replacement for this length preference. The adaptation changes the PE backbone, request pool, shared optimizer configuration, and evaluation renderer; it does not add SpecPE's intermediate semantic reward. Results are labeled an adaptation, not a reproduction of the original paper's reported scores.

PromptEnhancer-adapt. This adaptation fixes the CVPR 2026 version of PromptEnhancer (Wang et al., 2026), preserving CoT SFT followed by GRPO with its AlignEvaluator semantic image score. Following Eq. (1) of that version, the evaluator receives the generated image and rewritten prompt, $\mathrm{AlignEvaluator}(I, Q)$; this differs from SpecPE's request-conditioned image evaluation. The manifest must identify the evaluator's published reduction and immutable checkpoint. The common data and backbone replace the original policy training configuration. The evaluator is not

replaced silently by SpecPE's judge. The baseline configuration identifies the released or documented evaluator implementation; substituting a different reward defines a separate adaptation.

Prose and JSON plans with a shared score. The prose policy emits (T, L, Q), where L is a declared scene plan separated from the rationale by explicit delimiters. Its evaluator receives $C(P)$, L, and Q, with exact text-span evidence for L. The JSON policy emits (T, S, Q) and supplies JSON-pointer evidence for S. Both use the same request-level entailment, transmission, and contradiction rubric, excluding T. In both controls $d_{\mathrm{det}} = 0$, so $D_{\mathrm{shared}} = \max\{d_{\mathrm{req}}, d_{\mathrm{cross}}\}$; the typed internal-contradiction detector is not applied to either plan. Consequently, no learned prose-to-JSON conversion or unreported prose detector is needed to define the shared score.

The prose gate requires correctly nested unique delimiters, a non-empty plan, and a non-empty final prompt within the 256-policy-token ceiling. Its format score is $R_f = 0.2f_t + 0.6f_L + 0.2f_q$ and its gate is $g = f_t f_L f_q$, where f_L checks a non-empty prose-plan block and f_t, f_q retain the tag and prompt checks. Thus both representations have format score one exactly when their gate passes. The JSON control uses the strict canonical-schema gate with the same prompt ceiling. The format validators necessarily differ by representation, and their validity rates require separate reporting. Semantic supervision, image feedback, teacher information, and output budgets otherwise match. SpecPE restores d_{det} to the JSON configuration. Within the specified controls, comparing it with JSON plan (shared score) targets the detector's contribution, while the two shared-score plans compare representation without that detector. This comparison retains representation-specific serialization, token lengths, and validators, so it cannot isolate an intrinsic JSON advantage.

C.3 Rendered-Image and Optimization Budgets

The primary training comparison allocates $B = 80{,}000$ actual renderer invocations per run, with checkpoints at 20,000, 40,000, and 80,000 invocations. The learning-rate schedule is fixed to 1,250 rollout updates (160,000 sampled trajectories at the specified global batch), rather than retrospectively rescaled to a method-dependent stopping point. A request for a rendered image counts against the budget even if the rendering fails. Validation is stage-wise: invalid output stops subsequent rendering, while already issued calls, including an Interleaved draft, remain charged. Text-generation and optimization costs must be recorded in addition to rendering cost. A common safety cap of 160,000 sampled trajectories prevents a model with persistent invalid outputs from obtaining unlimited text-only optimization. A run that reaches this cap before the image budget is reported as incomplete at that budget, with the observed failure rate, rather than assigned an extrapolated score.

Each valid trajectory receives one independently seeded final reward image; Interleaved additionally receives one draft image before refinement. Infrastructure retries count as additional renderer invocations. At a budget boundary, an incomplete rollout batch is retained for logging and does not enter an optimizer update. The cost ledger must disclose all associated calls, including retries and discarded-group calls; the number of trajectory slots is not asserted to bound the number of renderer invocations. No further retry is issued after the image-call cap is reached. The same bookkeeping applies to every method. The primary endpoint uses the last completed update within the common actual image budget, not the checkpoint with the best test score. Intermediate checkpoints describe learning efficiency without selecting a winner retrospectively.

The format-only and semantic-only factorial arms follow the same rendering schedule for valid completions, but neither their image outputs nor image scores enter rewards, model inputs, sample filtering, or checkpoint selection. This controlled comparison intentionally spends the same image budget to identify the effect of observing image feedback. Omitting this diagnostic rendering would lower operational cost, but no numerical saving is inferred without measured costs. That operational variant is distinct from the matched-rendering control.

The reported comparison addresses performance under a common rendering budget. A secondary matched-optimizer-step comparison would answer a different question: performance at equal optimization exposure, with rendering costs allowed to differ. Its outcomes and budget-checkpoint curves are not reported, so the present efficiency claim is restricted to the stated resource comparison. The inference configuration specifies one enhancement trajectory per request, greedy decoding, a 4,096-token total generated-text ceiling, and four common final-image seeds (101, 211, 307, and 401). Interleaved has two policy stages and a separately specified draft seed (53). These seed settings

should be linked to the actual output records in the execution manifest. Four final images estimate one trajectory's expected rendering performance; they are not a best-of-four search budget. The raw-prompt generator reference uses the same image settings, with no PE call.

C.4 Endpoints and Statistical Analysis

The primary endpoint specified for evaluation is official T2I-ReasonBench overall accuracy. Its quality score and original WISE overall score are secondary endpoints; category-level results diagnose heterogeneity. Official aggregation is applied, including benchmark-prescribed sample weights rather than an assumed unweighted category mean. A malformed or missing final prompt is a failure for the corresponding test request, not a reason to remove it. A valid rendering that fails semantic criteria is scored normally. Infrastructure failures are retried once with the same input and seed, and persistent failures remain explicitly missing with a coverage count; any affected result requires both complete-case and worst-case sensitivity summaries. No missing case is silently dropped or converted into a claimed completed measurement.

For replicated runs, the aggregation protocol first averages the four images for each request within a training seed, then applies the official benchmark aggregation and averages across completed runs. The completed run count and per-seed outcomes must be reported alongside such an average; they cannot be inferred from the configured seed list. Paired differences use identical requests and image seeds. To quantify their uncertainty from existing output records, a paired bootstrap with 10,000 replicates would resample *requests as clusters*, retaining every image and method output for each sampled request and preserving official category weights. Its 95% interval estimates test-request variation conditional on the measured runs. Training variability is a separate quantity, summarized by per-seed scores and their standard deviation.

The core comparisons are JSON versus prose under the full shared score, SpecPE versus JSON full shared score, and SpecPE versus PromptEnhancer-adapt, the strongest reported external PE baseline on overall T2I-ReasonBench accuracy for each backbone. They address representation, the added detector, and practical competitiveness, respectively. The factorial additionally compares full with image-only reward and reports their difference-in-differences across representations. These comparisons already have reported point estimates; uncertainty analysis would establish the precision of those estimates rather than add a missing baseline. If formal significance tests are conducted, the tested family must be declared and multiplicity handled, for example with Holm adjustment. No such intervals, seed dispersion, or adjusted tests are reported here. Category breakdowns and factorial interactions are therefore interpreted descriptively. Absolute benchmark-score differences use points; relative changes divide the score difference by the stated comparator. Percentages attached to diagnostic rates retain their original population denominators.

Scope of the cost control. An 80,000-call rendering budget controls one expensive resource, not total computation. In the idealized case of valid trajectories and no retries, a 128-trajectory batch costs 128 final renders for SpecPE and 256 draft-plus-final renders for Interleaved. The budget then permits at most 625 and 312 complete updates, respectively, before partial-batch handling. These are arithmetic implications of the protocol, not observed update counts. Invalidity, discarded groups, and retries change the actual counts. Without execution records and matched-step results, the reported comparisons do not establish equality of optimization exposure or overall GPU cost.

D Factorial Reward and Representation Experiment

Table 4 and Section 5.3 present the factorial results and mechanism analysis. Here we specify the exact objectives and component controls. The 2×4 factorial uses a shared semantic score in both representations. Let A_{plan} denote A_S for JSON and the identically defined request-coverage score for prose L. With transmission $B_{\text{plan},Q}$ defined against the same frozen requirements, set

$$U_{\text{shared}} = \frac{A_{\text{plan}} + A_Q + B_{\text{plan},Q}}{3} \left(1 - \max\{d_{\text{req}}, d_{\text{cross}}\}\right). \tag{37}$$

Thus $d_{\text{det}} = 0$ in every factorial arm, including JSON. Let $R_f \in [0,1]$ and $g \in \{0,1\}$ be the representation-appropriate format score and gate; JSON requires the strict schema, and both require

a non-empty final prompt within its token ceiling. Write $R_i = R_{\text{image}}$. The four objectives are independently specified as

$$R^{\text{F}} = R_f, \tag{38}$$

$$R^{\text{F+U}} = 0.05R_f + 0.95gU_{\text{shared}}, \tag{39}$$

$$R^{\text{F+I}} = 0.05R_f + 0.95gR_i, \tag{40}$$

$$R^{\text{Full}} = 0.05R_f + 0.95g(0.3U_{\text{shared}} + 0.7R_i). \tag{41}$$

No arm sends T to a semantic evaluator. The F + U and F + I arms retain full normalized weight for their available substantive signal; they are not numerical zero-term substitutions into R^{Full}. The SFT-only anchors undergo no GRPO updates. Normalized reward-component means, standard deviations, gate rates, and zero-advantage group rates are needed to compare signal scale and informativeness. Only semantic-component means are reported among these training diagnostics.

The 2×4 experiment and both SFT anchors use Qwen3-VL-8B with FLUX.1-dev. Four Qwen2.5-VL-7B arms—prose and JSON with image-only and full shared-score rewards—are specified as a replication of the key contrasts; their results are not included, so the measured factorial comparison is limited to Qwen3-VL-8B. The prose and JSON full shared-score arms use the same supervision and budgets as the main comparison. A separate *JSON Full + detector* arm replaces U_{shared} with the method's U, restoring $D = \max\{d_{\text{req}}, d_{\text{cross}}, d_{\text{det}}\}$. This is SpecPE and reuses its main-table run for each backbone. It is not substituted for JSON full shared score in the factorial interaction. The full-versus-single-signal contrasts describe the value of the chosen mixture relative to each normalized single-signal objective. Because their weights differ, they do not isolate adding one reward while holding the coefficient of the other fixed. The difference-in-differences separately describes representation–reward interaction on the accuracy scale.

Two further controls modify SpecPE, not the shared-score factorial. The transmission-free version uses $U_{-B} = \frac{1}{2}(A_S + A_Q)(1 - D)$ with the full D including d_{det}. The penalty-free version uses $U_{-D} = \frac{1}{3}(A_S + A_Q + B_{SQ})$, removing all three contradiction penalties. Each replaces only U in the formal method's R_{full} and remains on $[0, 1]$. Transmission retains its fixed request denominator, $B_{SQ} = |C(P)|^{-1} \sum_j e_j(S)e_j(Q)t_j(S, Q)$; optional detail earns no extra credit. Non-deterministic judge disagreement about an optional detail does not become a hard veto. The controls are compared with SpecPE through the reported semantic-component summaries. A separate diagnostic set containing mismatched valid alternatives and contradictory commitments is needed to determine whether the effects generalize beyond the reward evaluator.

What the existing ablations identify. The factorial compares semantic and image supervision within explicit-plan policies; component removals further test transmission and contradiction handling. Together they support the utility of the tested plan-based reward design. They do not isolate whether equally weighted, request-conditioned semantic feedback on Q alone could obtain the same result: no same-configuration Q-only constraint-reward outcome is reported. The analysis therefore concerns the observed benefits of SpecPE's components, without claiming that an intermediate plan is necessary for every effective PE reward.

Complementary supervision across representations. The full reward gives the highest shared-score accuracy for both representations, suggesting that semantic and image feedback provide complementary supervision. Relative to image-only training, Full improves the accuracy score by 1.1 points for prose and 2.1 points for JSON, corresponding to relative improvements of 1.6% and 3.0%. The JSON-minus-prose gap consequently widens from 0.2 points under image-only training to 1.2 points under Full, giving a descriptive difference-in-differences of 1.0 point. Full also exceeds semantic-only training by 1.4 points for prose and 1.8 points for JSON, or 2.0% and 2.5% relative to those respective controls. Thus neither single signal reaches the full mixture's accuracy in this summary. This pattern is consistent with complementary supervision, but a positive point contrast alone does not establish a statistically reliable interaction.

The component comparisons align with the failures each term is designed to address. Adding the detector to JSON Full raises the accuracy score by 0.67 points, a 0.9% relative improvement, while lowering quality by 0.20 points. Removing transmission from SpecPE lowers accuracy by 1.67 points, or 2.3% relative to SpecPE: plan coverage stays at 93%, but final-prompt coverage falls from

91% to 87% and transmission from 88% to 81%. Removing contradiction penalties lowers accuracy by 1.27 points, a 1.7% relative decrease, and raises the contradiction score D from 0.03 to 0.08, while quality increases by 0.10 points. These quality tradeoffs rule out uniform improvement on all metrics. The aggregate associations connect reward components to their intended failure modes without independently validating the judge or establishing causal mediation.

The same progression appears in the SpecPE reward series: adding semantic feedback to format feedback raises the accuracy score by 2.23 points for Qwen2.5-VL-7B and 3.45 points for Qwen3-VL-8B, corresponding to relative improvements of 3.2% and 5.1%. Adding image feedback then raises accuracy by another 1.89 and 2.22 points, or 2.6% and 3.1% relative to the respective semantic-plus-format controls. These increments compare normalized reward mixtures rather than additive effects at fixed coefficients. The reported A_{plan}, A_Q, $B_{\text{plan},Q}$, and D satisfy their stated bounds, including $B_{\text{plan},Q} \leq \min(A_{\text{plan}}, A_Q)$. They are aggregate reward diagnostics, not independently adjudicated human accuracy. Mean semantic reward cannot be reconstructed by multiplying the displayed means because, in general, $\mathbb{E}[M(1-D)] \neq \mathbb{E}[M](1-\mathbb{E}[D])$. The shared-score factorial omits deterministic contradiction checks, whereas the SpecPE component series includes them unless explicitly removed; this distinction must be retained when comparing the results.

E Semantic Diagnostics and Evaluator Reliability

Table 6 reports stage-specific measurements, interface interventions, and evaluator checks for Qwen3-VL-8B with FLUX.1-dev. The protocol below defines their interpretation. Sample sizes describe the specified design; effective counts after eligibility screening and any aggregation across runs are not available in the summary table and are not inferred from rounded rates.

E.1 Stage-Specific Constraint Measurements

The diagnostic design targets a separately defined sample of 400 held-out requests, stratified across eight source domains at the scene-cluster level. This target does not define an additional corpus partition: the reported corpus split remains 90% training and 10% internal evaluation, while diagnostic sample membership and its relationship to that split are not specified in the aggregate results. The annotation protocol assigns two annotators to independently identify request-supported entities, attributes, directed relations, counts, exact strings, and scene facts, with an adjudicator resolving disagreements. Reference constraints are to be written independently of the training compiler. Comparing them with $C(P)$ measures extractor precision and recall, including omitted requirements that the training reward cannot penalize.

For the same requests, blinded annotators label each reference constraint in the plan, final prompt, and image as satisfied, omitted, contradicted, or not visually assessable. The image denominator excludes a constraint only if it was designated not visually assessable before inspecting method outputs. Additional compatible detail is distinguished from a contradiction. Stage satisfaction rates alone do not identify where a particular commitment was lost: the analysis must retain paired clause labels and separately count plan-to-prompt and prompt-to-image transitions. Plan–prompt agreement over declared commitments is a different quantity from B_{SQ}, whose denominator is the frozen request-constraint set.

Unconditional scores retain malformed outputs as failures. A conditional comparison may use the shared subset where both methods pass their respective gates, but must report the retained count and the difference in gate rates. Conditioning on validity can select a different population; it neither replaces the main endpoint nor by itself removes all formatting-related confounding. No completed paired-validity analysis is present in the supplied record, so the current evidence does not establish that the entire benchmark gain persists within this subset.

E.2 Field Interventions and Negative Controls

The primary intervention design targets 200 eligible diagnostic requests; this is a design target, not the effective annotated count. Edits select an alternative explicitly allowed by the original P, or alter a visually assessable attribute left unconstrained by P. A coupled alternative changes as one complete branch. Eligibility and edit construction must be established from the request and schema before

inspecting policy responses. The original request, preceding rationale, and unrelated plan content remain fixed; only the plan prefix changes before regenerating Q with paired decoding and rendering seeds.

Each eligible case includes an edited plan, an unedited control, and a meaning-preserving paraphrase or harmless serialization control. Measurements are edit realization in Q and I, preservation of original constraints, and stability under the harmless control. All eligible cases remain in the denominator, including invalid prefixes and output failures. Equivalent prose and JSON interventions compare the exposed interfaces, with SpecPE assessed separately. These interventions could establish behavioral use of a declared plan; they cannot certify faithful hidden reasoning.

A separate conflict stress test changes a requested commitment while retaining the original P and T as policy input. A counterfactual request P' describes the edit for scoring only. Counterfactual realization and original-request faithfulness must be shown together: rejecting a contradictory edit may be the correct behavior. Such cases cannot be pooled with request-compatible interventions. A further evaluator control holds P and Q fixed while corrupting one field in S and reuses the existing image. Image feedback is then unchanged by construction; only a measured change in the relevant semantic labels would demonstrate that the implemented judge detects the corruption. Without separate conflict-stress or fixed-Q corruption results, the reported request-compatible interventions and invariance controls cannot establish the theoretical blind spot empirically.

E.3 Judge Calibration and External Checks

The judge-calibration specification targets 400 requests, with scene-cluster disjointness from the diagnostic sample as a design requirement. It allocates half to prompt and label-map calibration and half to evaluation of the frozen judge; actual membership and effective counts are separate audit records. Faithful completions, omissions, contradictions, and meaning-preserving rewrites require verified clause-level reference labels. The evaluation measures are label precision/recall, macro F1, agreement, aggregate-score error, and errors by constraint type on the held-out labels.

Robustness controls change key order, consistently rename entity IDs, and vary prose verbosity without adding facts. Varying a rationale withheld from the judging API tests an input-isolation invariant, not semantic reasoning by the judge. The request compiler must remain frozen throughout. An evaluator specified to be from a different model family and blinded human preferences on a target of 200 benchmark requests provide the external comparisons reported in the table. Annotators receive the original request and shuffled images without method names or rationale text. The summary gives preference rates; exact evaluator revisions, effective request counts, and paired outcome records are required for reproducible uncertainty estimates.

Constraint fidelity across evaluation stages. The stage measurements distinguish strong textual fidelity from the remaining losses in extraction, rewriting, and rendering. Extraction reaches 95.2% precision and 91.8% recall against the human reference. A high downstream reward therefore remains conditional on an imperfect requirement set, with some human-reference requirements absent from that set. Plan satisfaction is 92.3%, with omission and contradiction rates of 5.2% and 2.5%; final-prompt satisfaction is 90.1%, with corresponding rates of 7.1% and 2.8%. The 2.2-percentage-point satisfaction gap coincides with 1.9 percentage points more omissions and 0.3 percentage points more contradictions in the prompt summary. This pattern is consistent with residual rewriting loss, although aggregate stage rates cannot identify which clauses changed status without paired labels. Image realization reaches 76.2% on visually assessable constraints, indicating incomplete end-to-end fulfillment. This marginal rate does not isolate renderer failures: its applicability denominator can differ, so subtracting it from prompt satisfaction would not estimate a paired prompt-to-image failure rate. Likewise, 94.4% plan–prompt agreement does not imply perfect transmission of the user's requirements; it measures declared commitments, whereas B_{SQ} is request-anchored.

Plan interventions and invariance controls. Request-compatible edits are usually reflected in the rewritten prompt and often in the resulting image, supporting behavioral use of the declared interface. Edit realization reaches 88.7% in prompts and 70.4% in images, while unaffected constraints are preserved at 94.5% and 81.6%, respectively. The harmless controls yield 96.8% stability under paraphrases and 99.1% under serialization changes. These outcomes are consistent with use of the interface and stability under the tested controls, while edit realization and constraint preservation

Table 6: Stage, intervention, and judge-validation measurements for SpecPE with Qwen3-VL-8B and FLUX.1-dev, followed by independent pairwise comparisons. Rates and normalized errors use $[0, 1]$. Confidence intervals and effective sample counts are not reported in this table.

Measurement	Value
Constraint fidelity and semantic transmission	
Request-constraint extraction precision / recall	0.952 / 0.918
Plan constraint satisfaction / omission / contradiction	0.923 / 0.052 / 0.025
Final-prompt constraint satisfaction / omission / contradiction	0.901 / 0.071 / 0.028
Plan–prompt agreement on declared commitments	0.944
Image realization of visually assessable constraints	0.762
Interventions and invariance controls	
Intervened-field realization: prompt / image	0.887 / 0.704
Unaffected-constraint preservation: prompt / image	0.945 / 0.816
Paraphrase / serialization-control stability	0.968 / 0.991
Judge validation	
Entailed-label precision / recall	0.941 / 0.956
Unknown-label precision / recall	0.889 / 0.862
Contradicted-label precision / recall	0.917 / 0.894
Clause-label macro F1 (derived from the three rows above)	0.910
Judge–reference agreement / normalized score MAE	0.913 / 0.041
External comparisons: win / tie / loss	
Independent evaluator: JSON shared vs. prose shared	0.462 / 0.244 / 0.294
Blinded-human evaluation: JSON shared vs. prose shared	0.435 / 0.285 / 0.280
Independent evaluator: SpecPE vs. JSON shared	0.397 / 0.276 / 0.327
Blinded-human evaluation: SpecPE vs. JSON shared	0.365 / 0.321 / 0.314

Plan–prompt agreement uses declared commitments; it is not B_{SQ}, whose denominator is the fixed request-constraint set. In each pairwise row, win means preference for the first method. Human rows report blinded image preferences. Image realization uses only visually assessable constraints, so its denominator differs from text-stage satisfaction.

remain incomplete. They do not establish simultaneous satisfaction of all original constraints or faithful hidden reasoning. The summaries also do not identify a conditional rendering-loss rate or the causal change relative to unedited outputs; those estimates require matched outcomes and explicit aggregation and stability criteria. Without matched prose interventions, the results cannot establish that this control is unique to JSON.

Evaluator agreement and uncertainty handling. The evaluator agrees with the reference on 91.3% of labels, with normalized score MAE of 0.041. Its strongest performance is on entailed constraints, where precision and recall reach 94.1% and 95.6%. Unknown labels are harder to recover: recall falls to 86.2%, with precision of 88.9%, identifying uncertainty handling as a remaining weakness. Contradiction precision and recall are 91.7% and 89.4%. The three label-wise F1 scores yield macro F1 of 0.910 after rounding, consistent with the table. Contradiction precision below 100% matters because incorrect penalties can suppress otherwise useful outputs. These are average audit properties, not uniform error guarantees over all RL rollouts, and they do not supply numerical constants for the conditional theoretical bounds. The deterministic detector's binary veto of semantic credit additionally warrants its own false-positive analysis; no detector-specific confusion matrix is reported.

External preferences favor the same ordering, with different margins. Independent evaluation and blinded human judgments favor the same ordering, with a larger margin for the representation contrast than for the detector increment. For JSON shared score versus prose shared score, independent-evaluator win, tie, and loss rates are 46.2%, 24.4%, and 29.4%; the corresponding human rates are 43.5%, 28.5%, and 28.0%. These give net win-minus-loss margins of 16.8 and 15.5 percentage points. SpecPE's additional advantage over JSON shared score is smaller: the corresponding net margins are 7.0 and 5.1 percentage points, with ties in 27.6% and 32.1% of comparisons. Thus external judgments favor both contrasts in aggregate, with a more modest incremental preference for the detector. The results reduce exclusive reliance on the training reward but do not prove a causal mechanism or a statistically reliable difference without paired request-level uncertainty estimates.

Table 7: Unchanged-checkpoint transfer from FLUX.1-dev to Qwen-Image with Qwen3-VL-8B, separate from the generator setting in Table 8. Costs are PE latency per request (seconds) and per-run SFT+RL GPU hours, excluding data construction. PE latency excludes final rendering. Appendix G describes cost accounting.

Method	Qwen-Image		Compute cost	
	Acc.	Qual.	PE latency (s)	Training (GPU h)
RePrompt (adapt.)	77.6	93.1	2.7	104
PromptEnhancer (adapt.)	79.9	94.2	3.2	116
Prose plan (shared score)	80.6	94.4	4.1	151
JSON plan (shared score)	82.1	94.3	3.8	143
SpecPE	83.4	94.1	3.9	158

The stage table reports SpecPE's absolute fidelity; it contains no baseline stage rates from which to claim fewer human-annotated omissions for every competing method.

F Synthetic-Data Audit

The data-audit specification targets 800 demonstrations, 100 per source domain, selected independently of reward scores. Two annotators and an adjudicator assess whether P justifies the required commitments in S and Q, whether image-derived structure agrees with the intended scene, and whether the request, plan, and prompt contradict one another. Optional stylistic detail and unsupported reasoning claims are recorded separately. The target sample size is distinct from an effective completed annotation count; the latter and the adjudication records are not supplied in the aggregate results.

The acceptance protocol preserves the intended request rather than treating a generated image as ground truth. A candidate with a recoverable scene description may be reserialized and checked against the request and final prompt; a candidate with unresolved contradictory or unsupported commitments is ineligible. Such preprocessing is distinct from repairing a policy completion during reward evaluation, which is disallowed. Appendix I illustrates the relevant failure types. Their prevalence and the numbers repaired or excluded require construction records; neither is inferred from the final corpus size.

The confirmed dataset facts are the final size of 32,599 and the 90:10 training/internal-evaluation allocation. An overlap audit is a separate check: exact-text hashing and semantic retrieval should cover each benchmark release, group related scenes and paraphrases, fix matching thresholds, and adjudicate flagged matches. The audit report would provide the corresponding overlap and exclusion counts, along with split identifiers and grouping rules. These outcomes are not reported here, so the use of public benchmarks is not itself presented as evidence of zero contamination. Likewise, no measured filtering benefit is claimed without a filtered-versus-unfiltered comparison.

G Checkpoint Transfer and Computational Cost

The transfer comparison in Table 7 trains PE with FLUX.1-dev and reuses the same policy checkpoints with Qwen-Image, changing the renderer and its documented settings. Baselines follow the same transfer protocol with common requests, PE decoding, and paired image seeds. This differs from the additional Qwen-Image setting in Table 8; target-generator RL, if used, is a separate adaptation with an additional training budget. A reproducible transfer record must identify the source and target checkpoint hashes so that unchanged reuse can be verified.

The accuracy advantage persists under checkpoint transfer. Reusing the FLUX-trained policy with Qwen-Image preserves SpecPE's leading accuracy score among the tested methods without further PE training. Its score of 83.4 exceeds JSON shared score by 1.3 points and prose shared score by 2.8 points, corresponding to relative improvements of 1.6% and 3.5%. Against the external PE baselines, it exceeds PromptEnhancer-adapt by 3.5 points and RePrompt-adapt by 5.8 points, or 4.4% and 7.5% relative to those respective baselines. Quality remains a tradeoff: SpecPE's score of 94.1 is 0.2 points below JSON shared score, 0.3 below prose shared score, and 0.1 below PromptEnhancer-adapt, while exceeding RePrompt-adapt by 1.0 point. The retained accuracy ordering supports

portability of the learned enhancement policy to this tested renderer; it does not establish generator-independent performance or uniform quality superiority. The 83.4 transfer score and 85.77 in the separate Qwen-Image setting do not identify a training benefit without a matched adaptation comparison.

Accuracy comes with a measurable cost tradeoff. SpecPE's accuracy advantage comes with measurable differences in source-checkpoint training cost and PE latency; transfer itself adds no further PE training. Reported PE latency is 3.9 seconds per request, compared with 3.8 for JSON shared score, 4.1 for prose shared score, 3.2 for PromptEnhancer-adapt, and 2.7 for RePrompt-adapt. Its original SFT+RL training takes 158 GPU hours, which is 15 hours above JSON shared score (10.5%) and 7 hours above prose shared score (4.6%). Against JSON shared score, the latency increase is 2.6%. Against PromptEnhancer-adapt, the 3.5-point accuracy gain (4.4% relative) accompanies 0.7 additional seconds of PE latency (21.9%) and 42 additional GPU hours (36.2%). SpecPE is therefore not uniformly cheapest, and its accuracy lead does not establish Pareto dominance when quality is included. Its latency is 0.2 seconds lower than the prose-plan control's; explaining this gap requires token counts and timing variability.

The table reports PE-only latency and SFT+RL GPU hours, excluding data construction. These are not end-to-end generation costs. A complete cost account also separates renderer calls, semantic/image-judge calls, initial constraint extraction, reward calibration, generated policy tokens, discarded completions, and optimizer updates. Cached constraints remain an initial computation cost. Cross-system comparisons require hardware type, precision, batch size, concurrency, warmup, and the latency aggregation statistic; latency tails and peak memory are unreported.

For Interleaved, pipeline latency includes two policy stages and a draft render; its absence from the timing table prevents a measured end-to-end latency comparison with SpecPE. The benchmark protocol generates four final images per request to estimate performance, whereas a single delivered image normally requires one final render. Fixed renderer-call budgets control one resource and do not establish equal total computation. Without budget-checkpoint curves or matched-step outcomes, these results cannot rule out unequal optimization exposure.

H Additional Generator and Benchmark Results

H.1 Qwen-Image and Direct Generation

Table 8: T2I-ReasonBench with Qwen-Image and Qwen3-VL-8B. Accuracy and quality are reported by category. This generator setting is distinct from the unchanged-checkpoint transfer test in Table 7.

Method	Idiom		Entity		Scientific		Textual		Overall	
	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.
W/O.	57.72	90.71	63.75	96.54	69.91	89.92	80.68	86.33	68.02	90.88
Text-Only	61.52	92.54	67.95	96.52	75.70	92.79	87.39	**91.58**	73.14	93.36
Interleaved	72.47	94.33	78.75	96.46	80.04	**94.08**	89.08	90.67	80.08	93.89
SpecPE	**84.74**	**96.08**	**85.91**	**97.71**	**82.24**	93.83	**90.18**	91.08	**85.77**	**94.68**

The additional Qwen-Image setting reproduces the largest accuracy gains on idiom and entity reasoning. With Qwen3-VL-8B, SpecPE raises the overall accuracy score from 80.08 for Interleaved to 85.77, a 7.1% relative improvement. The category gains are 12.27 points for idioms and 7.16 points for entities, corresponding to relative improvements of 16.9% and 9.1%, followed by 2.20 points for scientific reasoning and 1.10 points for textual reasoning. Overall quality increases from 93.89 to 94.68, a 0.8% relative improvement, although scientific quality decreases by 0.25 points and Text-Only retains the highest textual quality. The second generator therefore reproduces the category-dependent pattern without establishing uniform superiority on every metric. This setting remains distinct from unchanged-checkpoint transfer, which reuses the FLUX-trained PE checkpoint without target-generator training.

Direct generation provides a broader reference for the value of the complete enhancement pipeline (Table 9). It sends the original request to the renderer, whereas the main-table W/O. baseline enhances prompts without an explicit reasoning stage; their different inputs make them separate baselines.

Table 9: Direct-generation reference on T2I-ReasonBench. Each generator uses the original request. The main-table W/O. baseline instead enhances prompts without explicit reasoning.

Method	Idiom		Entity		Scientific		Textual		Overall	
	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.
FLUX.1-dev	37.32	87.25	60.50	81.28	44.95	92.42	48.44	84.17	47.80	86.28
Qwen-Image	47.13	91.79	58.40	97.00	58.44	89.50	70.88	86.54	58.71	91.21

Table 10: Supplementary WISE_Verified with FLUX.1-dev. Overall weights Cultural by 0.4 and each other category by 0.12, rounded to two decimals. This release uses different prompts and an evaluator distinct from original WISE in Table 2; scores are not pooled. Bold marks all per-backbone maxima, including ties.

Method	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
PE backbone: Qwen2.5-VL-7B							
RePrompt (adapt.)	0.69	0.55	**0.64**	0.57	0.53	0.50	0.61
PromptEnhancer (adapt.)	0.72	0.58	0.62	**0.60**	0.56	0.53	0.63
SpecPE	**0.76**	**0.60**	**0.64**	0.59	**0.57**	**0.56**	**0.66**
PE backbone: Qwen3-VL-8B							
RePrompt (adapt.)	0.71	0.57	0.63	0.55	0.58	0.51	0.62
PromptEnhancer (adapt.)	0.73	0.59	**0.64**	0.59	**0.59**	0.52	0.64
SpecPE	**0.77**	**0.62**	0.63	**0.62**	0.58	**0.57**	**0.67**

Relative to direct FLUX.1-dev, SpecPE's overall accuracy gains are 25.42 points with Qwen2.5-VL-7B and 25.77 points with Qwen3-VL-8B, corresponding to relative improvements of 53.2% and 53.9%. Their quality gains are 5.71 and 5.52 points, or 6.6% and 6.4% relative to direct FLUX.1-dev. With Qwen-Image, the accuracy and quality gains over direct generation are 27.06 and 3.47 points, corresponding to relative improvements of 46.1% and 3.8%. These larger contrasts combine all benefits of enhancement and training, while Interleaved and the shared-score comparisons provide more specific evidence about the proposed method. Qwen-Image's direct baseline is itself 10.91 accuracy points above FLUX.1-dev, a 22.8% relative advantage, so cross-generator absolute scores should not be attributed solely to the enhancer.

H.2 Supplementary WISE_Verified Evaluation

WISE_Verified is the official maintenance release announced on 19 April 2026 (`https://github.com/PKU-YuanGroup/WISE`). It updates prompts and uses a Qwen3.5-35B-A3B evaluator with binary WiScore. Its overall reduction weights cultural common sense by 0.4 and each remaining category by 0.12. We retain the six-category breakdown and fix the release, evaluator, and reduction in the manifest. The original WISE value of 0.65 in the main text is not a WISE_Verified measurement. The supplementary release preserves SpecPE's overall advantage over the external PE baselines. Its rounded scores are 0.66 with Qwen2.5-VL-7B and 0.67 with Qwen3-VL-8B, exceeding PromptEnhancer-adapt by 0.03 with either backbone and RePrompt-adapt by 0.05. These correspond to approximate relative score improvements of 4.8% and 4.7% over PromptEnhancer-adapt, and 8.2% and 8.1% over RePrompt-adapt, respectively. The advantage remains heterogeneous: with Qwen2.5-VL-7B, biology is 0.01 below PromptEnhancer-adapt and space ties RePrompt-adapt; with Qwen3-VL-8B, space and physics are each 0.01 below PromptEnhancer-adapt. Cultural common sense is an important contributor: its score rises by 0.04 over PromptEnhancer-adapt for both backbones, contributing 0.016 to the respective weighted differences under the benchmark's 0.4 category weight. The six displayed category scores reproduce every rounded overall value under the stated weights. These summaries support a consistent direction of overall improvement across WISE releases, not uniform category superiority, a paired significance claim, or a causal effect of changing the evaluator. Because prompts and scoring both change, absolute scores across releases are not directly comparable.

I Qualitative Examples and Data Consistency

I.1 Additional WISE Comparisons

Figure 6 shows a recognizable windmill, roses, and a koala in the SpecPE column. These examples illustrate the conversion of descriptive requests into concrete scene choices, but do not establish a unique answer to every request. In particular, the description of an Australian tree-dwelling marsupial does not uniquely identify a koala. Alternatives must be judged against all benchmark requirements; differing from the SpecPE image alone does not justify rejection.

A plausible rice field likewise does not verify a specific nutrient deficiency. Lighting in the market example can illustrate a temporal interpretation, but appearance alone cannot establish the intended date and time conversion. These cases separate readily visible identity or relation evidence from knowledge claims that require an independent reference. Selected images explain possible success and failure modes; category-level benchmark results provide the aggregate comparison.



Figure 6: Knowledge-dependent image generation. WISE examples compare cultural activities, landmarks, plant and animal identity, temporal context, and scientific knowledge. Column labels follow Figure 3. Appearance alone cannot verify every knowledge-dependent condition.

I.2 SYNTHETIC-DATA CONSISTENCY

The following figures preserve raw candidate records in their original, pre-canonical format. They illustrate inconsistencies addressed by the acceptance rules in Appendix C; neither is a conforming `specpe-2.0` demonstration as displayed. Their inclusion documents failure modes in data construction and does not establish that these records entered the retained training set. Appendix A.2 gives a conforming scene specification.

The butterfly example substitutes resting on a leaf for the requested emergence event. Because that event is explicitly required by P, a description faithful to the rendered image can still be an incorrect training target. The poster example illustrates two distinct issues: the structure omits the visible wording requested in P and selected explicitly in Q, while dawn in the structure conflicts with sunset in Q. The exact-text omission is request-anchored; the dawn–sunset difference concerns an optional scene choice unless time of day is separately required. Under the reward protocol, the latter does not by itself add a new mandatory constraint or positive-credit target. Thus data consistency review can be stricter about teacher commitments than the request-anchored semantic reward.

Case2 : Nature

User prompt: A monarch butterfly emerging from its chrysalis.

Thinking Text

Part 1: Understanding
The prompt describes a literal biological event: a monarch butterfly in the act of emerging from its chrysalis—a moment of metamorphosis. This is not metaphorical but a precise natural history scene requiring knowledge of lepidopteran life cycles. The viewer must understand that monarchs develop inside a jade-green chrysalis before eclosing as winged adults. Accurate depiction demands familiarity with entomology: the butterfly's iconic orange-and-black wings, its host plant (milkweed), and typical emergence behaviors. Cultural context also matters—monarchs symbolize transformation and fragility, often featured in educational and conservation media.
Part 2: Visual Planning
The core visual subject is the butterfly mid-emergence, positioned prominently per the rule of thirds to emphasize its vulnerability and beauty. Photorealism is chosen to honor scientific accuracy while evoking emotional resonance, blending Audubon's illustrative precision with National Geographic's immersive photography. Warm golden-hour lighting from the left enhances the dewy morning atmosphere, casting soft highlights on water droplets that signify freshness and new life. The color palette centers on warm orange (butterfly) against vibrant green (milkweed and chrysalis), creating natural contrast while maintaining harmony through emerald shadows and pale yellow highlights. A macro lens captures intricate textures—matte wings, glossy leaf veins, smooth chrysalis surface—while background bokeh isolates the subjects, reinforcing focus on transformation. Including both butterfly and chrysalis in balanced midground positions visually narrates the "before and after" of metamorphosis, aligning with themes of natural cycles and tranquility.

Image



Structured Vision


```
{
  "style_name": "Monarch Transformation in Dewlight",
  "inspiration": [
    "John James Audubon's natural history illustrations",
    "National Geographic wildlife photography",
    "Botanical art of the 19th century"
  ],
  "scene": "A sun-dappled meadow with flowering milkweed plants",
  "subjects": [
    {
      "type": "monarch_butterfly",
      "description": "orange wings with black veins and white spots, water droplets on wings and antennae",
      "position": "left midground",
      "size": "medium",
      "details": "wings partially open, resting on a leaf with visible water droplets"
    },
    {
      "type": "chrysalis",
      "description": "green chrysalis with subtle brown markings, suspended from plant stem",
      "position": "right midground",
      "size": "medium",
      "details": "water droplets on surface, attached to plant stem with visible texture"
    }
  ],
  "style": "photorealistic",
  "color_palette": {
    "primary": "warm orange",
    "secondary": "vibrant green",
    "highlight": "pale yellow",
    "shadow": "deep emerald",
    "background_gradient": [
      "soft green",
      "golden hue"
    ]
  },
  "lighting": {
    "source": "natural sunlight",
    "direction": "left",
    "tone": "warm golden hour",
    "intensity": "soft and diffused"
  },
  "mood": "serene and contemplative",
  "background": {
    "type": "scenery",
    "details": "softly blurred meadow with indistinct green foliage and distant light sources"
  },
  "composition": "rule of thirds with butterfly as focal point, chrysalis balancing the frame",
  "camera": {
    "angle": "eye-level",
    "distance": "close-up",
    "lens": "macro",
    "focus": "sharp subjects, blurred background"
  },
  "medium": "photographic realism",
  "textures": [
    "water droplets on surfaces",
    "matte wing texture",
    "smooth chrysalis surface",
    "glossy leaf veins"
  ],
  "details": {
    "weather": "dewy morning",
    "time_of_day": "early morning",
    "season": "summer"
  },
  "effects": [
    "water droplets",
    "bokeh",
    "soft light diffusion"
  ],
  "themes": [
    "transformation",
    "natural cycles",
    "tranquility"
  ]
}
```


Generative Prompt

Show a monarch butterfly in the precise moment it is emerging from its jade-green chrysalis, with crumpled wings still wet and clinging to the empty casing. The scene should be set on a milkweed plant in soft morning light, with dew on nearby leaves and no human interference visible.

Readable excerpt of the conflicting commitments

Request: "A monarch butterfly emerging from its chrysalis."
Final prompt: "in the precise moment it is emerging"; "clinging to the empty casing."
Raw structure: `details`: "resting on a leaf with visible water droplets."
Implication: separate object descriptions omit the requested emergence event.

Figure 7: Raw-pool inconsistency: the intended butterfly emerges from its chrysalis, whereas the image-derived structure describes it on a leaf. The acceptance rules require repairing or excluding this record.

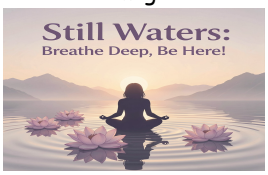
Case6 : Textual

User prompt: Create a mindful poster with the theme "Still Waters: Breathe Deep, Be Here!".

Thinking Text

Part 1: Understanding
The prompt "Still Waters: Breathe Deep, Be Here!" is a metaphorical call to mindfulness, where "still waters" symbolize inner calm and emotional stillness, contrasting with mental turbulence. The phrase "Breathe Deep, Be Here!" directly instructs present-moment awareness—breathing as an anchor to the now. This evokes Eastern philosophical traditions (especially Zen and Buddhist practices) where meditation, nature, and breath are interwoven. The core concept isn't just a scenic image but a psychological state: tranquility achieved through presence. Background knowledge in mindfulness philosophy, symbolic use of water and lotus flowers in Eastern cultures, and the significance of dawn (a time of renewal and clarity) is essential to interpret the prompt correctly.
Part 2: Visual Planning
The core visual subject is the meditating figure at the center, embodying presence. Placing them on the water's surface reinforces the idea of being grounded yet floating—emotionally unanchored from chaos. The lotus flowers around the figure represent purity and awakening, subtly reinforcing enlightenment through stillness. The hazy mountains and mist suggest impermanence and depth, while the low-lying sun creates a warm halo, symbolizing inner light and awakening consciousness. The soft lavender and periwinkle palette evoke calm, with warm gold highlights mirroring the sun's glow—visual metaphors for inner peace. Radial symmetry around the sun and a centered composition guide the eye inward, mimicking breath and focus. Gentle ripples and reflections deepen immersion, suggesting that even stillness contains subtle movement—a nod to mindful observation. Digital rendering with smooth gradients enhances the dreamlike serenity, making the scene feel both real and transcendent. Together, these choices translate abstract mindfulness into a sensory, emotionally resonant experience.

Image



Structured Vision


```
{
 "style_name": "Meditative Stillness Landscape",
 "inspiration": [
  "Zen minimalism",
  "Japanese landscape aesthetics",
  "contemporary mindfulness art"
 ],
 "scene": "A tranquil lake at dawn with misty mountains and a meditating figure",
 "subjects": [
  {
   "type": "meditating_figure",
   "description": "silhouetted person in lotus position with hands in mudra",
   "position": "center",
   "size": "large",
   "details": "hands forming meditation gesture, long hair, seated on water surface"
  },
  {
   "type": "lotus_flowers",
   "description": "pink water lilies with open petals",
   "position": "surrounding the figure",
   "size": "various (small to medium)",
   "details": "reflected in water, scattered across calm surface"
  },
  {
   "type": "mountains",
   "description": "hazy mountain range",
   "position": "background",
   "size": "large",
   "details": "mist-covered peaks, soft silhouettes against sky"
  },
  {
   "type": "sun",
   "description": "partially obscured sun",
   "position": "behind the figure",
   "size": "medium",
   "details": "creating a soft halo effect, warm glow"
  }
 ],
 "style": "digital illustration with smooth gradients",
 "color_palette": {
  "primary": "soft lavender",
  "secondary": "muted periwinkle",
  "highlight": "warm gold",
  "shadow": "deep plum",
 "background_gradient": [
    "light peach",
    "pale lavender"
  ]
 },
 "lighting": {
  "source": "low-lying sun",
  "direction": "from behind the figure",
  "tone": "warm golden hour",
  "intensity": "gentle, diffused"
 },
 "mood": "serene and introspective",
 "background": {
  "type": "scenery",
  "details": "misty mountain range, calm water surface with subtle ripples"
 },
 "composition": "centered subject with radial symmetry around the sun, balanced arrangement of lotus flowers",
 "camera": {
  "angle": "eye-level",
  "distance": "medium shot",
  "lens": "standard perspective",
  "focus": "sharp figure, soft background"
 },
 "medium": "digital rendering",
 "textures": [
  "smooth water surface",
  "subtle petal textures",
  "gentle ripple patterns"
 ],
 "details": {
  "weather": "clear with light mist",
  "time_of_day": "dawn",
  "season": "indeterminate"
 },
 "effects": [
  "gentle water ripples",
  "soft glow around sun",
  "subtle reflection effects"
 ],
 "themes": [
  "mindfulness",
  "tranquility",
  "presence",
  "nature connection"
 ],
 "optional_details": "Lotus flowers symbolize purity and enlightenment in Eastern traditions, reinforcing the meditative theme. The silhouette emphasizes universal applicability of mindfulness practices."
}
```


Generative Prompt

Show a mindful poster titled \"Still Waters: Breathe Deep, Be Here!\" with that exact phrase clearly legible, featuring ripples on calm water, lotus flowers, and a serene silhouette meditating at sunset in tranquil lavenders, soft greys, and pale peaches, arranged in a balanced, uncluttered layout with harmonious colors and seamless text integration.

Readable excerpt of the missing and conflicting commitments

Request: a poster with "Still Waters: Breathe Deep, Be Here!"
Final prompt: requires "that exact phrase clearly legible" and selects "meditating at sunset."
Raw structure: `time_of_day` is "dawn"; the requested visible phrase is absent.
Implication: missing wording loses required content. Dawn versus sunset is a separate optional-detail conflict, since the request does not specify a time of day.

Figure 8: Raw-pool inconsistency: the final prompt selects sunset and requires exact visible poster text; the image-derived structure specifies dawn and omits the exact wording. The timing conflict concerns optional detail; missing required wording reduces semantic coverage. This raw example predates the canonical schema.

J Extended Related Work and Scope

This section separates the research directions relevant to SpecPE by what they optimize, what intermediate evidence they expose, and what their evaluations establish. These works provide context; only systems listed in the result tables are evaluated.

Language reasoning and reinforcement learning. General-purpose language-model reports, including GPT-4, Qwen2.5, and Qwen3, document capabilities and training choices that motivate using language models as prompt enhancers (OpenAI, 2023; Yang et al., 2024; 2025a). They are distinct from the Qwen2.5-VL and Qwen3-VL reports for the actual PE backbones used here (Bai et al., 2025b;a). OpenAI o1 and DeepSeek-R1 develop reasoning-oriented systems (Jaech et al., 2024; Guo et al., 2025a); DeepSeekMath introduces the group-relative optimization adopted by SpecPE (Shao et al., 2024). Skywork-OR1 and Logic-RL further investigate RL for reasoning, including rule-based feedback (He et al., 2025; Xie et al., 2025). These results motivate learning a reasoning policy but do not establish that a plausible rationale preserves visual requirements.

Feedback and optimization choices. DAPO studies an open RL system at scale, while the analysis of R1-Zero-like training examines how optimization details affect reasoning outcomes (Yu et al., 2025; Liu et al., 2025d). Critique-GRPO combines natural-language and numerical feedback (Zhang et al., 2025c); process supervision instead checks intermediate solution steps (Lightman et al., 2023). SpecPE uses field- and span-grounded decisions to construct a scalar completion reward. It does not claim to implement token-level process rewards, natural-language critiques, or every optimization modification from these systems. This distinction also motivates reporting optimizer exposure separately from the number of rendered reward images.

Reinforcement learning for visual generation. DDPO and DPOK optimize diffusion models through image-level feedback (Black et al., 2023; Fan et al., 2023). DanceGRPO and Flow-GRPO investigate group-relative training for visual generation and flow matching (Xue et al., 2025; Liu et al., 2025a), while Pref-GRPO uses pairwise preference rewards (Wang et al., 2025b). Visual-CoG uses stage-aware rewards (Li et al., 2025b), while CoT-lized Diffusion couples layout updates to diffusion denoising (Liu et al., 2025c). These approaches optimize generator behavior or integrate planning into synthesis; SpecPE instead supervises the text interface to a frozen renderer, including whether a declared plan survives rewriting.

Planning and reasoning for synthesis. T2I-R1, PARM, and ReasonGen-R1 connect reasoning to autoregressive image generation through semantic planning, intermediate verification, or SFT/RL (Jiang et al., 2025a; Guo et al., 2025b; Zhang et al., 2025e). GoT and GoT-R1 make semantic and spatial information part of visual generation (Fang et al., 2025; Duan et al., 2026). Uni-CoT and Visual-Aware CoT investigate unified or visually consistent reasoning for synthesis (Qin et al., 2025; Ye et al., 2025). ThinkGen uses a decoupled MLLM and diffusion transformer with alternating reinforcement updates (Jiao et al., 2025). SpecPE shares the motivation to expose scene decisions, but supervises a text interface to a fixed generator. Its results therefore concern modular prompt enhancement rather than direct improvement of an autoregressive image decoder.

Interleaved generation and efficiency. ImageGen-CoT studies reasoning for text-to-image in-context learning (Liao et al., 2025); Interleaving Reasoning and DraCo alternate synthesis with reflection or draft verification (Huang et al., 2025a; Jiang et al., 2025b). ThinkMorph learns complementary text-image trajectories for visual reasoning (Gu et al., 2026). ShortCoTI reduces redundant reasoning prompts for autoregressive generation (Gu et al., 2025). Together, these directions motivate accounting for intermediate images and generated tokens. Our Interleaved control is the specific one-draft adaptation in Appendix C, not a claimed reproduction of all such systems. The common final-image sampling protocol does not make its inference cost equal to that of a text-only enhancer.

Grounding complex requests. LayoutGPT and LLM-grounded Diffusion turn descriptions into explicit spatial plans (Feng et al., 2023; Lian et al., 2023). LayerCraft organizes synthesis through layered objects, and DeCoT decomposes complex instructions (Zhang et al., 2025f; Lin et al., 2025). RealRAG addresses realistic generation with retrieval-augmented knowledge (Lyu et al., 2025b), a complementary source of information absent from SpecPE's request-only constraint compiler.

RePrompt and PromptEnhancer learn modular rewriters with image-derived rewards (Wu et al., 2026; Wang et al., 2026). SpecPE asks whether declared scene commitments are request-supported and survive rewriting; structure alone does not establish correct interpretation.

Thinking with images and visual tools. MVoT visualizes spatial reasoning, GRIT trains image-grounded thinking, and Thyme investigates reasoning beyond static images (Li et al., 2025a; Fan et al., 2025; Zhang et al., 2025d). OpenThinkIMG and VTool-R1 learn visual-tool use (Su et al., 2025a; Wu et al., 2025b), while DeepEyes learns active perception (Zheng et al., 2025). Pixel Reasoner encourages pixel-space exploration; VisTA studies visual-tool selection; ACTIVE-o3 and VisionThink investigate active or efficient perception (Wang et al., 2025a; Huang et al., 2025b; Zhu et al., 2025; Yang et al., 2025b). These methods obtain or manipulate visual evidence during reasoning. SpecPE's intermediate object is instead a declared specification that is evaluated before final rendering; no visual-tool interaction is required at deployment.

Grounded evidence and spatial limitations. Seg-Zero, Point-RFT, and UniVG-R1 connect reasoning with segmentation, visual grounding, or grounded reinforcement fine-tuning (Liu et al., 2025b; Ni et al., 2025; Bai et al., 2025c). DeepScan supplies a training-free framework for visually grounded reasoning, while Latent Sketchpad explores latent visual thoughts (Li et al., 2026; Zhang et al., 2025b). A position paper on spatial reasoning argues that scaling alone is insufficient and motivates changes to training and reasoning (Zhang et al., 2025a). These works support treating spatial understanding and evidence access as separate design questions. They do not justify interpreting SpecPE's JSON specification as an image-grounded internal state. Surveys place these distinctions within thinking with images and multimodal RL (Su et al., 2025b; Zhou et al., 2025).

Unified models and renderer scope. Unified multimodal pretraining and Understanding-in-Generation connect visual understanding with synthesis (Deng et al., 2025; Lyu et al., 2025a). HunyuanImage 3.0 documents another route to image generation through large multimodal models (Cao et al., 2025). Our evaluated renderers are FLUX.1-dev and Qwen-Image (Black Forest Labs, 2024; Wu et al., 2025a); mentioning other model families does not imply that SpecPE has been tested on them. Unchanged-checkpoint transfer between the two evaluated renderers establishes a limited compatibility result, not architecture-independent transfer to every unified model.

Preferences, instruction following, and evaluators. ImageReward and HPS evaluate human preferences in image synthesis (Xu et al., 2023; Wu et al., 2023); T2I-ReasonBench and WISE test reasoning-informed and knowledge-informed generation (Sun et al., 2025; Niu et al., 2025). VIBE studies visual instruction following in image editing, which is adjacent to but distinct from our text-to-image setting (Zhang et al., 2026). The LLM-as-a-judge literature motivates checking evaluator agreement and failure modes rather than equating model judgments with ground truth (Gu et al., 2024). We distinguish preference, request satisfaction, semantic-judge validation, and rendering; agreement among aggregate estimates does not establish statistical significance.

K AI USE STATEMENT

AI tools were used in this work solely to support manuscript preparation and code implementation. LLM assisted with improving clarity, refining phrasing, correcting grammatical errors, and identifying potential errors or inconsistencies in theoretical derivations. Coding agents provided additional assistance with debugging. These applications were limited to refining author-developed material; they did not involve generating original research ideas, establishing scientific claims, choosing research methods, interpreting experimental findings, or determining conclusions.

The authors independently conceived and developed the scientific contributions of this study, including its research questions, methodology, theoretical framework, experimental design, analysis, and conclusions. Human creativity and scientific judgment guided the research throughout, and AI assistance did not replace the authors' intellectual contributions or decision-making. Suggestions and outputs from AI tools were regarded as provisional rather than authoritative, and any material incorporated into the work underwent independent review by at least two authors. Beyond reviewing AI-assisted content, the authors verified the scientific claims, theoretical reasoning, experimental

results, and final manuscript in their entirety. The authors therefore retain full responsibility for the accuracy, originality, and integrity of the work, including any errors or omissions that may remain.