PRISM: SCOPE-AWARE MULTI-AGENT PROMPT INDUCTION FOR PROMPTLESS QUESTION GROUNDED VISUAL SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional Text-to-Image (T2I) synthesis and optimization assume that a visual prompt is already available. In reading comprehension (RC) illustration synthesis, however, the input consists of a passage and its questions, with no visual prompt specified. The system must first determine what evidence each illustration should depict and then induce prompts for synthesis. We formulate this ***prompt induction*** challenge as *Question Grounded Visual Synthesis* (QGVS). Its *dual-scope constraints* require question-scope evidence selection and answer secrecy for each illustration, alongside passage-scope visual consistency across the set. We introduce PRISM, a multi-agent system for RC illustration synthesis. Its scope-aware workflow jointly plans the full illustration set: question-scoped agents select evidence and construct renderer prompts, while passage-wide style planning and cross-question entity routing coordinate visual consistency. The direct mode uses frozen pretrained models. Across four RC dataset sources, direct PRISM improves visual quality over the strongest comparison (GeoMean 66.0 versus 60.6), with especially large consistency gains while preserving distinct question content. This advantage persists across the multiple tested VLM planners and T2I renderer sizes. Further experiments show that the synthesized illustrations yield larger question-answering gains for simulated developing and beginning readers. Finally, GRPO post training yields a PRISM variant with a different performance profile.

1 INTRODUCTION

Reading Comprehension (RC) is a fundamental capability for understanding and reasoning over written passages. In a typical RC task, answering a question requires readers to locate relevant evidence, connect events and relations, and integrate information distributed across the passage (Lai et al., 2017). Successful comprehension therefore depends on organizing distributed evidence into a coherent account. Visual representations, particularly evidence grounded illustrations, can make the relevant entities and relations more explicit and easier to inspect, thereby supporting evidence integration and question answering in RC (Chun & Plass, 1996; Hibbing & Rankin-Erickson, 2003).

Creating tailored illustrations manually for each passage and question requires substantial human effort and is difficult to scale to large RC collections. This naturally leads us to consider automatic illustration synthesis with Text-to-Image (T2I) models. T2I models synthesize images from natural language descriptions, and recent systems can render diverse, high quality scenes with strong correspondence to textual instructions. Despite this apparent suitability for illustration synthesis, applying T2I to RC exposes a fundamental obstacle: conventional T2I generation (Rombach et al., 2022; Black Forest Labs et al., 2025) and prompt optimization (Mañas et al., 2024; Liu et al., 2025; Wu et al., 2026b) **assume an image prompt that already specifies what the image should depict.** However, RC illustration synthesis provides no such image prompt. We refer to the required process of inferring a renderer ready visual specification from RC inputs as *prompt induction* (Figure 1). This absence arises because a passage and question are written to convey and test meaning, not to specify which evidence should become visual, which scene should express it, or which details should remain hidden. These content decisions must be resolved before T2I generation, whereas prompt optimization can only improve how an already specified visual intent is expressed.

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.

We formalize the challenge of synthesizing illustrations for RC as *Question Grounded Visual Synthesis* (QGVS). Given a passage and its associated questions, QGVS synthesizes one illustration for each question. Its defining difficulty lies in *dual-scope constraints*. Question-scope constraints require each illustration to select relevant passage evidence, support the corresponding reasoning process, and avoid answer text, option markers, or other selection cues. Passage-scope constraints require the complete illustration set to share a visual language and preserve the identities of recurring entities. Increasing consistency cannot come at the cost of giving distinct questions duplicated content. QGVS is therefore a structured illustration set synthesis task in which local semantic decisions and global appearance decisions must be coordinated without being conflated.

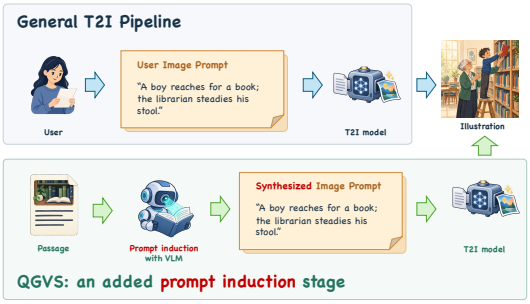

Figure 1: **Prompt induction in QGVS.** Conventional T2I starts from an existing prompt, whereas QGVS should infer visual intent before rendering.

With QGVS now clearly defined, we can return to a natural question: *why do existing approaches fall short?* Direct T2I synthesis models only map a supplied prompt to an illustration; applying them to QGVS still requires a separate model to construct that prompt from RC inputs (Rombach et al., 2022; Black Forest Labs et al., 2025). Prompt optimization retains the same assumption: it takes an initial prompt and refines its detail, composition, or alignment (Mañas et al., 2024; Liu et al., 2025; Wu et al., 2026b). Even recent reinforcement learning based and reward based methods train prompt rewriters from illustration level feedback, but likewise presuppose an initial prompt (Wu et al., 2026a; Wang et al., 2026). Thus, these paradigms operate only after visual intent has been specified; they neither perform *prompt induction* nor explicitly coordinate *dual-scope constraints*. Story visualization and other promptless visual synthesis tasks infer visual intent directly from task inputs (Wu et al., 2025; Zhang et al., 2025; He et al., 2025). Their objectives differ from QGVS, which must select question-specific evidence without revealing answers while coordinating appearance across a passage-wide illustration set.

We introduce PRISM, **P**assage Grounded **R**easoning Aware **I**llustration **S**ynthesis with **M**odular Planning, a multi-agent system for promptless RC illustration synthesis. To address the *dual-scope constraints*, PRISM factorizes *prompt induction* by decision scope. It first establishes a visual language shared by the full illustration set, then plans evidence and renderer prompts for distinct questions together. Recurring identities are consolidated across the passage, with relevant references routed to each question before rendering. A final review checks the illustrations for answer leakage. These steps keep question-specific content distinct while maintaining passage-wide appearance. Besides the training-free direct PRISM, which uses frozen pretrained models, we also explore optional GRPO post training to obtain a variant with a different performance profile.

We evaluate PRISM on four RC sources: RACE-High and RACE-Middle (Lai et al., 2017), RACE-C (Liang et al., 2019), and CosmosQA (Huang et al., 2019). Direct PRISM produces stronger illustration sets across the tested planners and renderers; ablations trace this advantage to planning question evidence jointly while keeping it distinct from shared appearance. In experiments with simulated student reading profiles, we find that these illustrations can improve question-answering accuracy, especially for developing and beginning readers. RL comparisons, in turn, yield different performance profiles with little aggregate improvement, revealing the difficulty of learning *prompt induction* from outcome feedback. Our contributions are:

- We formulate QGVS as a *prompt induction* task in which a passage and its questions determine what evidence to depict under question-scope and passage-scope constraints.
- We introduce PRISM, a multi-agent system that factorizes planning by scope, coordinating question-specific evidence, shared appearance, and cross-scope entity routing.
- We conduct extensive experiments across four reading comprehension data sources and multiple model backbones, showing that PRISM produces higher-quality, more consistent illustration sets and improves question answering for developing and beginning readers, as evaluated via simulated students.

2 Related Work

Visual Representations for Reading Comprehension. Educational research has examined multimedia reading materials and visual images as aids to comprehension, including for struggling readers (Chun & Plass, 1996; Hibbing & Rankin-Erickson, 2003). Computational RC, by comparison, has largely studied answering questions from passages, as in RACE (Lai et al., 2017); related question-generation work studies how to construct the questions themselves (Li & Zhang, 2024; Cheng et al., 2025). Together, these strands establish the value of visual support and passage-based assessment, but leave open how to choose and coordinate illustrations for a passage's questions.

Prompted T2I Synthesis and Prompt Optimization. Modern T2I models synthesize illustrations from supplied prompts (Rombach et al., 2022; Black Forest Labs et al., 2025). OPT2I, DialPrompt, and VisualPrompter refine them through optimization, interaction, or visual feedback (Mañas et al., 2024; Liu et al., 2025; Wu et al., 2026b), while PromptSculptor and GenPilot use specialized agents (Xiang et al., 2025; Ye et al., 2025). RePrompt and PromptEnhancer learn prompt rewriters from rendered outcomes (Wu et al., 2026a; Wang et al., 2026). Although these systems advance prompt-driven synthesis and optimization, they assume an initial prompt that specifies visual intent. This assumption makes them ill-suited to the *prompt induction* challenge in the promptless QGVS, where visual intent must first be inferred from the passage and questions.

Promptless Task Specific Visual Synthesis. Other promptless tasks infer visual intent from text: paragraph to image synthesis targets one scene (Wu et al., 2025), story visualization targets narrative continuity (Zhang et al., 2025; He et al., 2025), and document to diagram systems extract scientific structure (Mondal et al., 2024; Wang et al., 2025). QGVS instead requires question-specific evidence, shared appearance, and answer secrecy across an illustration set. These different task contracts preclude direct baseline comparisons.

Outcome Optimization for Structured Policies. GRPO estimates relative advantages from groups of sampled outputs without a learned critic (Shao et al., 2024). For QGVS, feedback arrives after the complete illustration set is rendered, making credit assignment across evidence, style, prompt, and entity decisions difficult. We study shared-adapter GRPO as an optional extension to PRISM; our results show only limited gains and a shift between consistency and diversity.

3 Methodology

We introduce PRISM, ***P**assage Grounded **R**easoning Aware **I**llustration **S**ynthesis with **M**odular Planning*. As illustrated in Figure 2, PRISM addresses the missing image prompt challenge in QGVS by factorizing *prompt induction* around its *dual-scope constraints*: question-scope evidence selection, passage-scope appearance planning, and cross-scope entity consolidation and routing, followed by rendering and answer secrecy review. By coordinating these decisions, PRISM synthesizes illustrations that convey distinct question-relevant evidence, maintain a coherent visual identity across the passage, and avoid direct answer leakage.

3.1 Problem Formulation

A QGVS instance is $x = (p, \mathcal{Q})$, where p is a passage and $\mathcal{Q} = \{(q_i, a_i)\}_{i=1}^{n}$ is a set of question answer pairs. The output is an illustration set $\mathcal{I} = \{I_i\}_{i=1}^{n}$, one per question. Its desired properties can be summarized by a conceptual objective:

$$\max_{\mathcal{I}} \quad \sum_{i=1}^{n} \big[\mathrm{Evidence}(I_i; p, q_i) + \lambda_{\mathrm{g}} \mathrm{Grounding}(I_i; p)\big] + \lambda_{\mathrm{c}} \mathrm{Consistency}(\mathcal{I}), \quad \text{s.t. } \ell(I_i; a_i) = 0 \ \forall i. \tag{1}$$

Here $\mathrm{Evidence}$ measures relevance to q_i, $\mathrm{Grounding}$ faithfulness to p, and $\mathrm{Consistency}$ shared style and recurring identities; the weights are conceptual, not a training loss. Illustrations should retain distinct question content. The answer a_i may guide *prompt induction*, but $\ell(I_i; a_i) = 0$ rules out answer text, option markers, and other direct selection cues.

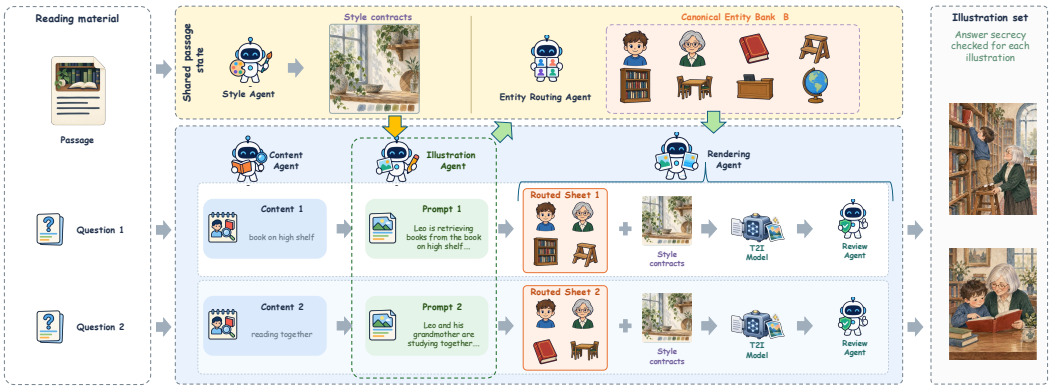

Figure 2: **PRISM overview.** Passage-wide style planning, question-specific prompt induction, and prompt-grounded entity routing feed two-reference rendering and answer-secrecy review.

3.2 PASSAGE-LEVEL SCOPE AWARE FACTORIZATION

Given x, PRISM factorizes *prompt induction* according to the semantic scope of each decision. Passage-scoped variables control visual styles and shared entities, whereas question-scoped variables preserve evidence and rendering intent for each QA pair. A passage-level multi-agent workflow realizes this factorization over the complete QA set: in each clean pass, every reasoning agent returns one ordered batch covering all questions, and a shared state is populated sequentially:

$$\underbrace{\mathsf{State}(x)}_{\text{joint passage state}} = \big(s, \kappa, \mathcal{E}, \mathcal{P}, \mathcal{B}, \mathcal{R}\big), \qquad \begin{aligned} \mathrm{scope}(s, \kappa, \mathcal{B}) &= p, \\ \mathrm{scope}(e_i, P_i, r_i) &= q_i, \end{aligned} \tag{2}$$

Here s is the shared *style contract* and κ the *style calibration prompt*. The question-specific outputs are the content plans $\mathcal{E} = \{e_i\}_{i=1}^n$ and completed *renderer prompts* $\mathcal{P} = \{P_i\}_{i=1}^n$. A passage-wide canonical entity bank $\mathcal{B}$ and question-specific entity routes $\mathcal{R} = \{r_i\}_{i=1}^n$ connect the two scopes.

Passage scope: shared appearance (s, κ). Visual consistency is a property of the complete illustration set, whereas scene content remains question specific. The *Style Agent* $\mathcal{A}_{\text{sty}}$ (Appendix A.4.1) plans the visual style of the complete illustration set through a 12-field *style contract* s (Table 7 in the Appendix) and a *style calibration prompt* κ. The latter specifies separated visual studies without an answer-bearing event or relation for later style-reference rendering. Formally,

$$(s, \kappa) = \mathcal{A}_{\text{sty}}(p, \{(q_i, a_i)\}_{i=1}^n)\,. \tag{3}$$

Question scope: evidence and realization $(\mathcal{E}, \mathcal{P})$. Each question requires its own evidence decision without losing awareness of the other questions. The *Content Agent* $\mathcal{A}_{\text{con}}$ (Appendix A.4.2) processes the complete QA set in one batch. For each question, it proposes three materially different, passage-grounded, renderer-feasible candidate scenes, assesses their strengths and faithfulness risks, selects one verbatim, and records the selection reason and excluded content. Subsequently, the *Illustration Agent* $\mathcal{A}_{\text{ill}}$ (Appendix A.4.3) realizes all selected scenes as an ordered batch of *renderer prompts*, preserving entities, actions, counts, and relations while applying the passage-wide *style contract* s only to appearance:

$$\mathcal{E} = \mathcal{A}_{\text{con}}(x), \qquad \mathcal{P} = \mathcal{A}_{\text{ill}}(x, \mathcal{E}, s, \kappa). \tag{4}$$

Cross scope: entity consolidation and routing $(\mathcal{B}, \mathcal{R})$. Recurring identities must be shared globally but supplied only where locally relevant. After all *renderer prompts* are complete, the *Entity Routing Agent* $\mathcal{A}_{\text{obj}}$ (Appendix A.4.4) identifies prompt-grounded visible entities under a configurable per-question budget K_{ent} and merges recurring identities across questions into the passage-wide canonical bank $\mathcal{B} = \{b_j\}_{j=1}^{N_{\mathcal{B}}}$, where $N_{\mathcal{B}}$ is the number of canonical entities. For each question q_i, it maps selected entities to their canonical entries in $\mathcal{B}$; the runtime derives the ordered route r_i

from these local mappings. We use $K_{\text{ent}} = 4$ in the reported experiments. Writing $\mathcal{A}_{\text{obj}}$ for the complete routing interface,

$$(\mathcal{B}, \mathcal{R}) = \mathcal{A}_{\text{obj}}(x, \mathcal{P}), \qquad \mathcal{R} = \{r_i\}_{i=1}^{n}, \quad |r_i| = K_{\text{ent}}. \tag{5}$$

Rendering: shared references and question-specific illustrations $(I^{\text{sty}}, \{J_j\}, \mathcal{I})$. The *Rendering Agent* $\mathcal{A}_{\text{rnd}}$ (Appendix A.4.5) renders one shared style reference image I^{sty} from s and κ and one canonical entity reference image J_j for each $b_j \in \mathcal{B}$, once per passage. For each question, it assembles a label-free identity sheet $S(r_i)$ from the routed references (a 2×2 sheet when $K_{\text{ent}} = 4$) and supplies that sheet alongside the shared style reference image to the renderer. The *renderer prompt* remains the semantic authority:

$$\left(I^{\text{sty}}, \{J_j\}_{j=1}^{N_{\mathcal{B}}}, \{S(r_i), I_i\}_{i=1}^{n}\right) = \mathcal{A}_{\text{rnd}}(s, \kappa, \mathcal{B}, \mathcal{R}, \mathcal{P}). \tag{6}$$

The final rendering call is $I_i = f_{\text{img}}(P_i, I^{\text{sty}}, S(r_i))$.

Answer secrecy review $\mathcal{L}$. Scope separation must remain valid after rendering. The *Review Agent* $\mathcal{A}_{\text{rev}}$ (Appendix A.4.6) checks the complete illustration set for directly visible answer disclosure. If it finds a leak, it identifies the visible cue and returns feedback to the *Illustration Agent*, which revises the affected *renderer prompt* before rerendering. Let $\ell_i = \ell(I_i; a_i)$ denote the leakage indicator for question q_i. An illustration set passes this check only when every indicator is zero:

$$\mathcal{L} = (\ell_i)_{i=1}^{n} = \mathcal{A}_{\text{rev}}(x, \mathcal{I}), \qquad \text{Release}(\mathcal{I}) \iff \sum_{i=1}^{n} \ell_i = 0. \tag{7}$$

3.3 Optional Outcome Post Training with GRPO

Direct PRISM requires no task-specific parameter updates. We additionally study whether outcome supervision can move this structured policy to a different balance among visual objectives. Rather than fitting a separate policy for each module, we attach one shared LoRA adapter $\boldsymbol{\theta}$ to the frozen planner and route the Style, Content, Illustration, and Entity Routing Agents through it. Their prompts and output contracts retain role specialization, while their parameter updates are coordinated in a common policy. For each passage, we evaluate K complete trajectories τ_k with an auxiliary QA-gain signal and the prompt-blind multimodal Judge:

$$R_k = \alpha \omega_{\text{qa}} z(R_k^{\text{qa}}) + (1 - \alpha) z(R_k^{\text{judge}}) - \lambda \Delta_k^{\text{qa}}, \qquad \widehat{A}_k = \frac{R_k - \mu_R(x)}{\sigma_R(x) + \epsilon}. \tag{8}$$

Here z standardizes each signal within the passage group, ω_{qa} discounts an unreliable low-variance QA signal, and $\Delta_k^{\text{qa}} = \max(0, -R_k^{\text{qa}} - \delta)$ penalizes QA regression beyond margin δ. The mean $\mu_R(x)$ and standard deviation $\sigma_R(x)$ are computed over valid candidate rewards. The same trajectory advantage is attached to every traced action from the four Agents. We then jointly optimize their union with one clipped GRPO objective, normalizing repeated visits so that each Agent contributes equal total weight per trajectory. The resulting adapter is rebound to all four interfaces for the next rollout round. This extension is tradeoff control rather than a prerequisite for PRISM or a guarantee of uniform improvement. Appendix G provides the reward construction, shared adapter objective, rollout schedule, and hyperparameters.

4 Experiments

Our experiments address five research questions. **RQ1** asks whether PRISM improves promptless QGVS while preserving question-specific content and passage-level consistency. **RQ2** examines whether the synthesized illustrations improve downstream reading comprehension across student VLMs and simulated reading profiles. **RQ3** tests which components of passage-level scope aware factorization account for the gains through controlled ablations. **RQ4** asks whether those gains persist across planner and renderer backbones. **RQ5** examines how optional outcome post training changes PRISM's balance of visual objectives.

4.1 EXPERIMENTAL SETUP

Benchmark Dataset. We construct one mixed RC benchmark from four sources: RACE-High and RACE-Middle (Lai et al., 2017), RACE-C (Liang et al., 2019), and CosmosQA (Huang et al., 2019). Each passage and all its questions form one bundle. After preprocessing, we sample 20% within each source and split, then pool the sampled bundles. All methods use the same mixed test set of 240 passage bundles and 856 questions; see the details in Appendix B.

Comparison Methods. Unless stated otherwise, methods receive the same passage and question inputs and use Qwen3.5-9B as the VLM planner and FLUX.2 Klein-9B KV as the T2I renderer. Both backbones remain frozen for direct inference; image resolution, denoising steps, and guidance scale are matched, while generation calls and elapsed time are reported to account for differing method budgets (Appendix D). We organize baselines by the information and computation available to them. *Direct generation* includes PFQS (Li & Zhang, 2024), which plans the questions before constructing renderer-ready prompts. *Iterative optimization* includes PromptRefine, our task adaptation of Self Refine (Madaan et al., 2023); OPT2I (Mañas et al., 2024); DialPrompt (Liu et al., 2025); VisualPrompter (Wu et al., 2026b); PromptSculptor (Xiang et al., 2025); and GenPilot (Ye et al., 2025). Methods that require an initial image prompt first induce one from the RC inputs before applying their respective optimization procedures. *RL based optimization* includes task adapted RePrompt (Wu et al., 2026a) and PromptEnhancer (Wang et al., 2026), together with PPO, DPO, and RLOO applied to a basic prompt induction policy. Finally, we report directly deployable PRISM and its globally optimized GRPO variant. Agent prompts and interfaces are documented in Appendix A; baseline adaptations are given in Appendix D.2, and model settings in Appendix D.

Evaluation dimensions. We report five quality dimensions and runtime. ***Faithfulness*** uses CLIP to measure semantic similarity between each illustration and its passage (F_P) or corresponding question (F_Q) (Radford et al., 2021). ***Diversity*** reports the fraction of cross-question pairs that are not near duplicates in *CLIP* and *DINO* feature spaces; a pair is a near duplicate when its cosine similarity exceeds 0.85 in the respective space. ***Consistency*** measures pairwise similarity among illustrations from the same passage; above 0.85, the score decreases as similarity rises, so near duplicates cannot inflate it. It is reported as *CLIP-P* and *DINO-P* (Radford et al., 2021; Caron et al., 2021). ***LAS*** is the geometric mean of six L1 category scores from our task-adapted hierarchical multimodal Judge, which extends the structure of Qwen-Image-Bench (Li et al., 2026): *Quality*, *Aesthetics*, *Alignment*, *Educational Fidelity*, *Creative Generation*, and *Style Consistency*. ***Aesthetic*** uses a CLIP-L/14 predictor (LAION-AI, 2022) to estimate the aesthetic score. ***Time*** is reported per q_i relative to PFQS. ***GeoMean*** combines 13 constituent scores: two faithfulness, two consistency, two diversity, six LAS L1 scores, and one aesthetic score.

Evaluation protocol. The Judge observes the passage, questions, category-relevant renderer prompts, and visible illustration set, but not method identity or hidden planner states. Correct answers are supplied only as private evaluation references for alignment and answer secrecy, never as visible output cues. Evaluators and generation seeds are fixed across methods. We use two-sided paired passage-level t tests, with $p < 0.05$ denoting significance. Appendix C specifies the metrics, Judge hierarchy, hard gates, aggregation, and statistical tests.

4.2 RQ1: DOES PRISM IMPROVE PROMPTLESS QGVS?

Table 1 compares directly deployable methods without task-specific post training; trained methods are analyzed under RQ5. The clearest pattern is that direct PRISM leads overall (*GeoMean*) while combining strong question faithfulness (F_Q) with substantially better passage-wide consistency (*CLIP-P/DINO-P*) and Judge quality (*LAS*). This combination is consistent with the scope decomposition: question-scope evidence planning preserves local relevance, while passage-scope style control and cross-scope entity routing stabilize shared appearances. We can also see that additional inference effort alone does not close the gap: OPT2I and PromptSculptor take longer (*Time*) than direct PRISM yet remain behind it on the aggregate.

The lower automatic diversity (*CLIP/DINO*) for PRISM also needs interpretation. A near-duplicate threshold can treat intended shared style and recurring identities as repetition, even when the scenes serve different questions. Human annotations later rate PRISM's diversity (*Diversity*) near the high-

Table 1: Main results on the pooled test set of the four-dataset mixture. Bold marks the best. Time is relative to PFQS (1.00×). All quality-metric differences from PRISM are significant ($p < 0.05$). Per-source results for the four constituent datasets appear in Appendix Table 18.

		Faithfulness		Consistency		Diversity		LAS	Aesthetic	Time
Method	***GeoMean***	F_P	F_Q	CLIP-P	DINO-P	CLIP	DINO	Score	Score	per q_i
PFQS (Li & Zhang, 2024)	57.4	21.5	23.6	59.5	34.7	97.3	99.4	70.8	57.6	**1.00**×
OPT2I (Mañas et al., 2024)	58.0	21.4	23.5	59.5	33.0	**98.4**	**99.9**	72.9	57.3	4.37×
VisualPrompter (Wu et al., 2026b)	58.4	21.6	23.6	60.1	33.5	97.6	**99.9**	73.6	57.8	3.73×
GenPilot (Ye et al., 2025)	59.2	21.9	23.3	65.7	44.1	93.9	99.1	71.7	58.8	2.95×
PromptSculptor (Xiang et al., 2025)	59.6	21.8	23.4	63.8	39.5	96.3	99.5	73.9	59.7	7.01×
DialPrompt (Liu et al., 2025)	60.4	21.5	23.3	62.1	34.8	96.8	99.4	78.5	58.6	3.64×
PromptRefine (Madaan et al., 2023)	60.6	21.6	23.4	63.3	38.6	93.3	97.3	78.4	57.6	3.17×
PRISM	**66.0**	**22.4**	**23.9**	**73.1**	**63.6**	86.0	91.9	**84.9**	**61.3**	3.86×

est in Table 5, suggesting that its visual coherence does not really erase the question-specific distinctions they perceive. Meanwhile, baselines with higher diversity cannot achieve the same coherence and LAS results across the illustration set.

4.3 RQ2: Do Synthesized Illustrations Improve Reading Comprehension?

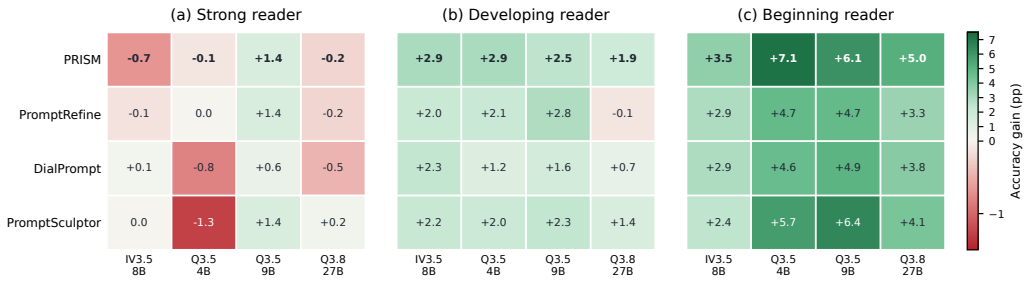

Figure 3: Multiple-choice question answering (MCQA) accuracy gain (percentage points) from adding an illustration to passage text across simulated reading profiles and student VLMs (Qwen3.5-4B/9B, Qwen3.8-27B, and InternVL3.5-8B). Appendix C.5 gives the full protocol.

To assess whether illustrations support RC beyond visual quality, we compare multiple-choice question answering (MCQA) accuracy with passage text alone and with the corresponding illustration across four student VLMs. Simulated strong readers receive the full passage; developing and beginning readers receive answer-blind selections of complete sentences guided by each question, preserving the original order and approximately half or one quarter of the passage words, respectively, along with profile-specific reasoning instructions. With the full passage, none of the illustration methods shows a consistent gain in Figure 3, likely because the text already supplies the question evidence and leaves little new information for an illustration to add. Gains grow as text access narrows. For developing and beginning readers, PRISM yields the largest gain among the four illustration methods on three of four student VLMs. This suggests that its question-specific visual evidence is most useful when readers have less textual evidence to work with, connecting the method's evidence planning to a measurable benefit in question answering.

4.4 RQ3: Why Passage-Level Scope Aware Factorization?

With the planner, renderer, and evaluation fixed, Table 2 isolates the effects of joint planning, the Agent decomposition, shared style, and entity references. Full PRISM has the strongest aggregate (*GeoMean*), but the tradeoffs are more revealing: planning questions independently or bypassing the Content Agent raises embedding consistency (*CLIP-P/DINO-P*) while reducing automated diversity (*CLIP/DINO*). This suggests that coordinating question content keeps shared appearance from obscuring differences among the requested scenes. Removing the Style Agent produces the largest aggregate loss despite high diversity, supporting passage-scope style planning. Removing

Table 2: Direct PRISM ablations. † denotes no significant difference from full PRISM ($p \geq 0.05$); ΔQ is the change in relative per-question time. Other conventions follow Table 1. Per-source results for the four constituent datasets appear in Appendix Table 19.

		Faithfulness		Consistency		Diversity		LAS	Aesthetic	Time	
Variant	***GeoMean***	F_P	F_Q	CLIP-P	DINO-P	CLIP	DINO	Score	Score	per q_i	ΔQ
PRISM	**66.0**	22.4	23.9	73.1	63.6	86.0	91.9	84.9	61.3	3.86×	0.00
Independent Question Planning	64.7	**22.7**	23.7†	**76.2**	**69.6**	63.7	77.5	86.0	61.1†	2.34×	-1.52
Monolithic Prompt Induction	65.5	22.4†	**23.9**†	73.2†	67.9	83.3†	87.6	84.0	59.9	2.19×	-1.67
w/o Style Agent	59.2	21.6	23.3	63.0	40.3	**97.4**	**99.4**	73.0	58.4	2.92×	-0.94
w/o Content Agent	65.8	22.7	23.8†	75.0	65.4	71.1	84.0	**87.2**	**61.8**†	2.25×	-1.61
w/o Entity References	62.8	22.3†	23.6	74.9	68.3	68.9	74.7	80.8	60.9†	**1.90**×	-1.97

entity references also weakens overall and Judge quality (*GeoMean*, *LAS*) even as embedding consistency rises, suggesting that similarity alone is an incomplete measure of identity control. Even the competitive Monolithic Prompt Induction variant does not preserve this balance. Together, the ablations show why the full PRISM is needed: its coordinated style, content, and entity interfaces jointly maintain question-specific evidence and coherent appearance across the set.

4.5 RQ4: DOES PRISM GENERALIZE ACROSS PLANNER AND RENDERER BACKBONES?

Table 3 shows that direct PRISM retains its overall lead (*GeoMean*) over PromptRefine, the strongest baseline in Table 1, with every tested planner and renderer pairing. Absolute quality varies across backbones, particularly with InternVL3.5-8B and Klein-4B, but the relative advantage persists. This suggests that scope-aware coordination is not tied to the default pairing, while still requiring a planner able to follow the task's structured instructions. We omit Qwen3.5-4B because it did not reliably follow those multi-step instructions in preliminary trials. Appendix E gives the metric breakdown.

Table 3: GeoMean across planners and renderers.

Planner	Renderer	PromptRefine	**PRISM**
Qwen3.8-27B	Klein-9B	60.1	**66.2**
Qwen3.5-9B	Klein-9B	60.6	**66.0**
Qwen3.5-9B	Klein-4B	59.2	**65.5**
InternVL3.5-8B	Klein-9B	57.8	**62.8**
InternVL3.5-8B	Klein-4B	57.6	**61.2**

The renderer comparison is limited to FLUX.2 Klein because PRISM depends on reliable reference-image conditioning for shared style and recurring entities. Many alternative renderers lack strong native support for such references, and an added adapter does not necessarily resolve the problem. In exploratory trials with Stable Diffusion 3.5 Large Turbo and IP-Adapter, weak reference adherence made shared appearances difficult to preserve. We therefore report both FLUX.2 Klein sizes without claiming generalization to other renderer families.

4.6 RQ5: HOW DOES OUTCOME POST TRAINING CHANGE PRISM'S OPERATING POINT?

Table 4: Outcome post training results. Bold marks the best trained result; † denotes no significant difference from PRISM (GRPO) ($p \geq 0.05$). Other conventions follow Table 1.

		Faithfulness		Consistency		Diversity		LAS	Aesthetic	Time
Method	***GeoMean***	F_P	F_Q	CLIP-P	DINO-P	CLIP	DINO	Score	Score	per q_i
RLOO (Ahmadian et al., 2024)	56.8	21.5	23.6†	59.1	33.9	**98.4**	99.5	69.5	57.4	**0.87**×
PPO (Schulman et al., 2017)	57.0	21.4	23.6	59.0	33.2	97.7	98.8	70.4	57.3	**0.87**×
DPO (Rafailov et al., 2023)	58.2	21.5	23.5	59.4	33.8	98.2	99.5	73.3	57.4	0.88×
PromptEnhancer (Wang et al., 2026)	58.7	21.6	23.3	61.0	34.2	98.2	99.6	73.9	58.4	1.55×
RePrompt (Wu et al., 2026a)	60.5	21.5	23.2	62.0	34.4	97.6	**99.6**	78.6	59.1	1.39×
PRISM (GRPO)	**66.4**	**22.4**	**23.8**	**72.0**	**62.7**	90.3	94.6	**85.3**	**61.4**	3.57×
PRISM (direct)	66.0	22.4	23.9	73.1	63.6	86.0	91.9	84.9	61.3	3.86×

Table 4 presents outcome post training as a change in operating point rather than a broad improvement. Relative to direct PRISM, the GRPO variant gains diversity (*CLIP/DINO*) but gives up some cross-illustration consistency (*CLIP-P/DINO-P*), leaving overall quality (*GeoMean*) nearly

unchanged. Direct PRISM therefore remains useful when stable shared appearance is the priority. The trained variant still leads every RL comparison, including RePrompt, the strongest trained baseline; high diversity alone does not compensate for weaker consistency and Judge quality (*LAS*) in those methods. The main advantage thus comes from PRISM's scope-aware prompt induction, while outcome learning shifts the balance among its coupled objectives without resolving them uniformly. Appendix G gives the training protocol.

4.7 HUMAN EVALUATION AND CASE STUDY

Table 5: Human rating scores use a 1–5 scale; Inter-annotator agreement is relatively high (mean pairwise quadratically weighted Cohen's $\kappa = 0.67$). Bold marks the best unrounded value.

Method	Faithfulness	Consistency	Diversity	Aesthetic	Helpfulness	Overall	Leakage ↓
PromptSculptor	3.61	2.71	**4.43**	3.77	3.51	3.60	1.7%
DialPrompt	2.96	2.41	3.89	3.52	2.71	3.10	**1.1%**
PromptRefine	3.50	2.55	4.42	3.80	3.36	3.53	2.3%
PRISM	**4.11**	4.91	4.39	**4.36**	3.91	**4.34**	1.7%
PRISM (GRPO)	4.10	**4.95**	4.35	4.33	**3.91**	4.33	3.4%

Human evaluation. We examine a fixed random sample of 50 held out passages (174 questions), comparing five relatively strong methods. Three annotators independently score each method's complete passage illustration set on faithfulness, consistency, diversity, aesthetics, and helpfulness using a five-point scale. Direct answer text is marked yes or no for each question illustration. Table 5 reports the mean of the 150 passage-level scores per method and dimension; Overall is the equally weighted mean of the five quality dimensions. Leakage is the fraction of 174 question illustrations marked yes by at least two of the three sheets. Appendix C.4 specifies the annotation details.

Human judgments broadly support the automatic comparison: direct PRISM has the highest sampled overall rating, while both PRISM variants receive higher consistency and helpfulness ratings than the baselines (*Consistency*, *Helpfulness*). They also clarify the apparent diversity tradeoff in RQ1. Despite PRISM's lower embedding-based diversity (*CLIP/DINO*), annotators rate its illustrations nearly as diverse as the highest-rated baseline (*Diversity*); shared visual identity need not make question-specific scenes feel repetitive. The fine differences between direct PRISM and PRISM (GRPO) are less clear: annotators slightly favor the trained variant for consistency, whereas automatic consistency (*CLIP-P/DINO-P*) favors direct PRISM. These small contrasts are descriptive rather than evidence that post training reliably improves the set.

Case study. To inspect behavior behind the aggregate ratings, we select five complete passage bundles from the Human Evaluation sample, covering all four benchmark sources. For each question in a selected passage, Appendix H shows the same five methods in the order of Table 5, alongside passage context and the questions. The cases are diagnostic, not an additional quantitative evaluation.

5 LIMITATIONS AND CONCLUSION

Limitations. *Renderer support.* Quantitative results cover only FLUX.2 Klein (9B and 4B). In exploratory trials with Stable Diffusion 3.5 Large Turbo and IP-Adapter, weak reference adherence made shared appearances difficult to preserve, limiting conclusions about other renderer families. *Outcome training.* Standard RL methods make limited progress on prompt induction, and GRPO changes PRISM's aggregate performance only marginally. Better feedback and credit assignment are needed to learn its coupled constraints.

Conclusion. The ablations show that visual similarity alone does not make a useful illustration set: higher embedding consistency often comes with less distinct question content. Joint evidence planning and separate appearance control give direct PRISM a better balance. Human ratings likewise distinguish shared visual identity from repeated scenes. This distinction matters for RC: simulated students gain most from illustrations when passage evidence is limited. Outcome post training shifts the consistency–diversity balance but adds little overall, leaving better feedback for coupled *prompt induction* decisions as an open direction.

AI USE STATEMENT

We used OpenAI Codex, Claude Code, multiple GPT models, and DeepSeek V4 to assist with manuscript organization, drafting and revising the text, especially the Appendix sections, and language editing. These tools also assisted with tool implementation and code. Generative AI was not used to fabricate experimental results or select outcomes for reporting. The authors reviewed the AI-assisted text, checked citations against their source records, and made all final scientific claims. The authors remain responsible for the manuscript, experiments, claims, and artifacts.

ETHICS STATEMENT

The benchmark is derived from existing reading comprehension resources. The human evaluation involved three graduate student annotators who independently rated the generated illustration sets. Their task was limited to rating the study outputs and involved no intervention or analysis of annotators' personal characteristics. We consider this evaluation to raise no additional ethical concerns. Generated images may nevertheless hallucinate unsupported details, reproduce social or cultural stereotypes, or reveal answer cues that undermine evaluation validity. PRISM therefore treats passage grounding and answer leakage as explicit constraints, but these checks do not guarantee that every output is appropriate for downstream use. Applications should retain human review, especially in high stakes settings or for populations not represented by the source corpora. We will distribute only identifiers, derived annotations, and generated artifacts permitted by the licenses of the underlying datasets and model checkpoints.

REPRODUCIBILITY STATEMENT

Section 4.1 specifies the comparison conditions. Appendix A gives the agent interfaces and leakage contract; Appendix B documents dataset construction; Appendix C defines the evaluation protocol; and Appendix D records model and optimization settings. Appendix G details the optional outcome optimization study. The supplementary material will include the retained data identifiers, preprocessing scripts, prompt schemas, evaluation rubrics, and configuration files needed to reconstruct the reported runs.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Cheng Cheng, Zhenya Huang, Guanhao Zhao, Yuxiang Guo, Xin Lin, Jinze Wu, Xin Li, and Shijin Wang. From objectives to questions: A planning-based framework for educational mathematical question generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 12836–12856, 2025.

Dorothy M. Chun and Jan L. Plass. Facilitating reading comprehension with multimedia. *System*, 24(4):503–519, 1996.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. DreamStory: Open-domain story visualization by llm-

guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11874–11891, 2025.

Anne Nielsen Hibbing and Joan L. Rankin-Erickson. A picture is worth a thousand words: Using visual images to improve comprehension for middle school struggling readers. *The Reading Teacher*, 56, 2003.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2391–2401, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://aclanthology.org/D19-1243/.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794. Association for Computational Linguistics, 2017.

LAION-AI. LAION aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. CLIP ViT-L/14 aesthetic regression model and checkpoints.

Kunze Li and Yu Zhang. Planning first, question second: An llm-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4715–4729, 2024.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-Image-Bench: From generation to creation in text-to-image evaluation, 2026. URL https://arxiv.org/abs/2605.28091.

Yichan Liang, Jianheng Li, and Jian Yin. A new multi-choice reading comprehension dataset for curriculum learning. In *Proceedings of the Eleventh Asian Conference on Machine Learning*, pp. 742–757, 2019.

Yilun Liu, Minggui He, Feiyu Yao, Yuhe Ji, Shimin Tao, Jingzhou Du, Justin Li, Jian Gao, Zhang Li, Hao Yang, Boxing Chen, and Osamu Yoshie. Taming text-to-image synthesis for novices: User-centric prompt generation via multi-turn guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8794–8811, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. SELF-REFINE: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *Transactions on Machine Learning Research*, 2024.

Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Boyd-Graber. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13342–13375. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.780. URL https://aclanthology.org/2024.findings-emnlp.780/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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.

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 (CVPR)*, pp. 14895–14904, June 2026.

Zihang Wang, Yilun Zhao, Kaiyan Zhang, Chen Zhao, Manasi Patwardhan, and Arman Cohan. SciSketch: An open-source framework for automated schematic diagram generation in scientific papers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 403–417. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-demos.28. URL https://aclanthology.org/2025.emnlp-demos.28/.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 14030–14057, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/176eb6c51427bd7581d514aaf2a365ed-Paper-Conference.pdf.

Shiyu Wu, Mingzhen Sun, Weining Wang, Yequan Wang, and Jing Liu. Visualprompter: Semantic-aware prompt optimization with visual feedback for text-to-image synthesis. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, and Di Zhang. Paragraph-to-image generation with information-enriched diffusion model. *International Journal of Computer Vision*, 133(8):5413–5434, 2025.

Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. Promptsculptor: Multi-agent based text-to-image prompt optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 774–786, 2025.

Wen Ye, Zhaocheng Liu, Yuwei Gui, Tingyu Yuan, Yunyue Su, Bowen Fang, Chaoyang Zhao, Qiang Liu, and Liang Wang. Genpilot: A multi-agent system for test-time prompt optimization in image generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 929–958, 2025.

Jinlu Zhang, Jiji Tang, Rongsheng Zhang, Tangjie Lv, and Xiaoshuai Sun. Storyweaver: A unified world model for knowledge-enhanced story character customization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9951–9959, 2025.

Table 6: Agent interfaces in the implemented Passage-Level PRISM graph. Structured outputs, rather than free-form dialogue history, define each interface.

Agent	Required input	Structured output	Principal invariant
Style	Passage and all QA pairs	12-field *style contract* s; *style calibration prompt* κ	Appearance is shared, while the calibration sheet contains separated studies and no answer-specific event, relation, label, or text.
Content	Passage and all QA pairs	Content plans $\mathcal{E}$: three candidate scenes, their strengths and faithfulness risks, one verbatim selected scene, the selection reason, and excluded content per question	The selected scene is copied verbatim from the candidates and remains passage grounded, renderer feasible, and nonrevealing.
Illustration	x, $\mathcal{E}$, s, κ, and optional review feedback	Ordered *renderer prompts* $\mathcal{P}$	Selected scenes control semantics; s controls appearance. Each prompt is at most 160 words; only leaked prompts are revised.
Entity Routing	Passage, all QA pairs, and completed $\mathcal{P}$	Canonical entity bank $\mathcal{B}$ and local mappings that determine routes $\mathcal{R}$	Identify K_{ent} visible entities per prompt, merge recurring identities globally, and preserve one route for each QA pair.
Rendering	s, κ, $\mathcal{P}$, $\mathcal{B}$, and $\mathcal{R}$	I^{sty}, $\{J_j\}$, $\{S(r_i)\}$, and $\mathcal{I}$	Render the style and canonical entities once, route exactly two conditioners per question, and retain P_i as semantic authority.
Review	Passage, all QA pairs, private answers, and $\mathcal{I}$	Ordered leakage decisions and reasons	Flag visible answer disclosure, not legitimate evidence from which a reader may infer an answer.

Table 7: Renderer-facing template of the 12-field *style contract* s.

Field	Required visual control	Example
`medium`	Rendering medium and material character.	Watercolor on textured paper.
`linework`	Edge weight, contour behavior, and stroke quality.	Fine contours with occasional ink accents.
`shape_language`	Geometric or organic form vocabulary.	Rounded organic forms with simple geometry.
`palette`	Shared hue range, saturation, and contrast.	Muted teal, ochre, and cream.
`lighting`	Direction, softness, and tonal organization.	Soft, diffuse light from the upper left.
`texture`	Surface detail and mark-making density.	Subtle paper grain and sparse brush marks.
`depth_treatment`	Perspective, layering, and spatial depth.	Layered foreground and background with shallow perspective.
`character_rendering`	Consistent treatment of people or creatures.	Simplified faces and consistent clothing treatment.
`object_rendering`	Consistent treatment of artifacts and props.	Clear silhouettes with restrained surface detail.
`background_treatment`	Environmental abstraction and background detail.	Simplified settings with low-contrast details.
`composition_language`	Framing, balance, and visual hierarchy.	Centered focal area with balanced negative space.
`detail_level`	Global granularity and simplification level.	Moderate focal detail; simplified surroundings.

A PROMPT CONTRACTS AND AGENT INTERFACES

A.1 PASSAGE-LEVEL EXECUTION CONTRACT

Each PRISM run processes a passage and its complete ordered QA pair set through the implemented graph `style_plan` → `content_plan` → `prompt_batch` → `object_plan` → `reference_bank` → `generate_batch` → `review_batch`. Reasoning Agents return schema-validated JSON batches covering every question. Each illustration should expose passage-grounded evidence and support independent reasoning without declaring the answer: private answers may guide planning, but prompts must exclude answer text, option markers, selection cues, and answer-key compositions. Review flags visible leakage and triggers targeted prompt revision before the affected illustration is rendered again.

A.2 STYLE CONTRACT TEMPLATE

Table 7 gives the exact keys enforced by the Style Agent schema. Each value must be nonempty, concise, renderer relevant, and concerned with appearance rather than answer content. Separating these controls makes the shared visual language inspectable and prevents a single prose description from silently mixing style with question-specific semantics.

A.3 PROMPT-AWARE GLOBAL-TO-LOCAL ROUTING

After all renderer prompts $\mathcal{P}$ are complete, the Entity Routing Agent selects K_{ent} prompt-grounded visible identity anchors for each P_i and records their evidence and selection reason. It merges

recurring anchors into the canonical bank $\mathcal{B}$, assigning each identity one shared name, ID, canonical specification, and reference image. The runtime validates K_{ent} distinct legal IDs per question, removes unused entities, renumbers the retained bank, and derives each ordered route r_i from its local mappings in their original order.

Each canonical entry stores its type, canonical appearance, silhouette, colors and materials, distinctive features, mutable and immutable attributes, negative constraints, and an isolated reference prompt. Rendering receives only $S(r_i)$, a label-free sheet of the routed references (2×2 for $K_{\text{ent}} = 4$), keeping unrelated entities out of I_i.

A.4 Agent Prompt Templates

The following templates expose the three components of each implemented call: the persistent System Prompt, the per-passage input template, and the required output contract. Bracketed terms are runtime placeholders; repeated JSON fields are abbreviated for readability. The boxes follow the same Agent order and notation as Section 3.2. They show the reported configuration with $K_{\text{ent}} = 4$. The JSON field `style_reference_prompt` is the style calibration prompt κ.

A.4.1 Style Agent Prompt

Style Agent $\mathcal{A}_{\text{sty}}$**:**

System Prompt

```
You are an art director for educational reading-comprehension
illustrations.  Produce one structured passage-wide Style Contract and
one safe passage-specific calibration-sheet prompt.  Return JSON only.
```

User Input Template

```
Inspect the complete passage and all QA pairs.
Passage:  [PASSAGE]
QA pairs (private planning context):  [ORDERED QA PAIRS WITH ANSWERS]
Define one coherent visual language and a calibration sheet of separated
visual studies.  Exclude answer-specific events, relations, diagrams,
labels, letters, numbers, and explanatory text.
```

Required Output

```
{"style_contract":  [12 FIELD CONTRACT], "style_reference_prompt":
[RENDERER READY PROMPT]}
```

A.4.2 Content Agent Prompt

Content Agent $\mathcal{A}_{\text{con}}$**:**

System Prompt

```
You are an educational illustration content director.  Decide what each
image should depict and return JSON only.
```

User Input Template

```
Passage:  [PASSAGE]
QA pairs with explicitly marked correct answers:  [ORDERED QA PAIRS WITH
ANSWERS]
For each QA pair, propose exactly three materially different,
passage-grounded, renderer-feasible scenes; assess strength and
faithfulness risk; select one verbatim; and record excluded content.
```

Required Output

```
{"plans":  [{"qa_id":  [ID], "learning_goal":  [GOAL],
"candidate_scenes":  [THREE CANDIDATES], "selected_scene":  [SCENE],
"selection_reason":  [REASON], "excluded_content":  [LIST]}, ...]}
```

A.4.3 ILLUSTRATION AGENT PROMPT

Illustration Agent $\mathcal{A}_{\text{ill}}$**:**

System Prompt

```
You are the Passage-Level T2I prompt writer. Produce one renderer-ready
prompt per supplied Question ID and return only valid JSON with no
Markdown fences. Follow the globally supplied task objective for all
semantic goals, optimization priorities, and task boundaries.
```

User Input Template

```
Reading material: [PASSAGE]
Targets: [ORDERED QA PAIRS WITH PRIVATE ANSWERS]
Content plans: [SELECTED SCENES AND EXCLUSIONS]
Style context: [STYLE CONTRACT; CALIBRATION PROMPT]
Review feedback: [VISIBLE LEAKAGE CUE OR NONE]
Preserve the selected scene's semantics, apply style only to appearance,
keep each prompt within 160 words, and revise only the affected prompt
if leakage is reported.
```

Required Output

```
{"prompts_by_id": {[QA ID]: [COMPLETE T2I PROMPT], ...}}
```

A.4.4 ENTITY ROUTING AGENT PROMPT

Entity Routing Agent $\mathcal{A}_{\text{obj}}$**:**

System Prompt

```
You are a visual asset director for educational illustrations. Build a
comprehensive reusable reference bank and return JSON only.
```

User Input Template

```
Completed T2I prompts (primary authority): [ORDERED RENDERER PROMPTS]
Passage (grounding context): [PASSAGE]
QA pairs (private grounding context): [ORDERED QA PAIRS WITH ANSWERS]
For each prompt, extract exactly four visible identity anchors; then
consolidate recurring anchors into one global canonical bank. Local
mappings must use the global name and ID.
```

Required Output

```
{"objects": [CANONICAL ENTITY CONTRACTS], "question_entity_plans":
[FOUR LOCAL MAPPINGS PER QA]}
The runtime deterministically derives [ORDERED FOUR-ID ROUTES] from
these mappings.
```

A.4.5 RENDERING AGENT INTERFACE

Rendering Agent $\mathcal{A}_{\text{rnd}}$**:**

System Prompt

```
N/A: this interface contains deterministic reference construction and
T2I rendering, not a language-model call. Its renderer-control
instruction assigns semantic authority to the concrete illustration
prompt.
```

User Input Template

```
Item: [PASSAGE ID]
Style inputs: [STYLE REFERENCE PROMPT; STYLE CONTRACT]
Identity inputs: [CANONICAL ENTITY BANK; ORDERED FOUR-ID ROUTES]
Synthesis inputs: [ORDERED RENDERER PROMPTS]
Render the style reference and each canonical entity once; construct one
label-free 2 × 2 routed sheet for each QA pair; supply only the shared
style reference and its routed identity sheet to each final renderer
call.
```

Required Output

```
[STYLE REFERENCE]; [CANONICAL ENTITY REFERENCES]; [ROUTED SHEETS];
[ORDERED ILLUSTRATION SET]
```

A.4.6 REVIEW AGENT PROMPT

Review Agent $\mathcal{A}_{\text{rev}}$**:**

System Prompt

```
You are the shared leakage-audit component. Apply
objective.private_answer_policy, objective.answer_leakage_control, and
objective.shared_evaluation_standards, and return only valid JSON.
```

User Input Template

```
Reading passage: [PASSAGE]
Items: [ORDERED QA IDS; QUESTIONS; PRIVATE CORRECT ANSWERS]
Illustrations: [ORDERED ILLUSTRATION FILES]
Inspect only directly visible content. Mark leakage only for readable
answer text, an option identifier, an explicit selection mark, or an
unambiguous formal answer annotation; ordinary visual evidence remains
safe.
```

Required Output

```
{"results": [{"qa_id": [ID], "leaked": [BOOLEAN], "reason": [VISIBLE
CUE REASON]}, ...]}
```

The Review Agent never inspects hidden reasoning. When it detects leakage, its reason identifies the visible cue without restating the answer. The full machine-readable objectives and JSON schemas accompany the code release.

B DATASET CONSTRUCTION

B.1 SOURCE COLLECTIONS AND UNIT OF GENERATION

The benchmark combines RACE-High and RACE-Middle (Lai et al., 2017), RACE-C (Liang et al., 2019), and CosmosQA (Huang et al., 2019). RACE-High and RACE-Middle draw from English examinations for high- and middle-school students; RACE-C adds a curriculum-oriented collection, while CosmosQA focuses on short narratives with causal, temporal, or social commonsense questions. We preserve source labels throughout preprocessing and evaluation, so "RACE" never denotes a merger of the three RACE subsets.

The sampling unit is a *passage bundle*: one passage and all retained multiple-choice questions associated with it. Bundling preserves shared style and recurring entity identities across a passage's illustrations while keeping their content question-specific; sampling questions independently would break this set structure. Table 8 shows three retained test bundles with paraphrased synopses and question focuses and abbreviated IDs.

Table 8: Illustrative test bundles. Each bundle contains one passage and multiple questions; answers and options are omitted.

Source	Record ID	Passage synopsis	Representative question focuses
RACE-High	high_20908	Competing explanations for the long term decline in crime in the United States.	Author intent, evidence for stricter sentencing, and the effect of demographic change.
RACE-C	racec_1126	Appreciation of the Indian rupee, growing reserves, and caution toward exchange control liberalization.	Effects on exporters, sources of the current account surplus, and reasons for policy caution.
CosmosQA	cosmos_3K3G..Q5L	A narrator visits MGM, waits while companions repeat a ride, and later attends a show.	A counterfactual about not visiting and the likely action before the ride.

B.2 NORMALIZATION AND SCREENING

We convert every source to a common schema containing a stable passage identifier, passage text, source name, and an ordered list of questions. Each question contains a local identifier, question

text, a dictionary of answer options, and the correct option label. Option labels are canonicalized to uppercase letters and question identifiers are reassigned consecutively within each passage. Records with missing passage text, malformed options, an invalid answer label, or duplicated identifiers are removed before sampling.

The source collections are screened before they enter the merged benchmark. For the RACE family, we retain passage bundles that are sufficiently long to support evidence selection and that contain between three and five usable questions; the released screened files contain passages of at least 200 whitespace delimited words. This removes very short passages for which several question specific images would be artificial, as well as unusually large question bundles that make generation cost incomparable. RACE-High, RACE-Middle, and RACE-C are processed independently, so curriculum labels are not collapsed during filtering.

CosmosQA provides questions as separate rows. We first group rows that share the same paragraph identifier, retain only paragraphs with at least two labeled questions, and apply passage and question length checks to remove degenerate or excessively long records. Since labels for the public CosmosQA test set are not released, we partition the screened original training data into our train and development pools and use the screened original validation data as the test pool. The unlabeled public test set is not used. These operations are performed before the common 20% sampling step. We will release the exact retained identifiers and preprocessing script so that selection does not depend on undocumented text matching.

The correct answer is preserved only as privileged planning and evaluation metadata. It is available for locating relevant evidence and auditing answer leakage, but is removed from every renderer facing prompt. Consequently, an image generator never observes a correct option label or answer string.

B.3 Experimental Sampling and Splits

Table 9 reports the normalized and screened source pools before the 20% sample.

Table 9: Screened source pools before stratified experimental sampling. Each cell reports passages/questions.

Source	Train	Development	Test
RACE-High	14,611 / 52,813	810 / 2,931	818 / 2,968
RACE-Middle	2,689 / 10,660	159 / 631	154 / 621
RACE-C	1,943 / 9,060	112 / 529	111 / 528
CosmosQA	2,405 / 5,487	129 / 296	129 / 258
Total	21,648 / 78,020	1,210 / 4,387	1,212 / 4,375

Full image generation over every retained passage is prohibitively expensive. We therefore sample 20% of the passages from every source and every split, without replacement, using a fixed seed of 20260803. If a source contains n passage bundles, we select $\lfloor 0.2n \rfloor$ bundles. All questions belonging to a selected passage are retained. Sampling is performed before any method is run, and every method receives exactly the same passage identifiers. Table 10 reports the resulting distribution.

Table 10: Final benchmark used by the experiments after deterministic 20% sampling. Each cell reports passages/questions.

Source	Train	Development	Test
RACE-High	2,922 / 10,609	162 / 590	163 / 580
RACE-Middle	537 / 2,159	31 / 118	30 / 119
RACE-C	388 / 1,806	22 / 100	22 / 107
CosmosQA	481 / 1,094	25 / 56	25 / 50
Total	4,328 / 15,668	240 / 864	240 / 856

For completeness, Table 11 reports the length and bundle statistics omitted from the compact main paper table.

Table 11: Detailed distribution of the held out test split. #W is mean passage length in whitespace delimited words.

Source	Passages	Questions	Q/P	#W
RACE-High	163	580	3.56	306
RACE-Middle	30	119	3.97	282
RACE-C	22	107	4.86	401
CosmosQA	25	50	2.00	88
Total	240	856	3.57	289

Training examples are used only by the post training variants. The directly deployable version of PRISM does not fit parameters on them. Development examples are used for selecting adapters and analysis settings, and the test set is evaluated only after these choices have been fixed. We report both macro averages over the four sources and pooled passage-level results so that the RACE-High subset cannot determine every conclusion solely through its size.

C Evaluation Protocol and Metric Definitions

C.1 Statistical Protocol

Because methods share the same examples, we assess significance with paired two-sided t tests over passage-level scores. For RQ1's geometric aggregates, we instead use 10,000 paired passage-level permutations that recompute the aggregate after swapping method labels within each passage.

C.2 Automatic Metrics

Table 12 summarizes the reported dimensions. All embedding models, thresholds, and aggregation rules are fixed before comparing methods. Faithfulness uses CLIP (Radford et al., 2021): for image encoder ϕ_{C} and text encoder ψ_{C}, we compute

$$F_P = \frac{100}{n} \sum_i \cos(\phi_{\mathrm{C}}(I_i), \psi_{\mathrm{C}}(p)) \,, \qquad F_Q = \frac{100}{n} \sum_i \cos(\phi_{\mathrm{C}}(I_i), \psi_{\mathrm{C}}(q_i)) \,. \tag{9}$$

The main table reports F_P and F_Q separately. Answer-text similarity is retained only as a diagnostic and is excluded from GeoMean.

For a representation $\phi \in \{\phi_{\mathrm{C}}, \phi_{\mathrm{D}}\}$, where ϕ_{D} is DINO ViT-B/8 (Caron et al., 2021), let s_{ij}^{ϕ} be the cosine similarity between I_i and I_j. Consistency uses 100 times the mean of $h(s) = s$ for $s \le \gamma$ and $h(s) = 2\gamma - s$ otherwise, with $\gamma = 0.85$; thus, near-duplicate pairs cannot inflate consistency without penalty. Diversity is

$$D_{\phi} = 100 \left(1 - \binom{n}{2}^{-1} \sum_{i<j} \mathbf{1}[s_{ij}^{\phi} > \gamma] \right) . \tag{10}$$

We report CLIP and DINO diversity in the main comparison. The aesthetic score applies the published LAION MLP regressor to normalized CLIP ViT-L/14 embeddings and rescales its 0–10 output to 0–100 (LAION-AI, 2022).

For reproducibility, the evaluation code uses the SentenceTransformer IDs `sentence-transformers/clip-ViT-B-32` for image embeddings and `sentence-transformers/clip-ViT-B-32-multilingual-v1` for passage, question, and answer text embeddings; the paper runner loads local cached copies of these checkpoints. DINO uses the local hub model `dino_vitb8` with checkpoint `dino_vitbase8_pretrain.pth`. The aesthetic branch loads the OpenAI CLIP checkpoint `ViT-L-14.pt` and the LAION predictor checkpoint `sac+logos+ava1-l14-linearMSE.pth`.

Table 12: Automatic evaluation dimensions. Pairwise metrics are computed within each passage before dataset-level aggregation.

Dimension	Backbone / input	Definition	Intended interpretation
Faithfulness	CLIP ViT-B/32; (I_i, p) and (I_i, q_i)	Mean cosine similarity for passage and question text, reported separately and jointly.	Passage grounding and question relevance.
Consistency	CLIP ViT-B/32 and DINO ViT-B/8; all $i < j$	Mean pairwise cosine after symmetric near-duplicate penalty at $\gamma = 0.85$.	Shared visual language and recurring identity stability without rewarding copies.
Diversity	CLIP ViT-B/32 and DINO ViT-B/8; all $i < j$	100 minus the percentage of pairs whose cosine exceeds 0.85.	Resistance to duplicated content across distinct questions.
Multimodal Judge	Qwen3.6 27B; passage bundle, category-specific context, and illustrations	Bottom-up L3→L2→L1 scoring; reported LAS is the geometric mean of six L1 category means.	Fine-grained quality, grounding, usefulness, secrecy, and set consistency.
Aesthetic	CLIP ViT-L/14 + LAION MLP	Predicted 0–10 preference score, linearly rescaled and clipped to $[0, 100]$.	Independent visual appeal estimate.

Table 13: L1–L3 design of the multimodal Judge. The L2 column is complete; the L3 column shows representative facets, with the full checklist retained in the released evaluator configuration.

L1 category	Scope	L2 subdimensions	Representative L3 facets
Quality	Each illustration	Detail; Resolution	Noise; Edge Clarity; Artifact Absence; Local Legibility
Aesthetics	Each illustration	Composition; Color Harmony; Lighting; Anatomical Integrity; Emotional Expression; Style Control	Composition; Lighting and Atmosphere; Anatomical Fidelity; Style Coherence
Alignment	Each illustration	Attributes; Actions; Layout; Relations; Scene; Passage Grounding; Question Grounding; Prompt Compliance	Attribute and Action Accuracy; Spatial and Relation Accuracy; Passage Faithfulness; Question Relevance; Evidence Coverage
Educational Fidelity	Each illustration	Fairness; Safety; World Knowledge; Educational Reliability; Answer Secrecy	Knowledge Fidelity; Reasoning Helpfulness and Validity; Direct Answer Text; Option Marker, Selection Cue, and Formal Answer Leakage
Creative Generation	Illustration and set	Controlled Imagination; Feature Matching; Logical Resolution; Text Rendering; Task Compliance; Visual Storytelling; Passage Set	Controlled Creativity; No Unwanted Text; Narrative Clarity; Cross-Image Complementarity; Question Distinctiveness
Style Consistency	Passage set	Visual Style; Character Consistency; Object Consistency; Scene Consistency; Narrative Continuity; Graphic Language	Rendering Style; Identity and Object Consistency; Spatial and Temporal Continuity; Layout Grammar

C.3 Hierarchical Multimodal Judge

Our Judge adapts the three-level organization of Qwen-Image-Bench (Li et al., 2026) to QGVS. L1 denotes a broad evaluation category, L2 a coherent subdimension, and L3 a directly observable checklist item. Table 13 lists the complete L1 and L2 structure and representative L3 facets. We add task-specific passage grounding, question grounding, answer secrecy, educational reliability, passage-set complementarity, and a separate passage-level Style Consistency category.

Each applicable L3 facet receives a raw score in $\{0, 1, 2, 3, 4\}$, mapped to $\{0, 25, 50, 75, 100\}$; N/A is excluded. Scores mean clear failure, borderline reject, baseline pass, clear acceptance, and observable excellence, respectively. L3 scores are averaged within each L2 group and L2 scores within each L1 category. The LAS score reported in the main tables is the geometric mean of the six dataset-level L1 means. Style Consistency is evaluated once on the complete passage set: the Judge first identifies anchors that genuinely recur and does not penalize changes required by different questions, viewpoints, times, or actions.

Passage contradiction, safety failure, and four forms of answer leakage are non-compensatory hard gates. A raw score of 1 caps the affected L1 and overall score at 50; a raw score of 0 caps them at 25. Alignment and Educational Fidelity receive private answer references solely to evaluate grounding and leakage. Aesthetics, Alignment, Creative Generation, and Style Consistency also receive renderer prompts where prompt compliance is relevant. The Judge never receives the method identity or hidden planner state. It must return the complete JSON schema. The evaluator also retains a separately capped per-passage overall score; L1, L2, L3, and hard-gate records are retained for analysis.

C.4 HUMAN EVALUATION RUBRIC

The human evaluation covers a fixed, randomly selected subset of 50 test passage bundles and all 174 associated questions. Each assessment sheet examines the passage, questions, answer options, correct-answer references, and the illustration set from one method at a time. The five quality dimensions are scored separately for each passage–method set; the 1–5 scale uses 1 for very poor, 2 for poor, 3 for neutral, 4 for good, and 5 for very good. Equal scores across methods are allowed.

Table 14: Criteria supplied for the human evaluation. The first five dimensions receive independent 1–5 scores for the passage illustration set; Answer Leakage is assessed separately for each question illustration.

Dimension	Criterion
Faithfulness	Whether the illustrations faithfully depict the passage and their corresponding questions.
Consistency	Whether illustrations for different questions in the same passage share a coherent visual style.
Diversity	Whether the illustrations depict meaningfully different content beyond stylistic variation, while each remains relevant to its own question.
Aesthetic	Whether the illustrations are visually appealing and suitable as illustrations, without ghosting or other visible artifacts.
Helpfulness	Whether the illustrations give useful visual cues for answering their questions and identifying the correct answers.
Answer Leakage	Yes only when visible text in a question illustration directly states or explicitly indicates the correct answer; otherwise no. Visual evidence alone is not leakage.

For each quality dimension and method, Table 5 averages all three recorded scores over all 50 passages (150 scores). Overall is the arithmetic mean of the five dimension means, with equal weight and no leakage term. To assess agreement on the ordinal ratings, we calculate quadratically weighted Cohen's κ for each of the three pairs of assessment sheets over the same 250 passage–method items within each dimension, then average the three coefficients. The resulting values are 0.64 (faithfulness), 0.93 (consistency), 0.53 (diversity), 0.59 (aesthetics), and 0.64 (helpfulness); their macro average is 0.67. For leakage, each question–method pair has three binary judgments; a pair counts as leaking if at least two judgments are yes. We divide this count by 174 questions per method. We report descriptive means and rates without inferential tests on the assessment sheets.

C.5 MCQA READING COMPREHENSION DIAGNOSTIC

We use 240 held-out passages to evaluate four student backbones: Qwen3.5 4B, Qwen3.5 9B, Qwen3.8 27B, and InternVL3.5 8B. For each question, method, and simulated reading profile, the student answers under three evidence conditions: passage text only, illustration only, and passage text plus the corresponding illustration. The reported gain is the difference between the last and first conditions in exact option-letter accuracy. The question and answer options are supplied in all conditions, but the correct answer is never given to the student. Text-only responses are shared across methods for a fixed student and profile; each illustrated condition uses the same question and passage budget as its text-only control.

The strong profile receives the full passage and instructions to integrate evidence across sentences. The developing and beginning profiles receive question-aware, answer-blind extractive passage budgets of approximately 50% and 25%, respectively, with instructions limiting the complexity of reasoning chains. Thus, the profiles jointly vary reading strategy and available text; they are not calibrated estimates of human ability. Each student-backbone/profile condition has a theoretical maximum of 856 questions. For Figure 3, we retain only questions for which each method returns a valid response in both the text-only and text-plus-illustration conditions; response failures therefore reduce the effective denominator. This filter may select easier questions, so the plotted, question-weighted gains are conditional on valid responses; they are not claims of significant method differences or effects on human learners. The plotting code and source MCQA results are retained with the paper artifacts.

Table 15: Main inference configuration.

Component	Setting
Planner	Qwen3.5 9B
Renderer	FLUX.2 Klein 9B KV
Serving	vLLM / Diffusers
Image size	512×512
Denoising steps	4
Guidance scale	1.0
T2I sequence length	512
Base image seed	42
Planning temperature	0.7
Review Agent temperature	0.2
Nucleus probability	0.95 planning; 0.90 review
Maximum prompt length	160 words

D Implementation and Optimization Details

D.1 Direct Inference Configuration

Table 15 records the main configuration. Qwen3.5 9B is served with vLLM and FLUX.2 Klein 9B with Diffusers. The two models remain frozen for every direct inference method. Planning uses nucleus sampling, whereas the Review Agent uses a lower temperature. All methods share the same renderer checkpoint and resolution unless an experiment explicitly studies a different backbone.

The maximum new token budgets are 1,024 for style planning, 6,144 for content planning, 4,096 for prompt construction, 8,192 for entity planning, and 512 for leakage review. Experiments are executed on two NVIDIA RTX Pro 6000 GPUs with 96GB memory each.

D.2 Adaptation of Comparison Methods

Every comparison method receives the same passage bundle, answer options, and private answers, and is subject to the same no leakage policy. PFQS adds question planning before constructing one renderer prompt per question. PromptRefine, OPT2I, DialPrompt, VisualPrompter, PromptSculptor, and GenPilot obtain an initial prompt from the RC inputs before applying their respective refinement steps. RePrompt and PromptEnhancer likewise begin from a task-derived prompt before method specific adaptation or post training. Methods originally designed for one image are executed once per question and receive the other prompts and images only when their published mechanism supports cross image feedback.

We match the maximum number of renderer calls when comparing iterative methods. A method may terminate early according to its own stopping criterion, but it cannot generate additional candidates after exhausting the shared budget. All baseline outputs pass through the same leakage audit for evaluation.

D.3 Post Training Configuration

Policy optimization changes only low rank adapters attached to the planner. The base VLM, renderer, evaluation models, and Review Agent remain frozen. Each controlled rollout group contains four complete candidates. One LoRA adapter is shared by all four planning Agents and optimized for two iterations of 240 sampled training passages. Preliminary experiments tested up to three GRPO epochs per optimization iteration. Across most tested settings, the best development-set checkpoint occurred after the first epoch; accordingly, all reported GRPO runs use one epoch per iteration. Table 16 lists the optimization settings.

The hybrid reward assigns weight 0.4 to an auxiliary QA signal and 0.6 to the prompt blind multimodal Judge. QA regression beyond 0.05 incurs a penalty equal to 0.1 times the excess. Groups whose reward standard deviation is below 0.02 are excluded because their normalized advantages are unstable; at most 20% of the groups may be excluded. Adapter selection uses the development

Table 16: Shared adapter GRPO hyperparameters. Learning rates follow the first and second optimization iterations.

Hyperparameter	Value
Candidates per group	4
Training passages per iteration	240
GRPO epochs	1 per iteration
Shared adapter learning rates	$5.0\mathrm{e}{-6}$, $2.5\mathrm{e}{-6}$
PPO clip range	0.2
KL coefficient	0.04
Target KL	0.01
Batch / gradient accumulation	1 / 4
Maximum sequence length	8,192
Gradient norm	1.0
LoRA rank / scale / dropout	16 / 32 / 0
Precision	bfloat16
Training and shuffle seed	20260914

subset and a combination of the Judge and auxiliary question utility signals. These selection signals are reported separately and are not part of GeoMean.

D.4 EVALUATION MODELS

The prompt blind Judge is Qwen3.6 27B FP8. The main visual evaluation uses CLIP ViT B/32 for text image alignment, CLIP and DINO ViT B/8 for cross image consistency and diversity, and a CLIP ViT L/14 based aesthetic regressor. For post training diagnostics only, an auxiliary QA signal compares correct answer probability with and without the image using Qwen3.5 9B and Qwen3.8 27B NVFP4; this signal is not included in GeoMean. Model identifiers, evaluator prompts, and per example outputs are stored with every evaluation artifact.

D.5 COMPUTE AND COST ACCOUNTING

We count calls rather than estimating cost from nominal context length. For a passage with n questions and b canonical entity references, a successful single pass through PRISM makes four planning agent calls, one review call, and $1 + b + n$ image generation calls: one style reference, b entity references, and one final image per question. We log planner tokens, renderer calls, wall clock time, and peak allocated GPU memory for every passage. Training cost additionally reports complete rollout trajectories and optimizer GPU hours.

E CROSS BACKBONE RESULTS

Table 17 expands the aggregate results in Table 3. The InternVL3.5 pairings yield higher DINO consistency but lower DINO diversity for PRISM, illustrating why the aggregate should be read alongside its components.

Table 17: Detailed RQ4 results with FLUX.2 Klein renderers. The Judge column reports the aggregate of its six dimensions.

				Faithfulness		Consistency		Diversity		Judge	Aesthetic
Planner	Renderer	Method	***GeoMean***	F_P	F_Q	CLIP-P	DINO-P	CLIP	DINO	Score	Score
Qwen3.8 27B	9B	PromptRefine	60.1	21.5	23.2	63.8	35.2	94.7	98.4	77.8	57.7
Qwen3.8 27B	9B	PRISM	**66.2**	22.3	23.8	74.6	69.7	81.5	81.7	87.0	58.7
Qwen3.5 9B	9B	PromptRefine	60.6	21.6	23.4	63.3	38.6	93.3	97.3	78.4	57.6
Qwen3.5 9B	9B	PRISM	**66.0**	22.4	23.9	73.1	63.6	86.0	91.9	84.9	61.3
Qwen3.5 9B	4B	PromptRefine	59.2	21.6	23.4	63.3	39.0	91.7	96.4	75.1	55.6
Qwen3.5 9B	4B	PRISM	**65.5**	22.4	23.6	72.7	60.6	86.0	93.2	84.4	60.8
InternVL3.5 8B	9B	PromptRefine	57.8	21.2	23.6	62.0	43.1	93.6	97.0	70.1	56.0
InternVL3.5 8B	9B	PRISM	**62.8**	22.4	23.8	73.7	71.1	67.0	68.1	82.1	60.8
InternVL3.5 8B	4B	PromptRefine	57.6	21.2	23.6	61.4	42.8	95.2	97.5	69.7	55.7
InternVL3.5 8B	4B	PRISM	**61.2**	22.3	23.7	73.5	68.8	64.8	65.5	79.2	60.5

F RESULTS BY READING SOURCE

Tables 18 and 19 break down the pooled results in Tables 1 and 4 by the four benchmark sources. Each component is averaged over valid passages within a source; LAS and GeoMean use the same geometric aggregation as the main tables. Runtime remains pooled. Source-specific GeoMeans need not average to the pooled GeoMean.

Table 18: Per-source direct QGVS results corresponding to Table 1. Bold marks the highest value within each source. All entries are quality scores; runtime is reported only for the pooled set in Table 1.

Method	***GeoMean***	F_P	F_Q	CLIP-P	DINO-P	CLIP	DINO	LAS	Aesthetic
RACE-High (163 passages, 580 questions)									
PFQS	56.9	21.4	23.6	58.8	33.9	**99.7**	99.7	69.7	57.6
OPT2I	57.5	21.4	23.4	58.8	31.8	98.8	**100.0**	72.3	57.3
VisualPrompter	58.2	21.6	23.5	60.0	33.2	98.5	**100.0**	73.2	57.9
GenPilot	58.7	21.9	23.3	65.2	43.8	93.5	99.6	70.6	58.9
PromptSculptor	59.2	21.8	23.3	63.6	39.1	96.4	**100.0**	73.1	59.5
DialPrompt	59.8	21.5	23.2	60.9	33.1	98.4	99.7	77.5	58.5
PromptRefine	60.4	21.5	23.3	62.6	37.3	94.6	98.8	78.0	57.5
PRISM (direct)	**65.9**	**22.4**	**23.9**	**72.6**	**63.8**	87.5	92.8	**84.2**	**61.0**
RACE-Middle (30 passages, 119 questions)									
PFQS	59.4	22.2	24.0	61.0	35.6	94.1	97.4	75.4	59.3
OPT2I	59.2	22.1	24.1	60.7	35.4	**97.7**	99.3	74.2	59.3
VisualPrompter	59.6	22.2	24.2	61.2	33.8	96.4	99.3	76.0	58.7
GenPilot	62.0	22.5	24.1	69.7	49.5	88.7	95.6	77.1	60.3
PromptSculptor	61.4	22.5	24.3	65.5	43.1	91.3	95.8	77.1	61.8
DialPrompt	63.1	22.2	24.3	65.6	40.1	97.0	**100.0**	82.0	60.0
PromptRefine	61.7	22.2	24.0	64.5	39.3	95.6	98.7	79.2	59.7
PRISM (direct)	**67.1**	**23.1**	**24.7**	**73.6**	**62.9**	80.9	91.4	**87.4**	**63.6**
RACE-C (22 passages, 107 questions)									
PFQS	55.3	21.1	23.2	56.3	32.5	**99.5**	99.5	66.8	57.4
OPT2I	53.7	21.0	**23.4**	55.6	26.8	**99.5**	**100.0**	64.9	56.9
VisualPrompter	53.6	21.2	22.8	57.5	30.0	99.1	**100.0**	63.1	58.3
GenPilot	54.4	21.3	23.2	62.4	40.3	96.8	99.5	61.3	58.0
PromptSculptor	56.1	21.3	23.0	61.8	35.8	97.4	**100.0**	66.6	59.5
DialPrompt	57.9	20.8	22.9	59.1	32.3	99.1	**100.0**	73.4	58.6
PromptRefine	57.6	21.2	22.9	62.1	36.2	95.9	99.1	71.0	57.4
PRISM (direct)	**66.0**	**21.9**	23.3	**71.6**	**61.2**	93.5	94.8	**85.1**	**59.7**
CosmosQA (25 passages, 50 questions)									
PFQS	59.6	21.5	**23.9**	64.8	40.2	84.0	**100.0**	75.7	55.8
OPT2I	62.0	21.6	23.5	65.7	43.1	96.0	**100.0**	80.1	55.6
VisualPrompter	60.8	21.6	23.8	61.4	38.2	92.0	**100.0**	79.5	55.8
GenPilot	61.3	21.7	22.7	66.4	43.3	**100.0**	**100.0**	77.6	56.7
PromptSculptor	61.9	**21.9**	23.3	64.9	41.0	**100.0**	**100.0**	79.2	58.4
DialPrompt	62.4	21.5	22.5	68.1	42.0	84.0	96.0	83.5	57.4
PromptRefine	62.7	21.5	23.5	67.4	47.6	80.0	84.0	84.9	56.2
PRISM (direct)	**65.5**	**21.9**	23.0	**77.1**	**65.4**	76.0	84.0	**86.0**	**61.8**

Table 19: Per-source outcome post training results corresponding to Table 4. Bold marks the highest value within each source, including the direct PRISM comparison in the final row.

Method	***GeoMean***	F_P	F_Q	CLIP-P	DINO-P	CLIP	DINO	LAS	Aesthetic
RACE-High (163 passages, 580 questions)									
RLOO	56.1	21.5	23.6	58.4	32.8	99.3	**100.0**	68.1	57.3
PPO	56.5	21.4	23.5	58.8	32.6	97.6	98.7	69.5	57.4
DPO	57.6	21.5	23.5	58.4	32.6	99.0	99.9	72.1	57.4
PromptEnhancer	58.2	21.6	23.3	60.5	33.9	99.0	99.6	72.7	58.4
RePrompt	59.7	21.5	23.2	61.4	33.0	**99.5**	99.8	76.8	59.0
PRISM (GRPO)	**66.4**	22.3	**23.9**	71.7	62.4	92.6	95.6	**85.0**	**61.2**
PRISM (direct)	65.9	**22.4**	**23.9**	**72.6**	**63.8**	87.5	92.8	84.2	61.0
RACE-Middle (30 passages, 119 questions)									
RLOO	59.8	22.1	24.3	62.4	36.9	**97.1**	98.0	75.3	59.2
PPO	58.9	22.1	24.2	61.7	36.1	95.0	97.3	73.6	58.4
DPO	60.8	22.0	24.2	62.9	37.3	95.4	96.9	78.2	59.2
PromptEnhancer	60.0	22.1	23.8	62.5	36.3	96.2	**99.0**	75.9	59.6
RePrompt	62.9	21.9	23.7	64.6	38.3	93.8	98.0	83.5	60.5
PRISM (GRPO)	**67.4**	**23.1**	**24.8**	72.1	62.0	88.3	93.8	87.3	62.9
PRISM (direct)	67.1	**23.1**	24.7	**73.6**	**62.9**	80.9	91.4	**87.4**	**63.6**
RACE-C (22 passages, 107 questions)									
RLOO	52.9	21.0	**23.4**	56.7	31.9	96.4	97.3	61.5	57.0
PPO	53.1	20.8	23.1	55.4	29.6	99.1	**100.0**	62.5	56.7
DPO	55.2	20.7	23.3	56.4	31.9	99.1	**100.0**	67.1	56.8
PromptEnhancer	55.1	21.1	23.1	58.3	28.0	97.7	**100.0**	67.6	58.1
RePrompt	58.9	21.1	22.8	58.6	31.3	**99.5**	**100.0**	76.3	59.0
PRISM (GRPO)	65.6	**22.0**	23.2	70.7	**62.6**	92.3	95.9	83.7	**59.9**
PRISM (direct)	**66.0**	21.9	23.3	**71.6**	61.2	93.5	94.8	**85.1**	59.7
CosmosQA (25 passages, 50 questions)									
RLOO	60.3	21.7	**23.3**	61.7	38.9	96.0	**100.0**	77.1	56.5
PPO	60.0	21.2	**23.3**	60.8	36.8	**100.0**	**100.0**	77.0	56.2
DPO	61.4	21.5	**23.3**	64.3	39.2	96.0	**100.0**	79.8	56.3
PromptEnhancer	62.2	21.4	23.1	64.8	39.1	96.0	**100.0**	82.0	56.8
RePrompt	62.7	21.7	23.0	65.9	41.7	88.0	**100.0**	83.1	58.6
PRISM (GRPO)	**65.5**	**22.0**	22.8	74.5	**65.9**	76.0	88.0	85.9	**61.9**
PRISM (direct)	**65.5**	21.9	23.0	**77.1**	65.4	76.0	84.0	**86.0**	61.8

Direct PRISM has the highest GeoMean among direct methods on each source. The margin over the strongest direct comparison ranges from 2.8 points on CosmosQA to 8.1 points on RACE-C. The GRPO variant is slightly higher on RACE-High and RACE-Middle, lower on RACE-C, and nearly unchanged on CosmosQA. These source-level comparisons are descriptive; we do not claim separate significance on the smaller splits.

G OUTCOME POST TRAINING PROTOCOL

We study post training as a diagnostic extension of PRISM, not as a component required by the inference framework. Let $\mathcal{M} = \{\mathrm{sty}, \mathrm{con}, \mathrm{ill}, \mathrm{obj}\}$ index the Style Agent, Content Agent, Illustration Agent, and Entity Routing Agent, and let $\pi_{\boldsymbol{\theta}}$ denote their common planner equipped with one shared LoRA adapter. Agent-specific instructions and schemas condition the same policy for each $m \in \mathcal{M}$. A complete execution produces a trajectory $\tau = (a_{\mathrm{sty}}, a_{\mathrm{con}}, a_{\mathrm{ill}}, a_{\mathrm{obj}}, \mathcal{I})$, where a_m collects Agent m's structured textual actions and $\mathcal{I} = \{I_i\}_{i=1}^{n}$ is the resulting illustration set. Rewards are computed only after reference generation, image synthesis, and safety review.

G.1 OUTCOME REWARD

For each passage, we sample a group of K complete candidate trajectories. Candidate k receives two complementary signals. First, an image aware QA model measures whether the generated images improve answer discrimination relative to solving from text alone. For question q_i, let m_i^0 and p_i^0 be the log probability margin and probability of the correct option without an image, and let m_i^k and p_i^k be their counterparts with image I_i^k. We define

$$R_k^{\mathrm{qa}} = \frac{1}{n}\sum_{i=1}^{n} w_i \left[\mathrm{clip}\left(\frac{m_i^k - m_i^0}{c_m}, -1, 1\right) + \eta\ \mathrm{clip}\left(\frac{p_i^k - p_i^0}{c_p}, -1, 1\right)\right], \tag{11}$$

where m_i is the correct-option log probability minus the largest incorrect-option log probability. We use $c_m = 2$, $c_p = 0.05$, and $\eta = 0.1$. The difficulty weight is $w_i = 0.25 + 0.75/(1+\exp(\mathrm{clip}((m_i^0 - 3)/2, -20, 20)))$, giving more weight to questions the QA model cannot already solve confidently from text. Second, a prompt blind multimodal Judge scores question relevance, evidence coverage, passage faithfulness, relation accuracy, reasoning helpfulness, reasoning validity, and visual quality. The six nonvisual dimensions each receive weight 0.15, and visual quality receives 0.10; if relation accuracy is N/A, the remaining weights are renormalized. Direct answer disclosure or passage contradiction sets its score R_k^{judge} to the minimum value.

Because the evaluators have different scales, we standardize them within each passage group. Let $z(R_k^{\mathrm{qa}})$ and $z(R_k^{\mathrm{judge}})$ denote the standardized scores. If σ_{qa} is the within-group standard deviation of the QA signal, we set $\omega_{\mathrm{qa}} = \min(\sigma_{\mathrm{qa}}/0.05, 1)$ to reduce its weight when variation is too small to be reliable. The outcome reward is

$$R_k = \alpha\omega_{\mathrm{qa}} z(R_k^{\mathrm{qa}}) + (1-\alpha) z(R_k^{\mathrm{judge}}) - \lambda\Delta_k^{\mathrm{qa}}, \qquad \Delta_k^{\mathrm{qa}} = \max\left(0, -R_k^{\mathrm{qa}} - \delta\right). \tag{12}$$

We use $\alpha = 0.4$, $\lambda = 0.1$, and $\delta = 0.05$. The final term penalizes regression beyond the margin in the auxiliary QA signal. The critic free group relative advantage is

$$\widehat{A}_k = \frac{R_k - \mu_R(x)}{\sigma_R(x) + \epsilon}, \tag{13}$$

where $\mu_R(x)$ and $\sigma_R(x)$ are the mean and standard deviation of valid candidates for instance x, with $\epsilon = 10^{-8}$ for numerical stability. For a zero-variance evaluator signal, its standardized score is set to zero.

G.2 SHARED ONE ADAPTER GRPO

All four planning interfaces are routed to the same adapter. For each passage, we retain the structured responses produced by every Agent in each complete trajectory; reference construction, rendering, and review are executed before the trajectory receives a reward. If $a_{k,m,j}$ denotes visit j to Agent m in candidate k, the shared adapter is optimized with

$$\mathcal{L}_{\mathrm{shared}} = -\mathbb{E}_{k,m,j,t}\left[w_{k,m,j}\min\left(\rho_{k,m,j,t}\widehat{A}_k, \mathrm{clip}(\rho_{k,m,j,t}, 1-\varepsilon, 1+\varepsilon)\widehat{A}_k\right)\right] + \beta\widehat{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\mathrm{old}}), \tag{14}$$

where $\rho_{k,m,j,t} = \pi_{\boldsymbol{\theta}}(a_{k,m,j,t} \mid h_{k,m,j,t})/\pi_{\mathrm{old}}(a_{k,m,j,t} \mid h_{k,m,j,t})$. Every action in a trajectory shares $\widehat{A}_k$. We set $w_{k,m,j} = 1/J_{k,m}$ for the $J_{k,m}$ visits to Agent m, so repeated review cycles do not give one interface disproportionate weight. Consequently, each Agent contributes the same total example weight per complete trajectory while all gradients update the same parameters $\boldsymbol{\theta}$.

Optimization uses two rounds. The first round initializes a fresh shared LoRA from the frozen base planner and reuses the fixed complete-trajectory batch prepared for the controlled global comparison. This is an explicitly offline first update: its stored actions, rewards, and advantages remain fixed, while their log probabilities are evaluated under a frozen copy of the initial policy. In the second round, the updated shared adapter is bound to all four Agents, produces a fresh K-candidate rollout group, and is further optimized using newly computed outcomes. The renderer, evaluators, Rendering Agent, and Review Agent remain frozen throughout. We evaluate the resulting policy as an alternative operating point rather than assuming that the aggregate visual score must increase.

H CASE STUDY

We draw five passage bundles from the 50 completed Human Evaluation cases, stratifying across the four sources. Each case presents the full original passage on a separate page, followed by its original questions and the same five methods in the order of Table 5. Each image row is one question from the same passage. No panel is edited or cherry-picked within a passage. These examples are descriptive. Correct options are highlighted in blue in the question tables; options preserve the source dataset wording.

Across Figures 5–13, PRISM generally maintains a common visual language while changing the depicted scene with the question. Yet consistency does not guarantee that a scene expresses an inference: the examples involving a writer's stance, negation, or the definition of a phrase remain difficult to communicate without text. Some outputs introduce labels or favor a repeated setting over question specific evidence. These cases illustrate why faithfulness, diversity, and answer leakage must be considered together rather than inferred from visual coherence alone.

Case 1: RACE High (high_23800). Original passage.

Joan, super visor of the town's morals recently accused George, a local man, of having a drinking problem because she noticed his car outside the town's only bar.She reasoned, it was a logical assumption .George stared at her for a moment and said nothing.Later that evening, he parked his car in front of her house and left it there all night. Why is it that so many of our logical assumptions are just plain wrong? And why do we want to act as if they must be true? In an age long before the Internet, a young American at a party found himself seated next to the famous Chinese diplomat Wellington Koo.Completely not knowing what to say to someone from such a different culture, this young man turned and asked, "Likee soupee?" After all, he assumed, don't all Chinese speak in broken English? Mr.Koo smiled and nodded.Later when called upon to speak, Wellington Koo delivered a wonderful talk in fluent English, sat down while the applause was still continuing, turned to the young man and whispered, "Likee speechee?" Do you know why red roses are so popular? Men keep buying them.They assume that all women prefer red roses.But when women are asked what color1 of rose is their favorite, they are more likely to answer that they prefer yellow, white or peach.And when asked what kind of flowers they would prefer to receive, the answer is generally not roses at all.The age-old assumption that most women like red roses best is simply not true. We all make assumptions.They're too often wrong and can easily get us into trouble.So here is what I've learned: when in doubt, check it out.And if I'm not in doubt, I still check it out.We may think we know.But what we know may not be so.So it's best to check it out.

Figure 4: Full original passage for Case 1.

Case 1: RACE High (high_23800). Questions.

Q1	According to the passage, we can know George is __ . **A. clever** B. honest C. kind-hearted D. bad-tempered
Q2	Wellington Koo's experience is mentioned to show __ . A. his talent in speaking English **B. people's assumption can be false** C. the Chinese were looked down on D. his humor in doing with difficult situations
Q3	Red roses are so popular at the market in that __ . A. women prefer red roses. B. women like the red color1 best. C. they smell pleasant than others. **D. men think women would like them.**
Q4	How is this passage mainly developed? **A. By giving examples.** B. By presenting research results. C. By analyzing causes. D. By making comparisons.



Figure 5: Complete RACE High passage comparison across the five Human Evaluation methods.

Case 2: RACE Middle (middle_3886). Original passage.

One day in the eighth grade, I was taking a Math Test on basic trigonometry . Now for a middle school student, there was nothing basic at all about it. So I had studied for about two hours the night before. After reviewing it for some time, I had started to understand it, so then I closed the book and _ . However, it was not until midnight that I fell asleep. The next day, when I got the test, it might as well have been written in Serbian (which I don't know how to read). I tried my best to work out the hard problems, but the numbers I came up with were strange. I sat back on my chair and looked for any possible answers, but I could not remember anything or think correctly. I started thinking about my dad coming home from work to find that I failed the test ... "How could you have failed the test? I am certain that nobody else in the whole class got as bad a grade as you did!" Naturally I didn't want that to happen. My dad was also really busy at work at present, so I was afraid that this might make him mad. When feeling hopeless, I noticed that my table partner was writing fast on his test. I could see smoke rising up from how fast he was writing. I was attracted to look over at his test, but then the many bad results I had heard about cheating came into my mind. I reasoned that if I started cheating now, it would be hard to give up that habit during high school. In the end, I decided not to copy his answers, and got a B- on that test. Even though my dad gave me a hard time about it, it would have been a lot worse if he found out that I had given in and cheated.

Figure 6: Full original passage for Case 2.

Case 2: RACE Middle (middle_3886). Questions.

Q1	The writer might feel __ before the Math Test. A. surprised B. relaxed **C. worried** D. excited
Q2	Which is the best expression to fill in the blank in the passage? A. had breakfast **B. went to bed** C. took exercise D. rode to school
Q3	The writer's father was __ . A. proud of him B. tired of him **C. strict with him** D. pleased with him
Q4	From the passage, we know that the writer __ . A. could read Serbian **B. didn't cheat at last** C. got a good grade at last D. didn't work hard that night



Figure 7: Complete RACE Middle passage comparison across the five Human Evaluation methods.

Case 3: RACE C (racec_3586). Original passage.

Millions of people pass through the gates of Disney's entertainment parks in California, Florida and Japan each year. What makes these places an almost universal attraction? What makes foreign kings and queens and other important people want to visit these Disney parks? Well, one reason is the way they're treated once they get there. The people at Disney go out of their way to serve their "guests", as they prefer to call them, and to see that they enjoy themselves. All new employees, from vice presidents to par-time workers, begin their employment by attending Disney University and taking "Traditions 1". Here, they learn about the company's history, how it is managed and why it is successful. They are shown how each department relates to the whole. All employees are shown how their part is important in making the park a success. After passing "Traditions 1", the employees go on to more specialized training for their specific () jobs. No detail is missed. A simple job like taking tickets requires four eight-hour days of training. When one ticket taker was asked why it took so much training for such a simple, ordinary job, he replied, "What happens if someone wants to know where the restrooms are, when the parade starts or what bus to take back to the campgrounds?... We need to know the answers or where to get them quickly. Our constant aim is to help our guests enjoy the party." Even Disney's managers get involved in the daily management of the park. Every year, the managers leave their desks and business suits and put on special service clothes. For a full week, the bosses sell hot dogs or ice cream, take tickets or drive the monorail (), and take up any of the 100 jobs that make the entertainment park come alive. The managers agree that this week helps them to see the company's goals more clearly. All these efforts to serve the public well have made Walt Disney Productions famous. Disney is considered by many as the best mass service provider in America or the world. As one long-time business observer once said, " How Disney treats people, communicates with them, rewards them, is in my view the very reason for his fifty years of success... I have watched, very carefully and with great respect and admiration, the theory and practice of selling satisfaction and serving millions of people on a daily basis, successfully. It is what Disney does best."

Figure 8: Full original passage for Case 3.

Case 3: RACE C (racec_3586). Questions.

Q1	The first day they come to Disney parks, all new employees __ . **A. begin by receiving on-the-job training** B. must learn several jobs C. begin as ticket takers D. have already attended Disney University
Q2	The main objective of the Disney employees is to __ . A. learn all parts of the business **B. see that their guests enjoy themselves** C. be able to answer all kinds of questions D. keep their important guests happy
Q3	Each year, managers wear special service clothes and work in the park to __ . A. set a good example for employees B. remind themselves of their beginnings at Disney **C. gain a better view of the company's objectives** D. replace employees on holiday
Q4	Which is the following is NOT true according to the passage? **A. Tourist learn the history of Disney in its entertainment parks.** B. Disney attracts people almost from all over the world. C. Parades are regularly held in Disney's entertainment parks. D. Disney's managers are able to do almost all kinds of work in the Disney parks.
Q5	This passage is mainly about __ . A. how Disney employees are trained B. the history and traditions of the Disney enterprises C. why Disney enterprises make a lot of money **D. the importance Disney places on serving people well**

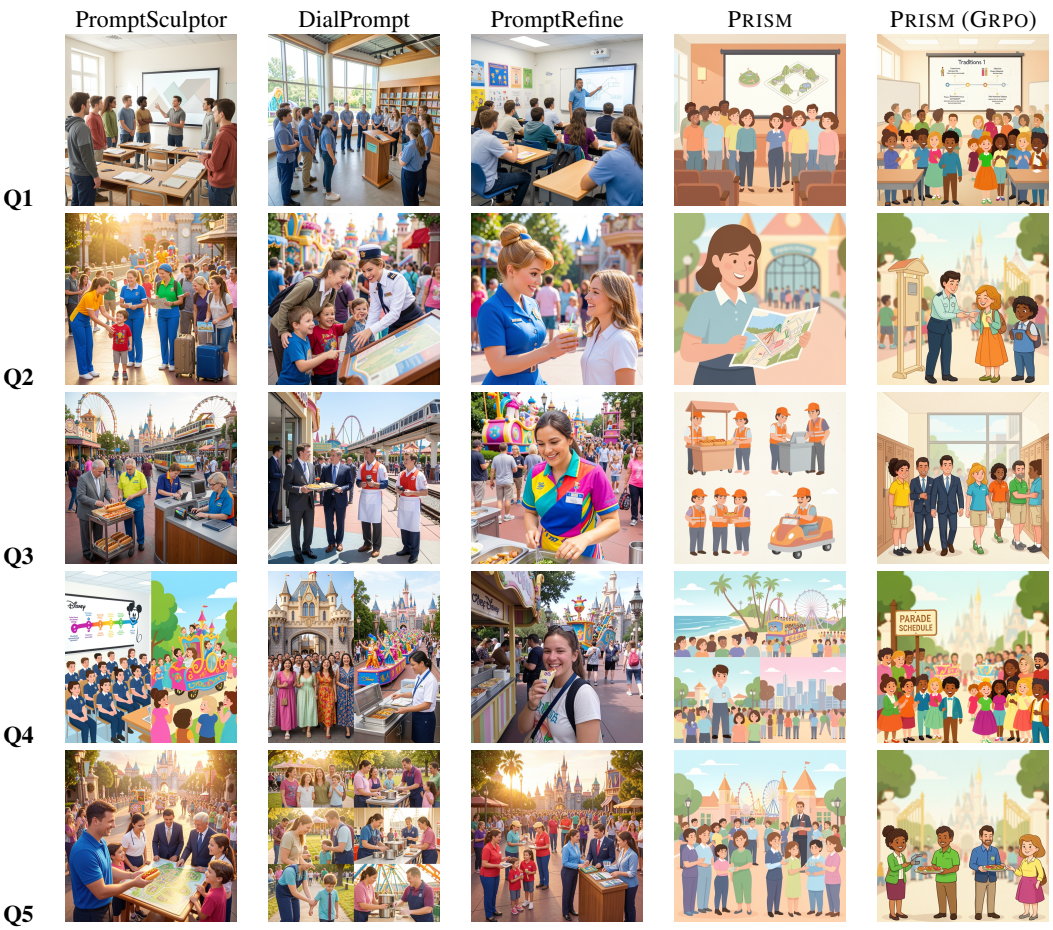

Figure 9: Complete RACE C passage comparison across the five Human Evaluation methods.

Case 4: CosmosQA (cosmos_30EV7DWJTX9DM0K8DPECBY0W0NMY6X). Original passage.

I have been no poo off and on for 3 years now . My hair type is baby fine , oily at scalp , dry at ends , and somewhat wavy and frizzy . Recently I decided to try CO again and finally remembered that any time conditioner meets my scalp I get dandruff .

Figure 10: Full original passage for Case 4.

Case 4: CosmosQA (cosmos_30EV7DWJTX9DM0K8DPECBY0W0NMY6X). Questions.

Q1	Why did she go no poo ? A. She wants to get her hairy really frizzy . B. Her thick hair is baby fine . **C. Conditioner gives her dandruff .** D. None of the above choices .
Q2	What is no poo and why is she doing it ? A. Because her hair is so think she does n't want to ruin it . B. The dryness she get at her scalp makes it hard to use no poo . C. She likes the idea of not washing her hair . **D. no poo stands for no shampoo .**

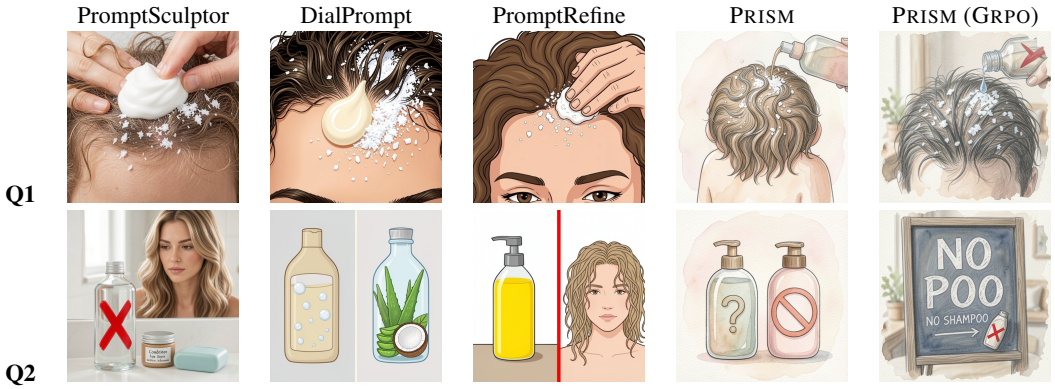

Figure 11: Complete CosmosQA passage comparison across the five Human Evaluation methods.

Case 5: RACE High (high_2938). Original passage.

The slogan "a dog is for life, not just for Christmas" has been going strong for 27 years now. Dogs and cats have been popular Christmas gifts in the gifts in the UK and Europe for many years. But some people still haven't quite got the message. The slogan still holds good. Some 105,000 dogs were _ by local government last year because they had been abandoned, many of them shortly after the festival season. The main reason given by owners for handing their dogs over to the Dogs Trust is that the dogs need more attention than they can give; behavioral problems and moving home are also reasons given. But Eleanor Silk, the Trust's communications director, says, "It's often four or five months after Christmas, in early summer, when owners find that the dog has got too big and too active, or it hasn't been trained properly and has caused damage around the home." In any case, some organizations that help pets are quick to point out that if you are serious about getting a puppy as a pet, Christmas is not the time to do it. Puppies need to get used to their new surroundings, to be introduced to a routine, and to be looked after like small children. It's hard to meet any of these needs at Christmas. What's more, pulling a puppy out of a Christmas stocking in itself sends the wrong message to children. It encourages children to think of the puppy as a toy, rather than a responsibility. Appreciating what you're taking on with a dog can be a sobering realization. "For life" means thinking about the effect they'll have on your life. It also infers you will need to care for the dog for its whole life.

Figure 12: Full original passage for Case 5.

Case 5: RACE High (high_2938). Questions.

Q1	What would be the best title for the passage? A. Dogs and Cats–the Most Popular Christmas Gifts. **B. A pet is for Life, Not just for Christmas.** C. How to Take Care of Your Pets over Christmas. D. Let's Help Those Abandoned Pets
Q2	According to the passage, which of the following is NOT a reason for pets to be abandoned? **A. Pets spread some diseases.** B. Pets damage some things in the house. C. The owners can't take good care of the pets. D. The owners have to move to other places.
Q3	Why is the author against the idea of putting a dog in the Christmas stocking as a Christmas gift? A. Because these dogs can still cause damage. **B. Because children will not take dogs seriously.** C. Because children don't like to have dogs as Christmas gifts. D. Because a Christmas stocking is not big enough to put a dog in.



Figure 13: Complete RACE High passage comparison across the five Human Evaluation methods.