BEYOND STARRY NIGHT: SHORTCUT-AWARE CONTROL-STATE PLANNING FOR ARTIST-GROUNDED TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

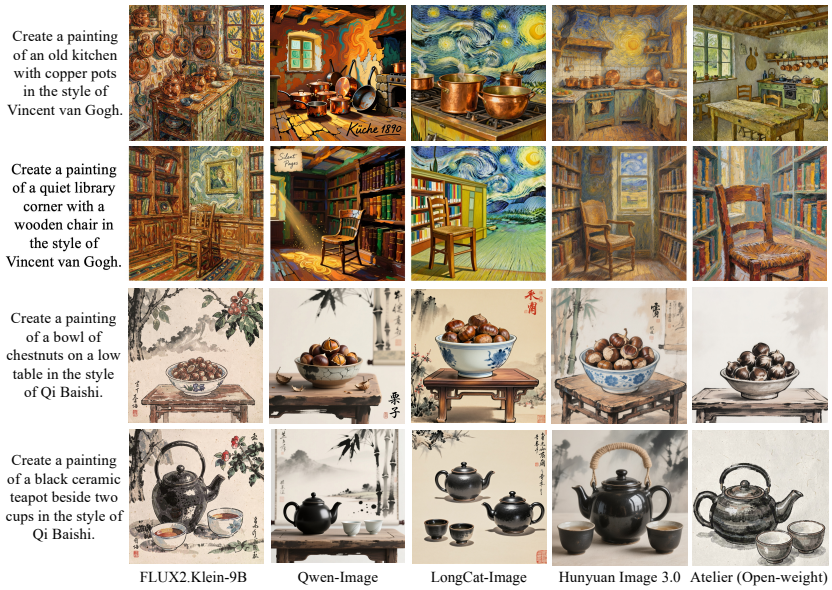

Figure 1: Artist-name conditioning can insert canonical cues unsupported by the requested scene. Direct generators and prompt-expansion agents add such motifs, whereas Atelier separates preserved content, transformed attributes, and cues to avoid.

ABSTRACT

Artist-grounded image generation requires more than appending an artist name to a prompt. Image models often respond to artist names through canonical shortcuts, such as recurring motifs, generic palettes, or overrepresented period signatures, rather than preserving the user's intended scene. We introduce Shortcut-Aware Artistic Control Planning (SACP), a formulation that translates underspecified artistic intent into an explicit control state separating scene anchors, preserve/transform decisions, style-regime hypotheses, role-bound artist evidence, and shortcut-avoidance constraints, and Atelier, its closed-loop runtime. Atelier grounds this state using artist-level knowledge and local patch references, compiles backend-aware generation plans, and iteratively refines candidates through global and local authenticity feedback. We further introduce ArtIntentBench, a benchmark covering Van Gogh and Qi Baishi across artwork re-rendering, period/style-controlled generation, historically unseen subjects, shortcut auditing, and human preference evaluation. Across open-weight and closed-source generators, Atelier improves artist-level style fidelity, preserves source structure more faithfully, and substantially reduces shortcut substitution compared with prompt-engineered, retrieval-augmented, and general-purpose agent baselines. These results suggest that artist-grounded generation is bottlenecked not only by image synthesis, but by the upstream inference of explicit, evidence-grounded artistic controls.

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.

Table 1: Frequent unrequested shortcuts across FLUX, Qwen, LongCat, and Hunyuan on 1,000 neutral prompts per artist. Frequencies are pooled across models.

Van Gogh		*Qi Baishi*	
Unrequested canonical pattern	Freq.	Unrequested canonical pattern	Freq.
Starry Night-like swirling sky	∼55%	Generic ink backgrounds (mountains, pines, mist)	∼45%
Saturated blue-yellow palette across scenes	∼52%	Unrequested calligraphic inscription	∼23%
Uniform, period-independent heavy impasto	∼26%	Generic Chinese decorative style replacing Qi Baishi's ink language	∼12%
Cypresses inserted into scene	∼17%		

1 INTRODUCTION

Modern text-to-image models (Rombach et al., 2022; Saharia et al., 2022; Esser et al., 2024) can produce visually plausible images from open-ended prompts, but artist-grounded generation remains difficult when the prompt specifies only a scene and an artist. A request such as "a quiet subway platform in the style of Van Gogh" does not specify which objects must remain stable, which attributes may be transformed, which period of the artist's work should guide the output, or which motifs are inappropriate for the scene. As shown in Figure 1 a model may produce generic painterly texture, overuse a small set of canonical visual cues, or replace the requested subject with motifs associated with the artist.

An important failure mode in artist-grounded generation is *canonical shortcutting*: a model relies on a small set of high-frequency visual associations rather than modeling the artist's visual language as a context-sensitive combination of composition, palette, brushwork, and motif selection. The term follows the broader notion of shortcut learning, in which models exploit spuriously correlated cues in place of the intended structure (Geirhos et al., 2020; Goyal et al., 2025). Not all canonical associations are shortcuts—a swirling sky is appropriate for a Van Gogh Saint-Rémy nocturne, and calligraphic inscription is appropriate when requested. A shortcut is an *unrequested* insertion: a motif or treatment applied outside its semantic, historical, or compositional context, or one that overrides the requested subject. These failures can appear plausible overall while replacing artist-specific visual structure with generic canonical cues.

To test whether shortcutting is systematic rather than anecdotal, we audit four strong baseline generators (FLUX (Black Forest Labs, 2026), Qwen (Wu et al., 2025), LongCat (Ma et al., 2025), Hunyuan (Cao et al., 2025)) on 1,000 neutral prompts per artist, avoiding explicit period or motif instructions. Each output is independently reviewed by three volunteer annotators, using the same majority-vote rule as the shortcut audit in Section 4.2: a pattern is counted only when selected by at least two annotators. Table 1 shows frequent unrequested shortcuts for both artists. These failures are visually plausible but artistically misgrounded because they replace the requested scene with high-frequency artist-name associations.

We frame this problem as *language-to-control translation* for artistic image generation. The goal is not merely to imitate an artist's surface style, but to infer an explicit control state from an incomplete natural-language request. A faithful output must preserve the intended scene while making artist-specific decisions about period, composition, motif selection, palette, brushwork, material treatment, and local texture. These decisions are prior to image synthesis: they specify what should remain stable, what may be transformed, which artist-specific evidence should guide the transformation, and which contextually inappropriate shortcut substitutions should be avoided.

We present **Shortcut-Aware Artistic Control Planning (SACP)**, a formulation of artist-grounded generation as control-state inference, and `Atelier`, its closed-loop runtime. SACP first predicts an artistic control state that decomposes the request into scene anchors, preserve/transform policies, style-regime hypotheses, retrieval targets, and anti-shortcut constraints. `Atelier` then retrieves artist-level knowledge and local patch evidence, binds patches to scene roles, compiles backend-aware generation plans, and iteratively refines candidates using global and local authenticity critics. Unlike generic retrieval-augmented prompting, it does not use references as undifferentiated style examples; it binds evidence to the specific scene roles and transformation decisions encoded in the control state.

Evaluating this problem also requires more than generic image quality or text-image alignment. Artist-grounded failures are often local: an image may be plausible overall while misrepresenting

brushwork, motif treatment, material texture, or period-specific surface language. We therefore evaluate both the intermediate control state and the final image, and we separate global image-level assessment from local authenticity diagnosis. To treat shortcutting as an empirical rather than a post hoc concern, we audit baseline failure patterns, verify each pattern by independent majority-vote annotation, and freeze the taxonomy before evaluating `Atelier` against it.

Contributions.

1. We identify and quantify canonical shortcutting, a systematic failure in which artist-name conditioning inserts high-frequency motifs or treatments that are unsupported by the requested scene. We construct frozen shortcut taxonomies for Van Gogh and Qi Baishi, validated by majority-vote human annotation, and use them to measure shortcut substitution across open-weight and closed-source generators.
2. We introduce SACP, which translates an underspecified artistic request into an explicit control state, and `Atelier`, its closed-loop runtime. The state separates scene anchors, preserve-versus-transform decisions, style-regime intent, role-bound artist evidence, and anti-shortcut constraints, and is iteratively revised using global and local evaluation feedback.
3. We evaluate Atelier on artist-specific control tasks covering artwork reconstruction, period-controlled generation, out-of-oeuvre modern subjects, and shortcut avoidance. Under matched backend and generation budgets, structured control improves artist-specific style fidelity and reduces shortcut substitution compared with prompt expansion, retrieval-augmented prompting, and generic iterative agents.

2 Related Work

Style transfer, personalization, and editing. Style-transfer methods render content using exemplar feature statistics or translate collections of an artist's works (Gatys et al., 2016; Huang & Belongie, 2017; Zhu et al., 2017). ChipGAN (He et al., 2018) targets Chinese ink painting through voids, brush strokes, and ink-wash diffusion. Personalization adapts generative models to specified subjects, concepts, or styles (Gal et al., 2023; Ruiz et al., 2023; Kumari et al., 2023; Sohn et al., 2023). Instruction-based editing uses natural-language, structural, and region-level controls (Hertz et al., 2023; Zhang et al., 2023; Ye et al., 2023; Mou et al., 2024; Brooks et al., 2023; Li et al., 2024; Yu et al., 2024). These methods assume that the style target, identity, or edit is already specified. Artist-grounded generation must instead infer, from a vague request, the regime within the artist's oeuvre, the content to preserve, and the substitutions to reject. A style loss or identity token does not express these decisions.

Agentic and tool-using generation systems. Language agents use decomposition, tool use, memory, and iterative feedback for complex tasks (Yao et al., 2023; Shen et al., 2023; Shinn et al., 2023; Qin et al., 2023). Perceive–plan–execute–evaluate agents extend this pattern to image generation (Yao et al., 2026), while hierarchical planner–executor agents use structured interaction memory for image editing (Ye et al., 2026). These frameworks do not ground the loop in artist-level knowledge. `Atelier` adds a control state for preserve/transform intent, scene-role-bound retrieval, and patch-level authenticity feedback.

Evaluation of image generation and artistic authenticity. Existing metrics provide global signals for quality, alignment, or human preference (Heusel et al., 2017; Radford et al., 2021; Xu et al., 2023), but do not directly diagnose local failures in brushwork, motif treatment, or material texture. Art-specific judgment remains difficult even for strong multimodal models (Bin et al., 2024; Hayashi et al., 2024; Strafforello et al., 2025), while patch-level VLM concepts partially align with art-historian judgment (Limpijankit et al., 2026). We therefore separate global image-level evaluation from localized authenticity diagnosis.

Shortcuts in text-to-image generation. Goyal et al. (Goyal et al., 2025) address adapter-training shortcuts in personalization. Their shortcut pathways prevent adapters from encoding confounding

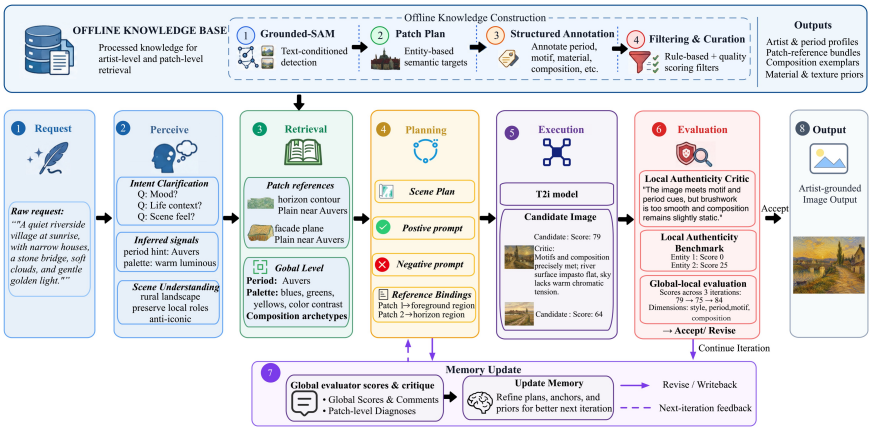

Figure 2: Overview of Atelier. An underspecified request is mapped to an explicit control state, grounded with global artist knowledge and role-bound local patches, and compiled into a backend-aware plan. A global critic and patch-level AUTHCRITIC write feedback to memory for iterative refinement and final selection.

attributes such as pose, expression, and lighting alongside the target identity. We study a complementary inference-time failure in general-purpose generators: artist names trigger canonical-motif substitution. Atelier addresses it with anti-shortcut constraints in the control state and iterative authenticity feedback rather than additional adapter training.

3 METHOD

We now specify **Shortcut-Aware Artistic Control Planning (SACP)** and its runtime. The method is *shortcut-aware* because anti-shortcut constraints and preserve/transform decisions are explicit fields of the control state, read directly by retrieval, planning, and evaluation rather than embedded in the natural-language prompt.

We instantiate SACP in Atelier (Figure 2), a closed-loop runtime that extends the perceive–plan–execute–evaluate paradigm (Yao et al., 2026) with two artist-specific representations: the structured control state, and a multimodal artist knowledge base whose evidence is bound to scene roles. The runtime can be summarized as

$$x \xrightarrow{\text{PERCEIVE}} z^{(0)} \xrightarrow{\text{DERIVE}} z \xrightarrow{\text{RETRIEVE}} k \xrightarrow{\text{PLAN}} (\tilde{z}, c, e),$$
$$e \xrightarrow{\text{EXECUTE}} \mathcal{Y} \xrightarrow{\text{EVALUATE}} \rho \xrightarrow{\text{REFLECT}} m.$$

Given an underspecified request x, Atelier first infers what should be preserved, what may be artistically transformed, which period or style-regime cues are relevant, and which artist-name shortcuts should be avoided. It then retrieves global artist knowledge and local patch references, binds them to scene roles, generates candidate images through one or more backends, and revises subsequent rounds using feedback from a global critic and AUTHCRITIC, a trained patch-level authenticity critic, stored in memory.

3.1 PERCEIVE: ARTISTIC CONTROL STATE

The perceiver, implemented as a prompted LLM, converts the raw request into an explicit artistic control state rather than a longer prompt. We write this state as

$$z = (s, q, h, r, b),$$

where s is the scene reading, q contains preserve-versus-transform decisions, h stores historical or style-regime intent, r records retrieval targets and anti-shortcut constraints, and b contains backend-facing execution preferences.

The scene reading identifies objects, regions, relations, and mood cues in the request. The preserve/transform component marks which elements should remain semantically stable and which

attributes may be translated into the target artist's visual language. The historical/style component records explicit or implicit period cues, such as Van Gogh's Paris, Arles, Saint-Remy, or Auvers periods, or Qi Baishi style regimes defined by motif, composition, ink economy, inscription, and seal usage. The anti-shortcut component specifies forbidden substitutions, for example replacing a requested modern object with a canonical artist motif. This state is the main interface between language understanding, artist evidence, generation, and evaluation. Appendix A.5.1 exhibits the full state z recorded for a held-out request. Appendix A.1.2 gives the field-level map from (s, q, h, r, b) to the concrete record paths. The perceiver resolves the request once into an episode-level interpretation context. Bootstrap inputs remain fixed, while the operational state used for planning and generation may change within a round. The planner may revise the working period hypothesis in h using retrieved evidence. A user-specified period is enforced deterministically by binding each compiled generation call to that period's resources, independent of the working hypothesis.

3.2 RETRIEVE: ARTIST KNOWLEDGE AND LOCAL EVIDENCE

`Atelier` grounds z with an artist-specific knowledge base $\mathcal{B}_a$, organized in three layers of evidence: a global layer of artist-level style priors, a supplementary layer of curated summaries that reinforce those priors, and a local layer of annotated patches from source artworks.

At runtime, the retrieval module produces an evidence bundle

$$k = K(x, z; \mathcal{B}_a).$$

The bundle contains both global references and local patches. Global references provide period, motif, composition, and palette priors. Local patches provide region-level evidence for scene roles such as foreground subject, background field, scene wrapper, material surface, or texture target. Unlike ordinary retrieval-augmented prompting, these references are not appended as generic style examples; they are explicitly bound to the scene roles and constraints in z.

The retriever conditions this bundle on x and the current scene reading. The planner uses it for global style decisions and local rendering.

3.3 PLAN AND EXECUTE: BACKEND-AWARE GENERATION

The planner, implemented as a prompted LLM, converts the control state and retrieved evidence into a working state view and a backend-aware generation plan. At iteration t, it computes

$$(\widetilde{z}^{(t)}, c^{(t)}, e^{(t)}) = P(x, z^{(t)}, k^{(t)}, m^{(t-1)}),$$

where $\widetilde{z}^{(t)}$ is the planner-emitted working view. $c^{(t)}$ contains scene anchors, period or style-regime targets, reference bindings, and negative shortcut constraints. $e^{(t)}$ specifies the backend execution policy.

The executor maps this plan to concrete generation calls. Backends with reference, adapter, or LoRA-style control receive local patch references, period-specific resources, patch-resource bindings, and scale parameters. General purpose backends receive a compact textual realization of the same control state, including preservation constraints, local style targets, and forbidden substitutions. Early rounds may explore multiple backends; later rounds can lock onto the strongest backend once the evaluator identifies a reliable direction.

Algorithm 1 (Appendix A.3) gives the complete runtime loop. At iteration t, DERIVE rebuilds the base state from $z^{(0)}$ and $m^{(t-1)}$, including the constraints in r and preferences in b. The planner may return a working state view with $c^{(t)}$ and $e^{(t)}$. Before execution, MERGE applies the artist-conditioned field policy. Revision updates the plan, evidence bindings, execution policy, and planner-revisable state fields, while leaving the episode request fixed.

Shortcut avoidance is enforced at four points in the runtime loop rather than left to prompt-level suggestion. (i) The plan $c^{(t)}$ encodes anti-shortcut constraints as explicit fields and describes style through decomposed visual attributes from $k^{(t)}$, rather than relying only on an artist name. (ii) Backends with a native negative-prompt channel receive the shortcut list in that channel; for backends without one, the executor appends the constraints to the text prompt as an imperative `Avoid:` block. (iii) The global critic receives the artist-specific shortcut guard. On unsupported-subject

routes, matched canonical aliases trigger a deterministic penalty to intent, motif, composition, and style scores, increase the artifact penalty, and downgrade an accept recommendation to revise. (iv) The resulting violation tag and repair actions enter memory and condition the next round's constraints.

3.4 EVALUATE: GLOBAL CRITIC AND AUTHCRITIC

Each candidate image is evaluated in a two-stage pipeline. AUTHCRITIC runs first on each candidate and produces per-patch authenticity scores; the low-scoring patches are surfaced to the global critic as focus signals. The global critic then judges the candidate as a whole for semantic fidelity, composition, period or style-regime match, palette, artist alignment, and visible artifacts, with its attention already focused on the regions AUTHCRITIC has flagged.

Global critic. We prompt Kimi K2.6 (Moonshot AI, 2026) as the global critic, without task-specific fine-tuning. The critic is conditioned on the candidate image and a fixed-schema projection $\phi(x, z, k)$ of the request, the current control state, and the retrieved evidence, together with the focus signals from AUTHCRITIC. Appendix A.4.7 specifies the projection in full. It returns a structured judgment with a score, a natural-language critique, and an accept/revise recommendation.

Because the visual evidence that constitutes artistic authenticity is artist dependent, the global critic uses a shared output contract with an artist-specific rubric. The Van Gogh instance evaluates period-conditioned oil-paint handling, directional brushwork, and canonical-shortcut avoidance; the Qi Baishi instance instead evaluates calligraphic stroke economy, ink-density variation, paper reserve, and avoidance of unrequested generic Chinese decorative conventions. The full rubrics appear with the critic prompts in Appendix C.1.5. The two instances share the same structured output fields and numerical aggregation (Appendix A.4.2), so the runtime policy is unchanged while the artistic interpretation of the fields is specialized to the target artist.

AUTHCRITIC. AUTHCRITIC is a patch-level authenticity critic obtained by supervised fine-tuning of Gemma-4 (Google DeepMind, 2026) (E4B-it) with a LoRA adapter. We use supervised fine-tuning rather than preference-based training because each patch carries a deterministic source label from its known origin image, and the runtime loop consumes calibrated per-patch scores rather than pairwise preferences. The training corpus contains $15{,}000$ patches evenly split across three source classes: real Van Gogh patches, synthetic Van Gogh-style patches, and other-painter patches. These three sources let AUTHCRITIC distinguish real Van Gogh from both the failure modes of current generators and from stylistically adjacent artistic traditions.

On a held-out test set of $1{,}485$ patches, AUTHCRITIC reaches 91.65% binary accuracy (real Van Gogh vs. others). At inference, each patch yields a source-type score and a proximity-to-real estimate. Appendix A.4.6 gives the candidate-level aggregation.

The Qi Baishi instantiation uses the same recipe; its corpus, split, and accuracies are given in Appendix A.2.

3.5 REFLECT: MEMORY AND REVISION

After each round, `Atelier` stores the selected candidate, backend choice, control plan, global critique, local patch evidence, failure tags, and repair targets in memory $m^{(t)}$. The record prevents the planner from repeating failed choices, such as reusing a shortcut motif or an unsuitable backend. It also preserves successful decisions, such as a strong period reference, stable composition, or effective patch binding.

The next round uses memory to revise the control state, update reference bindings, adjust backend selection, and focus generation on unresolved local failures. The loop terminates under the success, plateau, budget, or error conditions specified in Appendix A.4.3. FINALSELECT then returns the final image from the full trajectory (Appendix A.4.4).

4 EXPERIMENTS

4.1 ARTINTENTBENCH AND EXPERIMENTAL SETUP

ArtIntentBench favors artist-specific depth over broad coverage. Reliable evaluation requires period taxonomies, motif priors, local patch evidence, and human assessment of brushwork, material treatment, and shortcut substitution. We select Van Gogh and Qi Baishi because their media, traditions, and control challenges differ substantially.

The benchmark includes 100 Van Gogh paintings converted to structured captions for text-only re-rendering, 228 prompts that hold the scene fixed while varying the target across four periods, and 80 Qi Baishi paintings for re-rendering. It also includes 30 historically unseen subjects to test transfer beyond canonical motifs. All evaluations are non-interactive and infer intent from a single request.

Metrics. We report IntroStyle (Kumar et al., 2025) W_2 distance (↓) for artist-level style proximity and CLIP similarity (↑) for text–image alignment. DreamSim (Fu et al., 2023) (↓), DINO-cosine (Caron et al., 2021) (↑), and LPIPS (Zhang et al., 2018) (↓) measure fidelity to the source painting. For period-controlled generation, the W_2 reference set is restricted to the target period. This restriction penalizes canonical shortcutting: inserting a *Starry Night* sky for a Paris-period target increases the distance. Because W_2 alone does not measure content preservation or shortcut avoidance, we also use the human-annotated shortcut rate in §4.2. Human-preference results are reported in Appendix B.9. Full task protocols and metric rationale appear in Appendices B.1 and B.2.

Configurations and baselines. The open-weight variant uses DeepSeek V4 Pro (DeepSeek-AI, 2026) for perception and planning, followed by generation with four open image backends. The closed-source variant uses GPT-5.4 (OpenAI, 2026b) for both roles and two commercial backends (Appendix B.2). Both variants share the scene-reading schema, artist-knowledge retrieval pipeline, patch-reference bank, iteration budget, evaluator stack, and selection policy. This isolates structured control from model-family choice.

We compare `Atelier` with three baseline families. Direct-backend baselines use the same six image models with standardized prompt-engineered instructions rather than minimal "in the style of" prompts. Agentic prompt-expansion baselines, including Claude-based and GenArtist-style (Wang et al., 2024) systems, rewrite the request before generation but lack the explicit control state, artist-knowledge retrieval, scene-role patch bindings, evaluator feedback, and iterative memory used by `Atelier`. Retrieval-augmented prompting lets the planner select its own reference works. It uses the open-weight candidate pool and is reported with the prompt-construction comparison in Table 10.

4.2 SHORTCUT AUDIT

Setup. Using the frozen taxonomy introduced in Section 1 (Table 1), we audit unrequested canonical shortcuts on Van Gogh and Qi Baishi prompts. Each method is evaluated on 358 Van Gogh and 80 Qi Baishi prompts from §4.1. The taxonomy was frozen on separate neutral prompts before this evaluation. None of the audit prompts requests a taxonomy shortcut; an explicit period label in the period-controlled subset is a task condition, not a shortcut pattern. Three annotators apply the majority-vote protocol ($\geq 2/3$) defined in Section 1. SSR is the proportion of images containing at least one shortcut pattern from the corresponding artist's taxonomy.

Van Gogh. Baselines exhibit high shortcut rates in both configurations, with GenArtist reaching the lowest baseline SSR. The highest closed-source SSR comes from an agent wrapper, Claude (NBP); such wrappers inherit the underlying generator's canonical bias rather than mitigating it. `Atelier` reduces SSR to 51.96% (closed-source) and 31.56% (open-weight), improving over the strongest baseline in each category by 7.3 and 15.7 percentage points respectively. Heavy impasto remains the dominant failure mode across baselines, while cypress insertion peaks under closed-source systems (GPT-Image-2). `Atelier` attains the best rate in every column except two: for closed-source cypresses, GenArtist (NBP) is marginally lower, and for open-weight blue-yellow palette, Hunyuan is marginally lower.

The larger residual SSR under the closed-source configuration reflects a limitation of the underlying generation model rather than of the control state itself. `Atelier` produces the same intermediate control state regardless of downstream backend, so the gap between the open-weight and closed-source SSR indicates that closed-source models (GPT-Image-2, Nano Banana Pro) carry stronger pre-trained canonical priors that partially override the explicit avoidance constraints encoded in the control state. These backends also expose no dedicated negative-prompt channel; shortcut constraints are therefore carried as imperative statements inside the positive prompt, consistent with the SSR gap observed here. Structured control is therefore necessary but not sufficient: the downstream generator must also remain receptive to fine-grained control signals.

Qi Baishi. Baselines exhibit high shortcut rates in both configurations. The highest closed-source SSR again comes from Claude (NBP), matching the Van Gogh pattern. `Atelier` reduces SSR to 43.75% (closed-source) and 28.75% (open-weight), improving over the strongest baseline in each category by 11.25 percentage points. Generic ink-painting backgrounds remain a common failure mode across closed-source baselines: Claude (NBP) and Nano Banana Pro are most affected. In contrast, GPT-Image-2 and its wrappers exhibit high rates of unrequested calligraphic inscription. `Atelier` attains the best rate on every open-weight pattern and on two of three closed-source patterns.

Table 2: Shortcut audit on Van Gogh (358 images per method) and Qi Baishi (80 images per method) prompts. SSR is the fraction with at least one unrequested shortcut; pattern rates may overlap. Bold and underlined mark the best and second-best result within each configuration; lower is better.

Method	**Van Gogh**					**Qi Baishi**			
	SSR ↓	**Starry Night sky ↓**	**Blue-yellow palette ↓**	**Heavy impasto ↓**	**Cypresses ↓**	**SSR ↓**	**Generic ink BG ↓**	**Calligraphic insc. ↓**	**Decorative subst. ↓**
Closed-source configuration									
Nano Banana Pro (NBP)	69.55%	19.83%	15.08%	31.84%	19.83%	68.75%	51.25%	25.00%	10.00%
GPT-Image-2	76.82%	16.20%	13.97%	30.45%	33.80%	71.25%	25.00%	53.75%	16.25%
GenArtist (NBP)	59.22%	17.32%	14.25%	30.73%	**11.73%**	55.00%	35.00%	22.50%	16.25%
GenArtist (GPT-Image-2)	70.95%	14.25%	16.20%	31.84%	26.26%	63.75%	**21.25%**	45.00%	22.50%
Claude (GPT-Image-2)	75.70%	15.08%	24.02%	29.05%	25.98%	68.75%	35.00%	41.25%	17.50%
Claude (NBP)	77.37%	22.63%	18.72%	30.73%	23.18%	78.75%	56.25%	22.50%	16.25%
Atelier (Closed-source)	**51.96%**	**10.61%**	**10.34%**	**27.93%**	12.85%	**43.75%**	25.00%	**21.25%**	**15.00%**
Open-weight configuration									
FLUX	56.98%	27.09%	8.38%	24.58%	9.22%	53.75%	22.50%	20.00%	26.25%
Qwen	51.68%	13.97%	12.57%	31.84%	7.54%	48.75%	20.00%	32.50%	13.75%
LongCat	54.19%	19.55%	8.38%	30.73%	8.10%	48.75%	27.50%	25.00%	13.75%
Hunyuan	48.04%	17.88%	**4.75%**	30.17%	8.66%	40.00%	22.50%	25.00%	13.75%
GenArtist	47.21%	14.53%	8.66%	31.84%	7.82%	40.00%	18.75%	28.75%	15.00%
Atelier (Open-weight)	**31.56%**	**8.10%**	6.98%	**17.60%**	**4.19%**	**28.75%**	**13.75%**	**11.25%**	**11.25%**

4.3 Artist-Grounded Fidelity

We next test whether shortcut avoidance preserves artist-level fidelity. On artwork re-rendering, `Atelier` achieves the lowest IntroStyle W_2 in both configurations: 64.22 (closed-source) and 73.52 (open-weight), compared with 66.55 for GPT-Image-2. It also leads both configurations on DreamSim, DINO-cosine, and LPIPS (Table 3); all methods use the same structured VLM-derived prompt. Qualitative comparisons are provided in Appendix B.10.

Period-controlled generation. For period control, `Atelier` yields the lowest style distance in every period under both configurations (Table 3). Figure 3 shows period-cue reuse under the open-weight configuration: FLUX repeats swirling strokes and the blue-yellow palette across target periods, whereas `Atelier` changes palette, brushwork, and surface treatment while preserving scene structure and depth. Together, period-restricted references and source-fidelity metrics show that lower SSR does not merely reflect weaker artistic signals.

4.4 Control-State Quality and Component Ablation

We directly audit `Atelier`'s recorded control states (Appendix B.3). Period labels reach every compiled call, and preserve/transform fields are well formed in 73–100% of cases. Novel-identity and anti-shortcut checks pass in at least 93.3% of historically unseen cases, supporting the shortcut reductions in §4.2. The weakest step is inferring an unstated period from a caption, with 38–40% top-1 recovery. Two independent LLM judges agree at $r = 0.43$–0.66 on the same states (Appendix B.4).

Table 3: ArtIntentBench results. Artwork re-rendering reports IntroStyle W_2, CLIP, and source fidelity via DreamSim (DS), DINO-cosine (DINO), and LPIPS (LP); period control reports W_2 and CLIP. Bold and underlined mark the best and second-best result within each configuration.

Method	Artwork Re-rendering					Paris		Arles		Saint-Rémy		Auvers	
	$W_2\downarrow$	CLIP↑	DS↓	DINO↑	LP↓	$W_2\downarrow$	CLIP↑	$W_2\downarrow$	CLIP↑	$W_2\downarrow$	CLIP↑	$W_2\downarrow$	CLIP↑
Closed-source configuration													
GPT-Image-2	66.55	0.2861	0.3040	0.6847	0.5861	84.20	0.3428	75.77	0.3539	69.27	0.3504	71.47	0.3590
Nano Banana Pro (NBP)	70.61	0.2814	0.3128	0.6423	0.5882	93.38	0.3485	92.56	0.3410	78.41	**0.3610**	81.54	0.3571
GenArtist (GPT-Image-2)	67.77	0.2852	0.3293	0.6416	0.6052	85.55	0.3447	75.95	0.3539	69.59	0.3482	71.79	0.3532
GenArtist (NBP)	73.39	0.2835	0.3311	0.6535	0.5915	89.24	0.3391	87.06	0.3431	77.35	0.3535	79.28	**0.3616**
Claude (GPT-Image-2)	66.78	**0.2914**	0.3126	0.6619	0.5879	84.57	0.3445	75.01	**0.3544**	69.28	0.3534	70.05	0.3580
Claude (NBP)	72.48	0.2847	0.3205	0.6671	0.5855	92.90	**0.3508**	90.40	0.3352	77.45	0.3599	81.63	0.3566
Atelier (Closed-source)	**64.22**	0.2816	**0.2933**	**0.7075**	**0.5704**	**82.40**	0.3399	**73.42**	0.3494	**66.99**	0.3442	**69.83**	0.3493
Open-weight configuration													
FLUX	77.71	0.2546	0.3751	0.5741	0.6245	86.78	**0.3311**	84.30	0.3325	77.96	**0.3443**	82.18	**0.3376**
Qwen	76.27	0.2655	0.4068	0.5810	0.6191	93.50	0.3084	80.55	0.3209	74.09	0.3202	76.83	0.3170
LongCat	75.56	0.2863	0.3660	0.6098	0.6090	84.56	0.2687	75.32	0.2805	71.59	0.2726	72.83	0.2832
Hunyuan	74.82	**0.3008**	0.3579	0.6194	0.6086	84.36	0.3116	76.81	0.3293	67.74	0.3227	70.40	0.3308
GenArtist	81.99	0.2936	0.4915	0.4665	0.6368	84.67	0.3106	74.65	0.3211	70.38	0.3160	71.68	0.3238
Atelier (Open-weight)	**73.52**	0.2722	**0.3440**	**0.6398**	**0.5916**	**83.55**	0.3210	**73.70**	**0.3338**	**64.55**	0.3333	**69.22**	0.3320

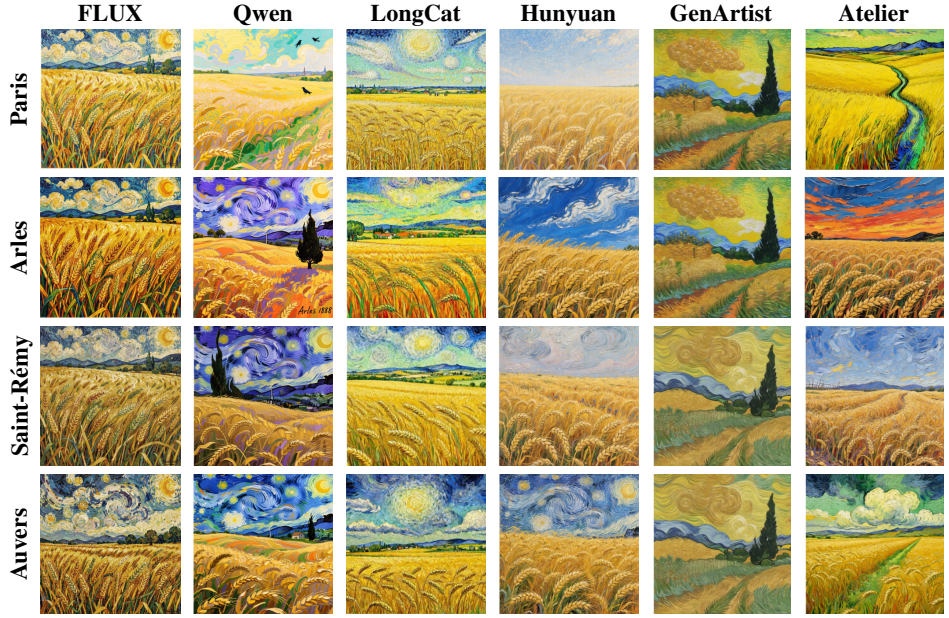

Figure 3: Period-controlled generation on a fixed scene in the open-weight configuration. Each column is a method; each row fixes the target period.

On 40 period-controlled cases (10 subjects $\times$ 4 periods), we ablate each mechanism while fixing the backend pool and iteration budget. Removing the structured intent representation raises W_2 from 81.79 to 85.96, the largest degradation, while all other removals also hurt performance (Table 9). The gain therefore cannot be attributed to iteration alone. Without preserve/transform decisions and anti-shortcut fields, the loop revises toward canonical output. The prompt-construction comparison in Table 10 further rules out better prompt wording as the sole explanation.

5 Discussion and Limitations

Generic metrics miss period fidelity, motif appropriateness, brushwork, material treatment, and shortcut substitution. We therefore combine style distances with blinded ranking and shortcut annotation. The two-artist benchmark limits broader claims, although Qi Baishi results show that artist-agnostic SACP can be retargeted by replacing artist-specific data rather than its control schema. Future work should expand curated evidence, validate taxonomies with broader expert panels, and let users inspect and revise control states.

REPRODUCIBILITY STATEMENT

The appendix specifies the runtime in full: perceiver, planner, and critic prompts (Appendix C.1), runtime constants, selection rules, and stop conditions (Appendix A.4), the control-state and memory schema (Appendix A.1), and the knowledge-base construction pipeline with AUTHCRITIC training details (Appendix A.2). Experimental configurations are specified in Appendix B.2. Benchmark construction and task counts are given in Section 4.1, with design and metric rationale in Appendix B.1, and the annotation protocol in Section 4.2. We will release the code, artist knowledge-base metadata, benchmark prompts, generated outputs, and evaluation scripts. Copyrighted source images will be linked rather than redistributed, as described in the Ethics and Copyright Statement.

ETHICS AND COPYRIGHT STATEMENT

The Van Gogh corpus uses public-domain reproductions from WikiArt and the National Gallery of Art; the Qi Baishi corpus uses public museum collections. We will release code, metadata, benchmark prompts, and evaluation protocols. Copyrighted source images will be linked rather than redistributed, and released records will include the control-state and evaluator identifiers needed to reproduce each artist-conditioned result.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, software implementation, and preparation of research artifacts. The authors reviewed and verified all AI-assisted text, code, analyses, and figures, and take responsibility for the final content and claims of this submission.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. GalleryGPT: Analyzing paintings with large multimodal models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7734–7743, 2024.

Black Forest Labs. Flux.2 [klein]: Towards interactive visual intelligence. Black Forest Labs Blog, January 2026. URL https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence. Accessed May 18, 2026.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

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*, pp. 9650–9660, 2021. URL https://arxiv.org/abs/2104.14294.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 8359–8388, 2024.

DeepSeek-AI. DeepSeek-V4 technical report. Hugging Face technical report, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12606–12633, 2024.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.09344.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Google. Gemini 3 Pro Image Preview (Nano Banana Pro). Google AI for Developers model documentation, 2025. URL https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-image-preview. Accessed May 18, 2026.

Google DeepMind. Gemma 4 model card. Google AI for Developers model documentation, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Accessed May 18, 2026.

Anujraaj Goyal, Guocheng Qian, Huseyin Coskun, Aarush Gupta, Himmy Tam, Daniil Ostashev, Ju Hu, Dhritiman Sagar, Sergey Tulyakov, Kfir Aberman, and Kuan-Chieh Wang. Preventing shortcuts in adapter training via providing the shortcuts. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/9323ea05f1764f94143392e9e4fb88ab-Abstract-Conference.html.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 705–729, 2024.

Bin He, Feng Gao, Daiqian Ma, Boxin Shi, and Ling-Yu Duan. ChipGAN: A generative adversarial network for chinese ink wash painting style transfer. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1172–1180, 2018.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Anand Kumar, Jiteng Mu, and Nuno Vasconcelos. Introstyle: Training-free introspective style attribution using diffusion features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14909–14918, 2025.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Zone: Zero-shot instruction-guided local editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6254–6263, 2024.

Marvin Limpijankit, Milad Alshomary, Yassin Oulad Daoud, Amith Ananthram, Tim Trombley, Elias Stengel-Eskin, Mohit Bansal, Noam M. Elcott, and Kathleen McKeown. Does AI see like art historians? interpreting how vision language models recognize artistic style. *arXiv preprint arXiv:2603.11024*, 2026.

Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. LongCat-Image technical report. *arXiv preprint arXiv:2512.07584*, 2025. URL https://arxiv.org/abs/2512.07584.

Moonshot AI. Kimi AI with K2.6. Moonshot AI product page, 2026. URL https://www.kimi.com/. Accessed May 18, 2026.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, pp. 4296–4304, 2024.

OpenAI. Introducing ChatGPT Images 2.0 (GPT-Image-2). OpenAI Blog, April 2026a. URL https://openai.com/index/introducing-chatgpt-images-2-0/. Accessed May 18, 2026.

OpenAI. Introducing GPT-5.4. OpenAI Blog, March 2026b. URL https://openai.com/index/introducing-gpt-5-4/. Accessed May 18, 2026.

Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Icebench: A unified and comprehensive benchmark for image creating and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16586–16596, 2025.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, et al. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023.

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.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks, 2024. URL https://arxiv.org/abs/2401.14159.

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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38154–38180, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

Ombretta Strafforello, Derya Soydaner, Michiel Willems, Anne-Sofie Maerten, and Stefanie De Winter. Have large vision-language models mastered art history? In *International Conference on Image Analysis and Processing*, pp. 524–544, 2025.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. doi: 10.48550/arXiv.2508.02324. URL https://arxiv.org/abs/2508.02324.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Mingde Yao, Zhiyuan You, King-Man Tam, Menglu Wang, and Tianfan Xue. Photoagent: Agentic photo editing with exploratory visual aesthetic planning. *arXiv preprint arXiv:2602.22809*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Ruijie Ye, Jiayi Zhang, Zhuoxin Liu, Zihao Zhu, Siyuan Yang, Li Li, Tianfu Fu, Franck Dernoncourt, Yue Zhao, Jiacheng Zhu, Ryan Rossi, Wenhao Chai, and Zhengzhong Tu. Agent banana: High-fidelity image editing with agentic thinking and tooling. *arXiv preprint arXiv:2602.09084*, 2026.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://arxiv.org/abs/1801.03924.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

A METHOD DETAILS

This section specifies the runtime components referenced in Section 3: the control-state and memory schema, knowledge-base construction and AUTHCRITIC training, the runtime algorithm, the evaluation and selection rules, and a recorded control-state walkthrough.

A.1 CONTROL-STATE AND MEMORY SCHEMA REFERENCE

This section documents how the abstract state $z = (s, q, h, r, b)$ (§3.1) and the memory m are realised as records in the runtime. The schema is enforced as a prompt-defined contract: the perceiver and planner prompts (Appendix C.1) specify the required keys and require strict JSON, and the runtime normalises the returned objects, substituting the corresponding default value for any field that is absent or malformed.

The abstract component s is called the *scene reading* in the main text. The recorded JSON key `world_model` is an implementation name for this component and should not be interpreted as a predictive dynamics model. The field contains the scene skeleton, local roles, relations, and scene-regime analysis used by DERIVE, PLAN, and the audit.

A.1.1 PLANNER OUTPUT CONTRACT

The planner returns a strict JSON object with the six top-level keys of Table 4; the runtime additionally records a derived `routing_plan`.

Key	Content
`thought_process_summary`	Rationale for the round's decisions.
`meta_strategy_update`	Revised strategy statement carried into memory (Appendix A.5.2).
`knowledge_evidence`	Evidence bundle: the `world_model` (scene-reading record), period policy, motif/patch bindings, retrieved works (Table 5).
`tool_plan`	Required keys `primary_tool_id`, `calls`; one entry per backend call.
`scene_plan`	Subject, setting, mood, and key visual elements.
`generation_plan`	Required keys `positive_prompt`, `negative_prompt`, `guidance_note`.

Table 4: Planner output contract. Missing or malformed keys are normalised to their default values.

A.1.2 FIELD MAP FOR $z = (s, q, h, r, b)$

Table 5 maps each component of z to the record fields that carry it. Paths under `knowledge_evidence` are abbreviated with `...` for space.

The full `world_model` scene-reading record additionally contains the following audit fields, which do not feed generation directly:

```
scene_logic_to_preserve, prompt_scene_regime, scene_regime_prior,
scene_spec, allowed_scene_roles, prompt_safe_constraints,
novel_object_identity_constraints, period_reasoning_inputs,
style_translatable_regions, role_coverage_audit, uncertainty_notes,
unresolved_uncertainties, frozen_user_intent_constraints,
confidence, build_trace
```

A.1.3 MEMORY STRUCTURE

The memory m persisted between rounds has three components. The *trajectory memory* stores one record per round, containing the plan executed (positive prompt, negative prompt, guidance note), the evaluation outcome (global-critic score and critique, failure tags, repair targets, AUTHCRITIC score and yes-probability), and the selection record. The *reflective memory* is the cross-round distillation read at the next round's DERIVE step and included in the planner's prompt context: it holds the current meta-strategy, the best score and its iteration, the locked backend, persistent failure tags, stable strengths, and the current repair focus. The *visual memory* records per-iteration visual observations plus a persistent summary, grounding the textual failure tags in image evidence.

z	Component	Realising fields	Lifecycle
s	Scene reading	`world_model.scene_skeleton`; `world_model.key_local_roles`; `...scene_regime_analysis`; `scene_plan`.	Base rebuilt from fixed inputs; planner-revisable under the artist merge.
q	Preserve / transform	`world_model.preserve_vs_translate` (with per-target `translation_axes`); `world_model.structure_anchors_to_preserve`.	Base rebuilt from fixed inputs; planner-revisable under the artist merge.
h	Historical / style-regime intent	`...selected_period_bias`; `...period_guidance`; `...period_policy`; `...clarifier_period_signal`.	Working hypothesis; planner-revisable.
r	Retrieval targets, anti-shortcut	`...selected_motif_cues`; `...selected_composition_cues`; `...selected_local_motifs`; `...motif_patch_binding_plan`; `...patch_reference_bundle`; `...subject_kb_support` (incl. `controller_policy`); `world_model.anti_iconic_risks`; `world_model.forbidden_entities` / `forbidden_scene_roles`.	Cues re-retrieved each round; constraints generated from memory.
b	Backend-facing preferences	`tool_plan`; `routing_plan`; `generation_plan`; `...artistic_control_signals`.	Generated each round; lock fields persist.

Table 5: Field map from z to runtime record paths and lifecycles. Paths prefixed with `...` are relative to `knowledge_evidence`; explicit user periods remain fixed across compiled calls.

A.2 Knowledge-Base Construction and AuthCritic Training

The knowledge base contains a global layer of artist- and period-level priors, a curated supplementary layer of art-historical summaries, and a local layer of annotated patches. The global layer captures artist-level style characteristics, period-specific tendencies, and recurring motif and composition priors. The supplementary layer contributes curated art-historical evidence that reinforces period and motif annotations. The local layer contains a bank of patches extracted from source artworks; each patch is annotated with a semantic role, source-work metadata, local style traits, material treatment, and brushwork description. The patch-grounded component is built through a four-stage offline pipeline. (1) Image-level patch planning associates each source artwork with semantically meaningful target regions and roles. (2) Grounded-SAM (Ren et al., 2024)-based extraction obtains candidate local patches conditioned on these targets. (3) A vision-language model performs structured annotation, producing semantic descriptions of content and localized style descriptors such as brushwork and palette. (4) Annotated patches are filtered for quality and grouped into family-level reference sets used during retrieval. At runtime, retrieval returns period guidance, selected motifs, reference works, and local patch bindings.

The Van Gogh AuthCritic adapter is a Gemma-4 E4B-it LoRA model with rank 16, $\alpha = 32$, dropout 0.05, trained for two epochs on a single H800 GPU over 15,000 balanced patches (real Van Gogh, synthetic Van Gogh-style, and other-painter). The real-patch class is drawn from the offline patch bank. The synthetic class is cropped from Van Gogh-style outputs of six text-to-image models spanning open-weight and closed-source families. The other-painter class draws from contemporaries and near-contemporaries. Training minimises cross-entropy over all target tokens, uses AdamW with learning rate 2×10^{-4} and effective batch size 8, and reaches 91.65% binary accuracy on 1,485 held-out patches. The Qi Baishi adapter reuses this configuration on 4,806 patches with a 3,846/480/480 split, three epochs, and effective batch size 16, selecting the checkpoint with the lowest validation loss; it reaches 93.54% binary and 89.38% three-class accuracy on 480 held-out patches. Its three source classes comprise real Qi Baishi patches, synthetic patches cropped from Qi Baishi-style generations, and other-painter patches from Wu Changshuo and Zhu Da, two stylistically adjacent ink painters. For both artists the three source classes are balanced within every split, and splits are disjoint at the level of source works. Candidate aggregation and selection thresholds are specified in Appendix A.4.

A.3 Runtime Algorithm

Algorithm 1 details the closed loop summarized by the state chain of §3.

Algorithm 1: Runtime loop of `Atelier`

Require: request x, artist knowledge base $\mathcal{B}_a$, iteration budget T
Ensure: final image $y^\star$, trajectory τ, and final memory $m^\star$

$m^{(0)} \leftarrow \emptyset$, $\tau \leftarrow \emptyset$
$z^{(0)} \leftarrow \textsc{Perceive}(x)$ {fixed interpretation context}
for $t = 1$ to T **do**
  $z^{(t)} \leftarrow \textsc{Derive}(z^{(0)}, m^{(t-1)})$ {base state; memory-derived constraints}
  $k^{(t)} \leftarrow \textsc{Retrieve}(x, z^{(t)}; \mathcal{B}_a)$ {bind global/local evidence}
  $(\tilde{z}^{(t)}, c^{(t)}, e^{(t)}) \leftarrow \textsc{Plan}(x, z^{(t)}, k^{(t)}, m^{(t-1)})$
  $z^{(t)} \leftarrow \textsc{Merge}(z^{(t)}, \tilde{z}^{(t)})$ {artist-conditioned merge}
  $\mathcal{Y}^{(t)} \leftarrow \textsc{Execute}(e^{(t)}, c^{(t)})$
  $\rho_A^{(t)} \leftarrow \textsc{AuthCritic}(\mathcal{Y}^{(t)})$ {patch scores and flagged regions}
  $\rho^{(t)} \leftarrow \textsc{GlobalCritic}(\mathcal{Y}^{(t)}, \phi(x, z^{(t)}, k^{(t)}), \rho_A^{(t)})$ {holistic critic with projected inputs}
  $\hat{y}^{(t)} \leftarrow \textsc{Select}(\mathcal{Y}^{(t)}, \rho^{(t)})$
  $\tau \leftarrow \tau \cup \{(\hat{y}^{(t)}, c^{(t)}, e^{(t)}, \rho^{(t)})\}$
  $m^{(t)} \leftarrow \textsc{Reflect}(m^{(t-1)}, \hat{y}^{(t)}, \rho^{(t)})$
  if $\hat{y}^{(t)}$ satisfies the acceptance criterion **then**
    break
  end if
end for
$y^\star \leftarrow \textsc{FinalSelect}(\tau)$
$m^\star \leftarrow m^{(t)}$
return $y^\star, \tau, m^\star$

At the start of each round, the runtime rebuilds a base scene-reading record from the request, its clarification, and the visual anchor, then combines that base with the planner's memory-conditioned working view. The merge is artist-conditioned: the Qi Baishi path retains the rebuilt scene skeleton, structural anchors, preserve/transform policy, local roles, and anti-shortcut fields, whereas the Van Gogh path retains a planner-emitted scene-reading record when one is present.

A.4 Runtime Evaluation and Selection

This section specifies the numeric thresholds and decision rules of the closed loop of Algorithm 1: the stop condition, the per-round winner rule, the trajectory-level FinalSelect rule, and the input contracts of the two critics. All thresholds and policy rules are held fixed across the experiments reported in the main text. Algorithm 1 and the rules below describe the full `Atelier` configuration.

A.4.1 Runtime Constants

Symbol	Value	Meaning
$\tau^\star$	85	Accept-side score threshold.
T	3	Iteration budget per episode.
$N^{(t)}$	1, 2, or 4	Backend calls in round t. Round 1 and re-exploration use the full track pool (four open-weight or two closed-source backends); backend-locked rounds use one.
δ	5	Global-score rerank margin.
τ_ℓ	40	Low-quality tie-break threshold.
τ_p	40	AuthCritic per-patch score threshold.
π_p	0.40	AuthCritic per-patch yes-probability threshold.
K_p	4	Maximum patches escalated to the global critic.

Table 6: Runtime thresholds, budgets, and routing cardinalities, applied uniformly to all main-text experiments.

Round 1 explores a track-specific backend pool. Its winning backend is locked for subsequent rounds unless the selected outcome is classified as low-quality: $s_g < \tau_\ell$, a `reject` recommendation, or an `uncertain` recommendation with low confidence. There is no additional 60-point lock gate. The two track-level pools are

$$\mathcal{B}_{\text{open}} = \{\text{FLUX2-LORA}, \text{QWEN-T2I}, \text{LONGCAT-T2I}, \text{HUNYUAN-T2I}\},$$
$$\mathcal{B}_{\text{closed}} = \{\text{GPT-IMAGE-2}, \text{NANO-BANANA-PRO}\}.$$

A.4.2 GLOBAL-SCORE AGGREGATION

The global critic returns eight subscores $x_j \in [1, 5]$ and an artifact penalty $a \in [0, 3]$. The base score is

$$s_{\text{base}} \;=\; \text{round}\Big(\,\text{clamp}\,\Big(\tfrac{1}{5}\textstyle\sum_j w_j x_j \;-\; 7a,\; 0,\; 100\Big)\Big),$$

with weights w_j: style authenticity 18, period match 18, intent preservation 12, brushwork directionality 12, impasto texture 12, motif match 10, composition match 10, and palette match 8. The base score is then capped by the critic's own recommendation (59 for reject, 69 for uncertain, 84 for revise) and passed through a second calibration step that adjusts for stated confidence, blocking failure tags, actionable gaps, and request-anchor violations; the calibrated value is the score s_g used by the stop condition and FINALSELECT. The score is therefore deterministic given the critic's structured output, but it is not a simple mean of the subscores. Both artist schemas share this aggregation. For schema compatibility, the Qi Baishi rubric interprets `period_match` as agreement with the selected artistic regime and retrieved evidence rather than a chronological period, and `impasto_texture` as ink-on-paper material control, including absorption, wet and dry transitions, edge variation, paper reserve, and ink-density modulation; a high value does not denote thick oil paint. The Qi Baishi runtime uses the grounded-reasoning AUTHCRITIC adapter `qi_baishi_gemma4_grounded_reasoning_checkpoint-482`.

A.4.3 STOP CONDITION

Let $s_g^{(t)}$, $\rho^{(t)} \in \{\text{accept}, \text{revise}, \text{reject}, \text{uncertain}\}$, and $F^{(t)}$ denote the score, recommendation, and canonical failure-tag set at round t. Define $o(A, B) = |A \cap B|/|A \cup B|$, with $o(\varnothing, \varnothing) = 1$. The deployed policy applies the following terminal conditions in order.

Success. The episode terminates when

$$s_g^{(t)} \geq \tau^\star \quad \wedge \quad \rho^{(t)} = \text{accept}. \tag{1}$$

Both terms are required.

Plateau. A hard near-threshold plateau is detected after three consecutive `revise` recommendations when

$$\max_{j=t-2}^{t} s_g^{(j)} \geq \tau^\star - 1, \quad \max_{j=t-2}^{t} s_g^{(j)} - \min_{j=t-2}^{t} s_g^{(j)} \leq 1, \quad o(F^{(t)}, F^{(t-1)}) \geq 0.5. \tag{2}$$

The implementation also contains a four-revision generic plateau guard with score span at most 2, mean overlap with the preceding two tag sets at least 0.6, and maximum score below $\tau^\star - 3$; this branch is unreachable for the reported budget $T = 3$. A two-round near-plateau with score span at most 1, overlap at least 0.5, and maximum score at least $\tau^\star - 3$ is logged as a risk but is not terminal.

Terminal reject, budget, and error. The episode stops early after at least two recent `reject` decisions at scores no greater than 59 when the current score does not improve on the preceding recent scores. Otherwise it stops at $t = T$, or when a step remains failed after all configured retries. FINALSELECT (Appendix A.4.4) then selects $y^\star$ from the completed trajectory.

A.4.4 CANDIDATE SELECTION

Let $s_g(y)$ be the global-critic score, $s_a(y)$ and $p_a(y)$ the AUTHCRITIC candidate-level score and real-patch probability, and $i(y)$ the candidate index within a round. When at least one candidate has a usable AUTHCRITIC sidecar, within-round selection first forms

$$\mathcal{S}_t = \{y : s_g(y) \geq \max_{y'} s_g(y') - \delta\}, \tag{3}$$

then applies

$$\hat{y}^{(t)} = \arg\operatorname*{lex\,max}_{y \in \mathcal{S}_t} \big(s_a(y), p_a(y), s_g(y), \mathbb{1}[\rho(y) = \text{accept}], \mathbb{1}[\text{conf}(y) = \text{high}], -i(y)\big). \tag{4}$$

Without a usable sidecar, candidates are ordered by global score, accept flag, high-confidence flag, backend preference, and earlier index. An exact tie under a low-quality outcome prefers a non-FLUX candidate; otherwise the default backend tiebreak favours FLUX.

FINALSELECT operates over the full trajectory and uses a related but distinct rule. Let $s_2(y)$ denote the stored second-stage score and $t(y)$ the generation round. With usable local evidence, it applies the same δ-margin global gate and selects

$$y^\star = \arg\operatorname*{lex\,max}_{y \in \mathcal{S}} \big(s_a(y), p_a(y), s_g(y), s_2(y), -t(y)\big). \tag{5}$$

Without usable local evidence, it selects lexicographically by $(s_g(y), s_2(y), -t(y))$ without the margin gate. A round is marked low-quality, and therefore remains in multi-backend exploration, when its winner has $s_g < \tau_\ell$, is rejected, or is both uncertain and low-confidence.

A.4.5 AUTHCRITIC OUTPUT CONTRACT AND ESCALATION

AUTHCRITIC is queried as a grounded reasoning model rather than a binary classifier. For each patch, it must emit a strict JSON object whose keys, in order, are:

```
source_type, grounded_title, grounded_period,
source_global_summary, patch_visual_description,
patch_style_description, reason_short,
difference_from_real_master
```

The `source_type` field is one of `real_master_patch`, `synthetic_master_style_patch`, or `other_painter_patch`; artist-specific aliases (e.g., `real_vangogh_patch`) are normalised at parse time. For real-master predictions, the model grounds the label in a work title and period; otherwise it articulates the axis on which the patch departs from the target artist's handling (brushwork, palette, contour, surface, or motion).

The two-stage evaluation described in Section 3.4 escalates any patch that falls below τ_p or π_p: the crop and its reasoning record are attached to the global-critic call as focus signals, so the global critic receives both a suspect region and a stated cause. At most $K_p = 4$ patches are escalated per candidate, ordered by ascending score.

A.4.6 AUTHCRITIC CANDIDATE AGGREGATION

Context-only crops are excluded when at least one local style-evidence patch is available; otherwise all sampled patches are used. Let f_r be the fraction predicted as real-master patches, $\bar{p}_s$ the mean proximity-to-real among synthetic-master predictions, f_v the fraction with a valid proximity parse, n the number of scoring patches, and e the proximity parse-error rate. If no synthetic-master patch has a valid proximity, $\bar{p}_s$ is taken as the mean over all valid proximities. The candidate score is

$$s_a = \operatorname{round}\Big(100\ \operatorname{clamp}_{[0,1]}\Big[0.5 f_r + 0.4\bar{p}_s + 0.1 \min\Big(\frac{n}{3}, 1\Big) f_v - 0.2e\Big]\Big). \tag{6}$$

The associated probability p_a is f_r.

A.4.7 GLOBAL-CRITIC INPUT PROJECTION

The global critic is Kimi K2.6 (Moonshot AI, 2026); its fixed-schema projection contains the request, clarification, period hypothesis, selected motif and composition cues, and any low-scoring patch evidence.

The global critic does not receive the full state z or evidence bundle k. Its prompt is built from the projection $\phi(x, z, k)$ of the request and the current round's state: the raw request, the clarified request, the working period hypothesis and its candidate alternatives, the selected motif cues, and the selected composition cues, together with the per-patch metadata of any low-scoring patches returned by AUTHCRITIC (identifier, crop box, score, predicted label, and grounded reasoning). The candidate image and the flagged patch crops are attached as visual inputs. The remaining fields of z and k condition retrieval, planning, and generation but are not part of the critic prompt.

A.4.8 GLOBAL-CRITIC RESPONSE CONTRACT

The global critic returns a `local_patch_review` array together with the scalar fields: one review item per escalated AUTHCRITIC hypothesis, in the order the hypotheses were supplied. Each item

marks its hypothesis confirmed or rejected, carries a region-specific repair hint when confirmed, and states why the local hypothesis does not hold in full-image context when rejected. Scalar fields, failure-tag vocabularies, and score aggregation follow Appendix A.4.2.

If a response violates the JSON or local-patch contract, each subsequent attempt appends the suffix in Prompt 15 to the same user message. The deployed runtime permits at most three total attempts.

The parser rejects a missing, non-array, misordered, or identity-mismatched `local_patch_review` rather than treating it as an empty review. The prompt-template and schema identifiers emitted with each stored evaluator record are listed in Prompt 16.

A.5 Control-State Walkthrough

This section traces two recorded episodes. Appendix A.5.1 exhibits all five components of the state $z = (s, q, h, r, b)$ (§3.1) on a single held-out request. Appendix A.5.2 follows a three-round refinement trace on a second request, showing how critic feedback is transported through the reflective memory into subsequent plans.

The runtime does not serialise z as one JSON object. Its components are distributed across the perceiver's scene specification, the recorded `world_model` field (the implementation record for the scene reading), the planner's `knowledge_evidence` block, and the compiled plans; the field-level mapping is given in Appendix A.1 (Table 5). The excerpts below group the recorded fields by the component of z they realise; evaluation fields keep the implementation's names (e.g., `heavy_judge_score` is the global-critic score s_g of Appendix A.4). Elisions are marked with `...`. Values and text are copied from the records; the backend excerpt omits compatibility-only arguments that the corresponding backend does not consume.

A.5.1 Single-Round Control State

The episode uses a held-out still-life request:

> *A low-angle still life of harvested vegetables—pale cabbage, reddish root crops, and scattered yellow maple leaves—resting on dark soil, rendered in Paris-period muted earth tones with directional impasto brushwork, arranged in a diagonal mass from lower-left to upper-right against a deep shadowed background.*

Scene reading s. The perceiver assigns the request to the closed scene-regime catalogue (Prompt 3) and emits three allowed local-role names. The scene-reading builder (recorded under the implementation field `world_model`) enriches these roles with the style axes along which they may be translated:

```
"scene_skeleton": {
  "primary_regime": "object_still_life",
  "active_regimes": ["object_still_life"],
  "setting_detail": "still_life_tabletop",
  "mood": "calm", ...
},
"key_local_roles": [
  {"scene_role": "foreground_object_mass",
   "translation_axes": ["surface_relief", "palette_relation", "
  contour_pressure"]},
  {"scene_role": "field_surface_band",
   "translation_axes": ["brush_rhythm", "palette_relation", "
  contour_pressure"]},
  {"scene_role": "background_halo_field",
   "translation_axes": ["directional_motion", "palette_relation", "
  surface_relief"]}
]
```

Preserve/transform decisions q. The same record separates preserved structure from translatable style axes:

```
"preserve": [
  {"target": "object_silhouette_integrity"},
  {"target": "tabletop_support_plane"}
],
"translate": [
  {"target": "foreground_object_mass",
   "translation_axes": ["surface_relief", "palette_relation", "
  contour_pressure"]},
  {"target": "field_surface_band",
   "translation_axes": ["brush_rhythm", "palette_relation", "
  contour_pressure"]},
  {"target": "background_halo_field",
   "translation_axes": ["directional_motion", "palette_relation", "
  surface_relief"]}
]
```


Historical/style-regime intent h. The request names the Paris period, so the perceiver records an explicit period signal; the downstream period policy consequently collapses to a single route:


```
"clarifier_period_signal": {
  "explicit_period_preference": "paris_period",
  "period_clarification_mode": "explicit", ...
},
"period_policy": {
  "mode": "single_route",
  "dominant_period": "paris_period",
  "candidate_periods": ["paris_period"],
  "reason": "explicit_period_preference"
},
"period_guidance": [
  {"period_label": "paris_period", "score": 37.4,
   "top_form_tags": ["expressive brushwork", "complementary colors", "
  greens", ...],
   "dominant_palette_relations": ["complementary colors", "greens", "
  lighter palette", ...], ...},
  {"period_label": "nuenen_period", "score": 33.4, ...}
]
```


Retrieval targets and anti-shortcut constraints r. Each scene role is bound to a patch family, or explicitly left unbound. Here only one role finds a role-consistent family; the other two are bound to work-level evidence instead:


```
"motif_patch_binding_plan": [
  {"motif_name": "foreground object mass",
   "scene_role": "foreground_object_mass",
   "family_id": "nuenen_period::foreground_object_mass::
  foreground_object_mass",
   "binding_status": "bound_patch_family",
   "representative_patch_ids": ["patch_03958"],
   "family_cluster_size": 6, ...},
  {"motif_name": "field surface band",
   "binding_status": "unbound_patch_family_fallback_to_work_evidence",
   ...},
  {"motif_name": "background halo field",
   "binding_status": "unbound_patch_family_fallback_to_work_evidence",
   ...}
]
```


The bound family is indexed under `nuenen_period` although the generation route is Paris: the period label enters retrieval as a scoring term rather than a hard filter, so an adjacent-period family

can win the rerank while the generation route is unchanged. The per-candidate audit records the score decomposition; each column is a term of the family scoring rule (§3.2):


```
"selected_candidates": [
  {"patch_id": "patch_03958",
   "source_work_title": "Still Life (F.1972.44.P)",
   "score": 61.8,
   "score_breakdown": {
     "period_score": 24.0,
     "reference_work_prefilter_bonus": 8.0,
     "patch_role_score": 20.0,
     "canonical_work_bonus": 2.2,
     "asset_priority_bonus": 4.0,
     "world_model_alignment_bonus": 3.6,
     "anti_iconic_penalty": 0.0, ...}}
], ...
```


The knowledge base has no subject-level evidence for this request; the record encodes this as a policy over what the planner may use, not as a substitution:


```
"subject_kb_support": {
  "subject_kb_support_level": "none",
  "has_subject_patches": false,
  "matched_motif_count": 0,
  "controller_policy": {
    "allow_style_guidance": true,
    "allow_compatible_geometric_anchors": true,
    "allow_motif_style_analogy": false,
    "allow_subject_content_anchors": false
  }, ...
}
```


Backend-facing preferences b**.** The planner compiles the state into $N = 3$ parallel backend calls with a shared generation plan. These calls are the case-specific subset selected for this episode, rather than an enumeration of the track-level backend pool in Appendix A.4.1:


```
"tool_plan": {
  "primary_tool_id": "remote_flux_lora",
  "calls": [
    {"tool_id": "remote_flux_lora",
     "arguments": {"artist": "van_gogh",
                   "period_resource": "paris_period",
                   "style_strength": 0.9, "lora_scale": 0.9, ...}},
    {"tool_id": "qwen_t2i",
     "arguments": {"period_resource": "paris_period",
                   "style_strength": 0.8, ...}},
    {"tool_id": "longcat_t2i",
     "arguments": {"period_resource": "paris_period",
                   "style_strength": 0.8, ...}}
  ]
},
"generation_plan": {
  "positive_prompt": "Low-angle still life oil painting in the style of
  Vincent van Gogh's Paris-period. A mass of harvested vegetables
  arranged diagonally from lower-left to upper-right ... Muted earth
  tones, warm ochres, olive greens, deep browns ... Thick, directional
  impasto brushstrokes ...",
  "negative_prompt": "glossy, photorealistic, smooth, flat, cartoon,
  bright saturated colors, pure black background, signature, watermark,
  ..., bright blue sky, sunflower, rural landscape, over-decorative halo
  .",
  "guidance_note": "... Preserve structural anchors:
  object_silhouette_integrity, tabletop_support_plane. Translate style
```

```
  through: foreground_object_mass via surface_relief, palette_relation,
  contour_pressure; field_surface_band via brush_rhythm, palette_relation
  , contour_pressure; background_halo_field via directional_motion,
  palette_relation, surface_relief."
}
```

The internal identifier `remote_flux_lora` denotes the FLUX2-LORA backend. Only this call consumes `lora_scale` as an adapter weight. The raw planner record also places a compatibility `lora_scale` field on the Qwen and LongCat calls; the executor ignores that field for these backends, so it is omitted above. Their `style_strength` values provide prompt-level guidance and do not load a LoRA adapter. Hunyuan belongs to the open-weight pool but was not selected in this episode's three-call plan.

The compiled plan carries the fields of r and q into the backend text: canonical Van Gogh shortcuts appear as negative-prompt terms (`sunflower`, `bright blue sky`), and the guidance note restates the preserve targets and per-role translation axes verbatim, so the same constraints reach the critic in §3.3.

A.5.2 THREE-ROUND REFINEMENT TRACE

This subsection follows a completed open-weight Qi Baishi episode. The recorded request is:

> *Qi Baishi-style ink-and-wash composition featuring bold, wet-ink lotus leaves with expressive white highlights, a vivid magenta lotus bloom, and two colorful mandarin ducks rendered in simplified forms with yellow, blue, and purple washes; vertical calligraphy and a red seal appear on the left, while ample blank paper space enhances the dynamic asymmetry.*

The evaluator provenance matches the contract in Appendix A.4.8: the record names the Qi Baishi global-critic prompt and schema together with the checkpoint-482 AUTHCRITIC adapter. The first round explores the four-backend pool. Hunyuan supplies the winner and is then locked for the remaining rounds; the trace therefore holds the selected generator fixed while the execution prompt is revised.



Figure 4: Three-round refinement trace for a Qi Baishi request, shown from left to right: Round 1 ($s_g = 47$), Round 2 ($s_g = 49$), and Round 3 ($s_g = 57$). Hunyuan is locked after Round 1.

Round 1. The candidate preserves the requested subjects and layout, but the critic identifies weak calligraphic energy, limited ink-density transitions, and a drift toward generic decorative illustration. These observations are stored as canonical tags together with localized repair targets:

```
"global_score": 47,
"canonical_failure_tags": [
  "style_calligraphic_energy_weak",
  "material_ink_density_transition_weak",
  "shortcut_generic_decorative_illustration",
  "brushwork_too_uniform"
],
"repair_targets": [
```

```
  "leaf edges: replace soft wash boundaries with tapered, dry-brush
  frayed terminations",
  "lotus bloom: replace solid magenta with fine ink contour + diluted
  magenta wash over graded ink base",
  "ducks: reduce color saturation, integrate ink underwash, simplify
  forms to 3-5 expressive strokes per duck"
]
```

The reflective memory records the canonical tags verbatim and separates the repair targets from preserved strengths, including the subject count, the left-side inscription and seal, and the blank-paper reserve. It also records the backend lock:

```
"locked_backend": "hunyuan_image_t2i",
"backend_locked_from_iteration": 1,
"canonical_failure_tag_history": [
  {"iteration": 1,
   "target_artist_id": "qi_baishi",
   "failure_tags": [
     "style_calligraphic_energy_weak",
     "material_ink_density_transition_weak",
     "shortcut_generic_decorative_illustration",
     "brushwork_too_uniform"]}
], ...
```

Round 2. The next generation prompt instantiates the recorded repairs as material and stroke instructions: dry-brush fraying on leaf edges, graded wash beneath the magenta bloom, ink underwash for the ducks, and pressure-varying calligraphic strokes. The subject and layout constraints are retained:

```
"generation_prompt": "Traditional Chinese ink wash painting in the xieyi
  freehand manner of Qi Baishi. Bold wet-ink lotus leaves with dry-brush
  frayed edges, irregular white paper reserve highlights. A vivid magenta
   lotus bloom with fine dark ink contour and diluted magenta wash. Two
  simplified mandarin ducks rendered in 3-5 expressive calligraphic
  strokes each ... Vertical column of calligraphy on the left ... Small
  red seal beneath calligraphy. Ample blank paper space ... Ink density
  variation from deep black to pale gray ..."
```

The global score increases to 49, and the candidate score rises from 42 to 67. The remaining local review calls for layered petal washes, dry-brush stippling, tapered stroke endings, and less uniform saturation.

Round 3. The third prompt carries these revisions forward. It specifies ink gradation within the leaves, translucent rather than solid petal color, dry-brush fraying, and abbreviated duck forms while preserving the requested count and composition:

```
"generation_prompt": "Master Qi Baishi-style ink-and-wash painting on
  absorbent xuan paper. A lively waterside scene: three bold lotus leaves
   executed with large wet-ink brushstrokes that show rapid pressure
  release and dry-brush fraying at the edges ... Leaf interiors show rich
   ink gradation from deep black to watery gray. A single magenta lotus
  bloom composed of fine calligraphic ink contours and translucent
  magenta wash over a graded ink underlay ... Two simplified mandarin
  ducks reduced to xieyi essentials ... Vertical column of calligraphic
  script on the left side ... a small red seal stamp placed near the
  lower-left corner ..."
```

The global score reaches 57 and FINALSELECT returns the Round 3 candidate. The episode remains below $\tau^\star = 85$ and terminates at the budget $T = 3$ with recommendation `revise`. Figure 4 is therefore a monotonic three-round refinement trace, not an accepted outcome. It isolates the

transport of localized critic findings through reflective memory into successive execution prompts while the selected backend remains fixed.

B Experimental Details and Additional Results

This appendix gives the benchmark design and experimental configurations, followed by supplementary experiments and qualitative comparisons. Each experimental unit includes its protocol, quantitative results, qualitative evidence where applicable, and interpretation.

B.1 Benchmark Design

Compared with style-transfer and stylization evaluations, which typically condition on a single exemplar to a few dozen reference images per style and report a single global metric (Gatys et al., 2016; Huang & Belongie, 2017; Sohn et al., 2023), ArtIntentBench is larger in case count and richer in task diversity, covering re-rendering, period-controlled generation, cross-tradition transfer, and unseen-subject generalization within a single evaluation protocol. General generation and editing benchmarks evaluate instruction following and compositional control at scale (Wu et al., 2024; Ye et al., 2025; Pan et al., 2025); to our knowledge, none couples artist-grounded intent with period control, content preservation, and shortcut auditing.

Metric rationale. IntroStyle embeds images in a style-attribution feature space designed to isolate stylistic execution from depicted content; the W_2 distance is computed between the feature distribution of a method's outputs and that of a reference set of authentic paintings, so a low score reflects proximity to the artist's stylistic range rather than to any single exemplar. For period-controlled generation the reference set is restricted to the target period, and for unseen subjects it contains 20 authentic Van Gogh works balanced across the four periods, five per period.

B.2 Experimental Configurations and Baselines

The open-weight configuration dispatches generation to FLUX2.Klein-9B (Black Forest Labs, 2026), Qwen-Image (Wu et al., 2025), LongCat-Image (Ma et al., 2025), and Hunyuan Image 3.0 (Cao et al., 2025); the closed-source configuration dispatches to Nano Banana Pro (Gemini 3 Pro Image Preview) (Google, 2025) and GPT-Image-2 (ChatGPT Images 2.0) (OpenAI, 2026a). The same six systems also serve as direct-backend baselines. In artwork re-rendering, each source painting is converted into a structured caption by Qwen3-VL-Plus (Bai et al., 2025), and all methods generate from the caption without access to the original image. No baseline receives a minimal prompt of the form "a scene in the style of artist"; each receives a standardized prompt-engineered instruction containing the scene description, the target artist and period or style regime, explicit preservation requirements, and negative instructions against unsupported subject substitution. Baselines are run at the same generation budget as Atelier.

In the non-interactive evaluation of §4, the perceiver's follow-up-question output is never exercised; the operative role of the clarification and intent-inference prompt is intent inference, which feeds the control state directly.

The explicit-reference planner block (Appendix C.1.4) was inactive in the located controller roots of the re-rendering, period-controlled, and Qi Baishi experiments.

B.3 Automatic Control-State Audit

The final image alone cannot reveal whether the system inferred the right controls. We therefore automatically audit Atelier's intermediate state across all four task settings. The audit checks period propagation, preserve/transform well-formedness, role-specific evidence binding, and, for unseen subjects, identity protection and anti-shortcut checks. Each metric checks field consistency rather than semantic quality; the complementary LLM artistic-signal audit is reported in Appendix B.4.

The state propagates explicit period labels perfectly, while caption-based period recovery is harder (top-1 38–40%; top-2 62–66%). Preserve/transform fields are well formed in 92.0–100.0% of cases. Evidence binding is strong for Van Gogh but falls to 45.0% for Qi Baishi because its patch bank is

Table 7: Automatic audit of Atelier's intermediate control state. Each metric checks whether a field is populated and internally consistent; it does not judge the semantic quality of the selected content.

Property	Open-weight	Closed-source
Van Gogh artwork re-rendering		
Source-regime recovery (top-1)	40.0%	38.0%
Source-regime recovery (top-2)	66.0%	62.0%
Preserve/transform structural well-formedness	92.0%	100.0%
Patch-backed transform	80.0%	84.0%
Patch soft-fail	20.0%	16.0%
Van Gogh period-controlled generation		
Source-regime recovery	100.0%	100.0%
Preserve/transform structural well-formedness	94.7%	95.6%
Patch-backed transform	74.1%	95.6%
Patch soft-fail	8.8%	4.8%
Historically unseen subject (Van Gogh)		
Novel identity guard present	96.7%	93.3%
Anti-shortcut pass	100.0%	96.7%
Preserve/transform structural well-formedness	96.7%	73.3%
Evidence-backed transform	100.0%	90.0%
Patch soft-fail	1.3%	3.3%
Qi Baishi artwork re-rendering		
Preserve/transform structural well-formedness	96.3%	93.8%
Evidence-backed transform	45.0%	45.0%
Patch soft-fail	52.5%	52.5%
Patch hard-fail	3.4%	2.5%

smaller; the low hard-fail rate shows that this is a coverage limitation rather than a binding failure. On unseen subjects, identity guards and anti-shortcut checks pass in at least 93.3% of cases, directly supporting the shortcut results in Section 4.2.

B.4 Artistic Control-Signal Audit

The automatic audit above checks field-level consistency. We complement it with two independent LLM judges, DeepSeek V4 Pro and GPT-5.5, using task-matched 0–10 protocols. For period-controlled generation, judges score palette, brushwork, composition, and motif match. For artwork re-rendering, they score source retention, preserve/transform policy, period-style adaptation, and patch-evidence usefulness. For Qi Baishi, the same four axes are interpreted with respect to its stylistic conventions. No fixed keyword list is supplied; judges apply the corresponding artistic criteria case by case.

Judge protocols. The Control State Quality evaluation (Section 4.4) audits the intermediate control state with artist-instantiated LLM-judge protocols. The Van Gogh audits use three task-specific protocols: period-controlled generation (Prompt 19), source-preserving artwork re-rendering (Prompt 20), and historically unseen subjects (Prompt 21). The Qi Baishi artistic-signal panel uses the dedicated protocol in Appendix C.2.2. Each protocol is given to two independent judge models; the protocols fix the scoring dimensions, the composite rule, and the output schema. The re-rendering and unseen protocols additionally include few-shot calibration examples. The prompts in Appendix C.2.1 retain the dimensions and fields used in the reported analysis.

Within the deployed period-judge prompt, "no fixed rubric applies" denotes the absence of a scene-independent reference answer or prescribed stylistic treatment; the evaluated dimensions, scoring scale, and output schema remain fixed by the protocol.

The Qi Baishi panel of the artistic-signal audit uses a dedicated judge protocol rather than a reuse of the Van Gogh protocols; the deployed system prompt is reproduced verbatim in Prompt 22.

On Van Gogh re-rendering, source retention is the strongest dimension ($\geq$ 8.4), while patch-evidence usefulness is weakest (4.79–6.01), identifying knowledge-base coverage as the main limitation. On period control, brushwork is strongest and motif/composition weakest; Paris is the weakest period for both judges. On unseen subjects, composite scores are highest (open-weight

Table 8: LLM-judge evaluation of Atelier's intermediate control state on a 0–10 scale. DeepSeek V4 Pro and GPT-5.5 score each task independently.

	Atelier (Open-weight)		**Atelier (Closed-source)**	
	DeepSeek	GPT-5.5	DeepSeek	GPT-5.5
Panel A: Artwork re-rendering (Van Gogh)				
Source retention	8.50	8.81	8.42	8.72
Preserve/transform	6.69	7.24	6.90	7.65
Period-style adaptation	7.83	7.99	7.70	7.90
Patch evidence usefulness	4.79	5.68	5.00	6.01
Composite	6.95	7.43	7.00	7.57
Inter-judge r	0.59		0.66	
Panel B: Period-controlled generation (Van Gogh)				
Palette match	7.66	5.83	7.62	5.70
Brushwork match	8.63	6.77	8.61	6.58
Composition match	7.36	5.02	7.33	4.83
Motif match	7.37	4.87	7.36	4.72
Paris	7.39	4.73	7.20	4.55
Arles	7.82	5.92	7.74	5.72
Saint-Rémy	8.21	5.88	8.22	5.66
Auvers	7.58	5.98	7.75	5.89
Inter-judge r	0.43		0.49	
Panel C: Historically unseen subject (Van Gogh)				
Novel content coverage	9.18	7.86	9.00	8.07
Preserve/transform decomposition	9.27	8.14	8.55	6.80
Period-style translation	9.00	8.20	8.96	7.98
Shortcut avoidance	9.91	8.99	9.41	7.71
Composite	9.34	8.30	8.98	7.64
Panel D: Artwork re-rendering (Qi Baishi)				
Palette match	7.18	5.72	7.96	6.28
Brushwork match	7.60	6.34	8.22	7.04
Composition match	5.28	3.54	5.62	5.36
Motif match	4.34	3.16	4.78	4.54
Composite	6.10	4.69	6.65	5.81

8.30–9.34; closed-source 7.64–8.98), while shortcut avoidance remains high. On Qi Baishi, palette and brushwork transfer are stronger than composition and motif, and the lower composite scores expose the difficulty of transferring evidence from the smaller ink-painting patch bank. Inter-judge correlations range from 0.43 to 0.66; we therefore use these scores as convergent diagnostics rather than as absolute ground truth.

B.5 Ablation Studies

Table 9 reports the component ablation of §4.4 in full. We then ask whether a stronger prompt, rather than structured control, would suffice. We use the same open-backend candidate pool for all prompt-construction baselines and compare them with Atelier's native selector. Free-form, expert-template, and LLM-reference expansion do not use the curated knowledge base, patch bank, or explicit control-state schema; the last of these is the retrieval-augmented prompting family of §4.1, in which the planner proposes its own reference works. The prompted-critic variant retains the full pipeline but replaces the trained AuthCritic with a prompted vision–language critic. Table 10 shows that Atelier achieves the best score (75.64) and that letting the planner propose its own references is the weakest of the three expansion variants (81.87), so the curated knowledge base contributes more than reference retrieval alone.

B.6 Historically Unseen Subject Generation

This experiment tests whether Atelier can apply Van Gogh's visual language to modern subjects absent from the artist's oeuvre without replacing those subjects with canonical motifs. We use 30 prompts describing laptops, train stations, highways, skyscrapers, and nighttime cities. The same prompt, backend, and selection budgets are used as in the main-text evaluation.

Table 9: Component ablation on 40 period-controlled cases (10 subjects $\times$ 4 periods), with the backend pool and iteration budget held fixed.

Ablation	**IntroStyle** $W_2 \downarrow$
w/o Artist KB & Patches	84.43
w/o Iterative Writeback	<u>84.41</u>
w/o Structured Intent	85.96
w/o Both Feedback	84.84
Atelier	**81.79**

Table 10: Prompt-construction baseline comparison on ArtIntentBench. All variants share the same open-weight candidate pool.

Method	**IntroStyle** $W_2 \downarrow$
Free-form expansion	79.99
Expert-template expansion	77.54
LLM-proposed references	81.87
Atelier (prompted critic)	<u>77.19</u>
Atelier	**75.64**

Table 11 reports IntroStyle W_2 and CLIP. Atelier obtains the lowest W_2 in both settings (86.61 closed-source and 91.28 open-weight). Its CLIP scores are mid-range in the closed-source setting and lowest in the open-weight setting, reflecting the expected tradeoff between literal text-token alignment and artist-specific material translation.

The qualitative comparison in Figure 5 shows the same pattern. Direct baselines apply surface texture without always preserving modern identity, and agent baselines sometimes replace a highway with a rural road or add arched structures to an airport. Atelier retains the glass ceiling and crowd depth of an airport terminal and the urban density of a skyscraper scene without defaulting to a Starry-Night composition.

B.7 Cross-Artist Fidelity Transfer to Qi Baishi

We evaluate whether the control-planning formulation transfers from Van Gogh's period-sensitive oil painting to Qi Baishi's sparse ink-and-color painting. The task is artwork re-rendering from 80 original Qi Baishi paintings, using text-only structured captions and the same matched evaluation protocol.

Table 12 reports style proximity, text–image alignment, and source fidelity. Atelier obtains the lowest IntroStyle W_2 in both tracks (57.09 closed-source and 68.33 open-weight) and the best DINO-cosine and LPIPS values in both tracks. DreamSim is led by Claude (NBP) in the closed-source track and Hunyuan in the open-weight track. These results show that the framework's control-state and evidence-binding formulation transfers, while also exposing the weaker coverage of the Qi Baishi patch bank.

As shown in Figure 6, FLUX often applies Western impasto, whereas Nano Banana Pro and Qwen recover ink-wash character less consistently. Atelier more reliably captures sparse composition, calligraphic placement, restrained color economy, and motif-centered brushwork across flowers, frogs, chrysanthemums, and willows.

B.8 Iterative Refinement

We measure the effect of reflective memory over three rounds on 45 cases. IntroStyle W_2 decreases from 83.02 in Round 1 to 79.61 in Round 2 and 79.50 in Round 3. The quantitative improvement therefore largely occurs by Round 2, while later rounds consolidate local corrections.

The qualitative trace shows that the first round establishes scene structure, whereas later rounds use critic feedback to improve local details, material handling, and style alignment. The trajectory-level selector can still return a candidate below the acceptance threshold when the budget is exhausted; the trace is therefore evidence of iterative correction, not a claim that every episode reaches acceptance.

Table 11: Historically unseen-subject generation on 30 modern prompts. IntroStyle W_2 uses 20 authentic Van Gogh references (five per period); CLIP measures text–image alignment.

Method	**Closed** $W_2 \downarrow$	**CLIP** $\uparrow$	**Method**	**Open** $W_2 \downarrow$	**CLIP** $\uparrow$
GPT-Image-2	89.96	0.3505	FLUX	98.49	0.3377
Nano Banana Pro	90.70	0.3418	Qwen	93.56	0.3225
GenArtist	92.71	0.3226	LongCat	102.68	**0.3505**
GenArtist (NBP)	88.21	0.3467	Hunyuan	94.36	0.3217
GenArtist (GPT-Image-2)	91.86	<u>0.3511</u>			
Claude (NBP)	<u>87.65</u>	0.3370			
Claude (GPT-Image-2)	91.34	**0.3559**			
Atelier (Closed-source)	**86.61**	0.3357	**Atelier (Open-weight)**	**91.28**	0.3137



Figure 5: Historically unseen-subject examples. Atelier preserves modern subject identity while applying period-appropriate color, brushwork, and composition.

B.9 Human Preference

The blinded user study used an online questionnaire with 10 art-trained non-specialists and 5 experts. Each participant completed 35 ranking questions covering three tasks: 20 period-controlled questions (five subjects across four Van Gogh periods), five artwork re-rendering questions, and ten historically unseen-subject questions. Each question displayed five anonymized outputs labelled A–E; method identities were independently randomized, and the source painting was shown for re-rendering. Participants ranked the outputs from 1 (best) to 5 (worst).

Table 12: Qi Baishi re-rendering on 80 cases, evaluated with IntroStyle W_2, CLIP, DreamSim (DS), DINO-cosine (DINO), and LPIPS (LP). Bold and underlined mark the best and second-best result within each configuration.

Method	$W_2 \downarrow$	CLIP↑	DS↓	DINO↑	LP↓	**Method**	$W_2 \downarrow$	CLIP↑	DS↓	DINO↑	LP↓
	Closed-source						*Open-weight*				
GPT-Image-2	59.71	0.3216	0.3653	0.4864	0.5998	FLUX	79.73	0.3125	0.4425	0.4427	0.6392
Nano Banana Pro	63.50	0.3119	0.3629	0.5083	0.6129	Qwen	83.33	0.3042	0.4431	0.4561	0.6353
GenArtist (GPT)	<u>57.98</u>	<u>0.3317</u>	0.3552	0.5097	0.6008	LongCat	75.42	**0.3244**	0.4280	<u>0.4631</u>	<u>0.6248</u>
GenArtist (NBP)	60.44	0.3221	0.3429	<u>0.5274</u>	0.5980	Hunyuan	<u>73.58</u>	0.3185	**0.4092**	0.4507	0.6278
Claude (GPT)	58.46	**0.3330**	0.3439	0.5110	<u>0.5943</u>	GenArtist	78.42	0.3171	0.4486	0.4158	0.6618
Claude (NBP)	61.09	0.3219	**0.3381**	0.5274	0.6005						
Atelier (Closed-source)	**57.09**	0.3194	<u>0.3397</u>	**0.5452**	**0.5804**	**Atelier (Open-weight)**	**68.33**	<u>0.3193</u>	<u>0.4133</u>	**0.4731**	**0.6154**



Figure 6: Cross-artist transfer to Qi Baishi. Atelier adapts artist-specific resources to sparse composition, calligraphic brushwork, restrained ink-color rendering, and motif-centered expression.

Table 13: Per-round IntroStyle W_2 averaged over 45 cases.

	Round 1	**Round 2**	**Round 3**
Mean $W_2 \downarrow$	83.02	79.61	**79.50**

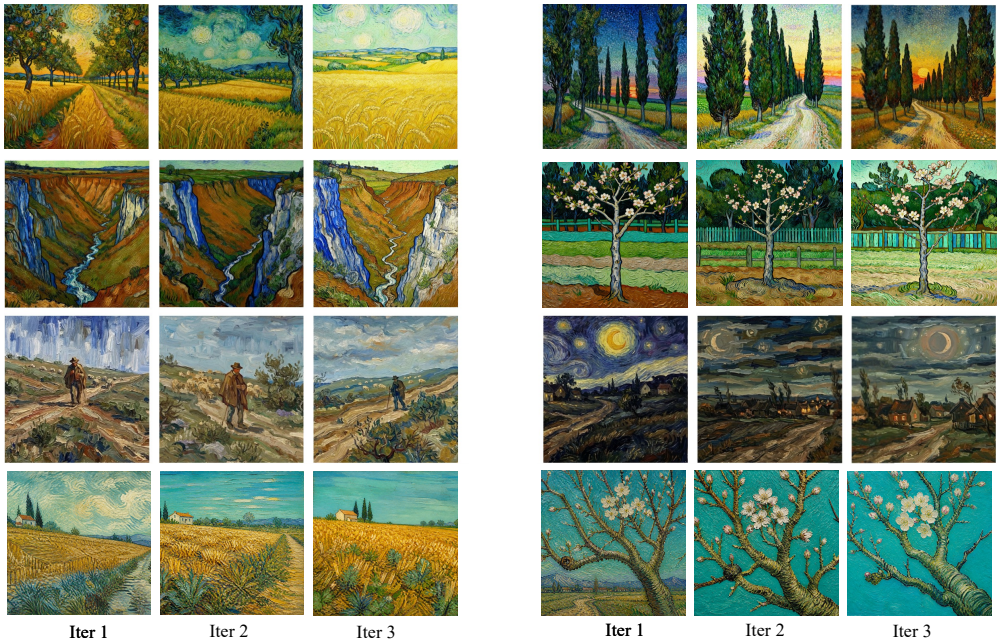

Figure 7: Three-round qualitative refinement. Each example shows the input followed by Rounds 1–3; later rounds improve local fidelity and target-style alignment.

Bradley–Terry analysis (Bradley & Terry, 1952), widely used for aggregating pairwise human preferences (Chiang et al., 2024), shows that preferences for Atelier and GPT-Image-2 over the three remaining methods are consistent across both rater groups, with pairwise preference probabilities exceeding 0.87 in all comparisons against Qwen and FLUX. As shown in Figure 8, `Atelier` receives the highest Rank-1 preference rate among both non-experts and experts, at 42% and 41% respectively, and the lowest Rank-5 rate among both groups, at 3% and 1% respectively. GPT-Image-2 ranks second among both groups; FLUX receives the most last-place rankings (65% from non-experts, 74% from experts). The near-identical preference profiles of the two groups suggest that the qualities separating `Atelier` from the baselines are visible without formal training, while the experts' harsher treatment of the weakest outputs indicates that their agreement does not reflect insensitivity to artistic detail.

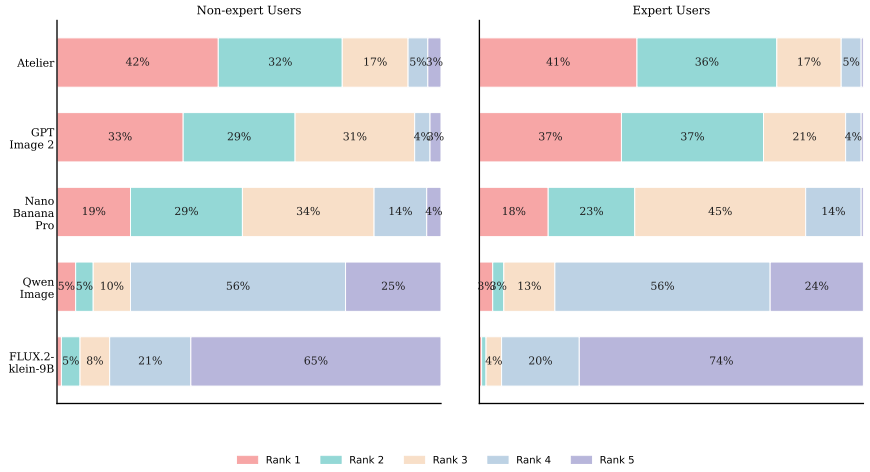

Figure 8: Human preference study: rank distribution across methods.



Figure 9: User-study interface. Each question shows five anonymized outputs and asks participants for a complete ranking.

We also fit a Bradley–Terry model (Bradley & Terry, 1952) to estimate pairwise preference probabilities. Values above 0.5 indicate that the row method is preferred to the column method; distance from 0.5 measures preference consistency. The pairwise probabilities for the expert and non-expert groups are shown in Figure 10.

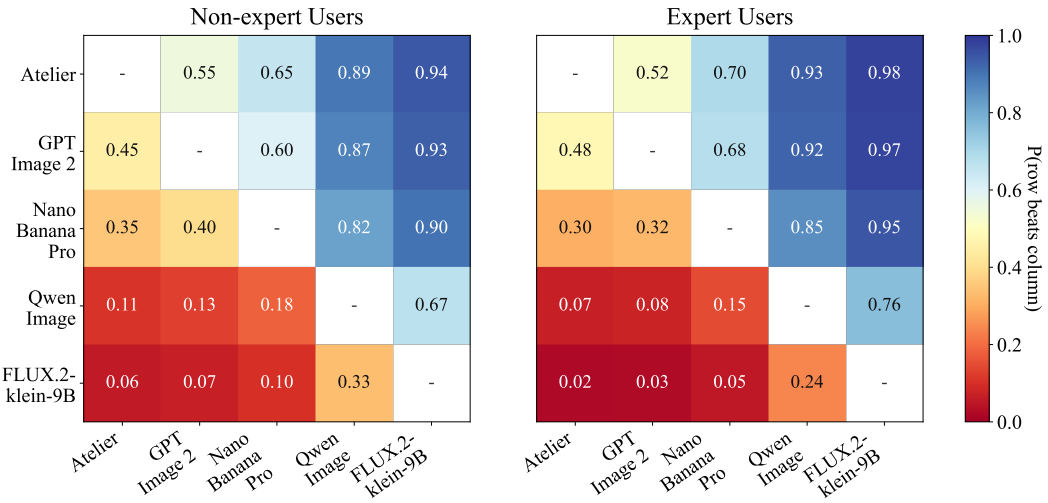

Figure 10: Bradley–Terry pairwise preference probabilities for expert and non-expert groups.

B.10 Qualitative Comparison

Figure 11 shows the full artwork re-rendering comparison. Claude (GPT-Image-2) preserves scene semantics better than direct baselines but adds *The Starry Night*'s swirling strokes and blue-yellow palette to two source paintings that contain neither cue. This response to the artist name contrasts with `Atelier`'s stronger preservation of composition, subject scale, spatial relations, brushwork, palette, and tonal layering.



Figure 11: Caption-based, text-only artwork re-rendering across all methods and both configurations.

Figures 12 and 13 show period-controlled generation on two scenes, with only the target period varied across rows. Claude (NBP) repeats the composition of *Wheatfield under Thunderclouds* across target periods, whereas FLUX and GPT-Image-2 repeat swirling strokes and the blue-yellow palette. `Atelier` changes palette, brushwork, and surface treatment while preserving scene structure and depth.

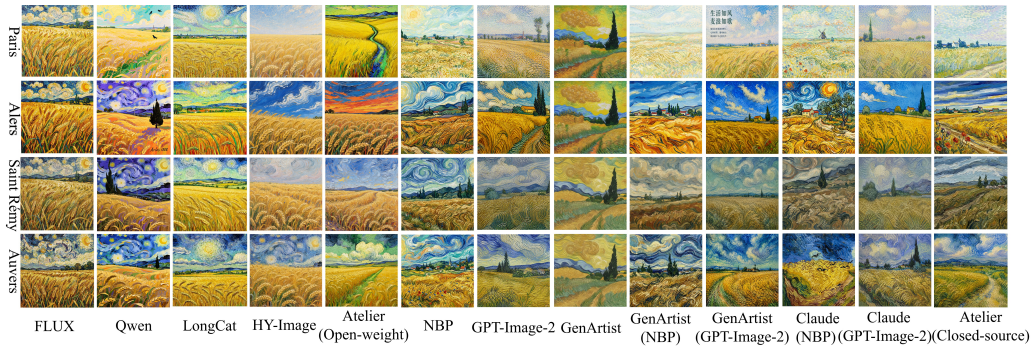

Figure 12: Period-controlled generation on the wheatfield scene across periods and methods.

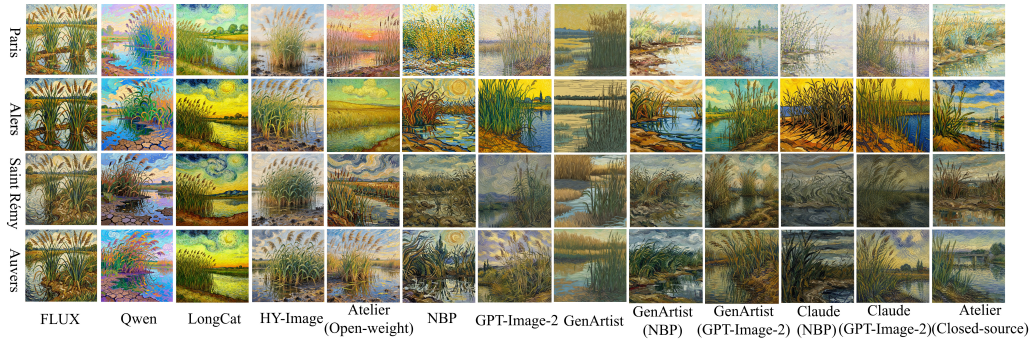

Figure 13: Period-controlled generation on the reeds scene across periods and methods.

B.11 Additional Qualitative Results

Figures 14 and 15 provide additional high-resolution outputs across subjects, periods, and artistic traditions. They are supplementary visual evidence rather than new evaluation cases.



Figure 14: Additional Van Gogh qualitative results across periods and subjects.



Figure 15: Additional Qi Baishi qualitative results across diverse subjects.

C Prompt Library

This appendix reproduces the prompts used by every LLM-based component of `Atelier`: the perceiver (Appendix C.1.1), the planner (Appendix C.1.2), the global critic (Appendix C.1.5), the local AuthCritic (Appendix C.1.8), and the judge protocols behind the Control State Quality evaluation, for Van Gogh (Appendix C.2.1) and for Qi Baishi (Appendix C.2.2). The two judge prompts marked as limited cover only the fields included in the reported analysis; the remaining prompts reproduce the deployed text.[1] Prompts are shown as templates: fields in curly braces, such as `{user_request}`, are filled at runtime; everything else is fixed instruction text. In the implementation the instruction sentences are concatenated into a single string; line breaks are inserted here for readability only.

C.1 Runtime Prompts

C.1.1 Perceiver Prompts

The perceiver issues two prompts. The *scene-spec extraction prompt* (Prompt 1) converts the raw request into the scene-facing fields of the control state: the scene regime, subject type, must-have and forbidden entities, and the scene roles that later drive patch-role binding. The *clarification and intent-inference prompt* (Prompt 2) additionally decides whether follow-up questions are needed, infers reference intent (whether the user implicitly anchors on a specific work or work family), and produces the period-preference signals used by retrieval. Their non-interactive use is specified in Appendix B.2. Both prompts constrain the model to closed catalogs of scene regimes and scene roles, reproduced in Prompt 3; the model may not invent regime or role names.

For compactness, Prompt 2 denotes the shared block from `EVIDENCE RULES` through the four `BOUNDARY EXAMPLES` in Prompt 1 by `{scene_spec_policy}`. The implementation inserts that block verbatim. The target-artist route supplies `{artist_specific_scene_spec_rules}`.

Prompt 1: Perceiver scene-spec extraction prompt.

[1] Module names in the implementation predate the terminology of this paper: the perceiver corresponds to the `clarifier` module and the planner to the `controller` module. Except for the two explicitly scoped judge prompts, prompt text is reproduced verbatim from the implementation, so these internal names still appear inside the prompt strings (e.g., "You are the Agent Controller ...").

```
You are the scene-spec extractor for an artistic image-generation system.
   Convert the user request into a minimal, evidence-grounded scene
  specification for downstream retrieval and planning.

Return one valid JSON object only. Do not use Markdown, comments, or text
   outside the JSON object. Return exactly these keys:

{
  "primary_regime": "<one value from the supplied regime catalog>",
  "subject_type": "<short noun phrase>",
  "must_have_entities": ["<explicitly requested entity>"],
  "forbidden_entities": ["<explicitly excluded entity>"],
  "allowed_scene_roles": ["<one to four values from the role catalog>"],
  "period_preference": "<explicit period label or empty string>",
  "confidence": <number from 0.0 to 1.0>,
  "reasoning_note": "<one concise sentence>"
}

EVIDENCE RULES

1. Base every output field on the user request.
2. Do not add objects, settings, periods, or scene roles merely because
  they are common in the target artist's work.
3. Put explicitly required subjects and objects in must_have_entities.
4. Put only explicitly negated or excluded entities in forbidden_entities
  . Use [] when the request contains no exclusions.
5. Negated content must never appear in must_have_entities.
6. subject_type describes the requested compositional subject, not the
  artistic style.

PERIOD RULE

Set period_preference only when the request explicitly names a period,
  date range, or unambiguous period alias. Otherwise return an empty
  string. Do not infer a period from the artist name, subject matter,
  palette, brushwork vocabulary, or resemblance to a canonical work.

REGIME SELECTION

Select primary_regime from the supplied regime catalog only. Classify the
   requested image by its dominant compositional organization, not by
  incidental content. Apply this precedence: (1) explicit framing or
  scene-type statements in the request; (2) the requested compositional
  subject and its framing; (3) the surrounding setting. The specific
  boundary rules below override this general precedence.

BOUNDARY RULES

- portrait_subject applies only to explicitly face-focused, bust-length,
  sitter-led, close-up portrait, or self-portrait framing.
- A person appearing in a larger scene, including a farmer, worker,
  reader, walker, or figure in a field, room, courtyard, path, street, or
   waterside setting, is not sufficient evidence for portrait_subject.
  Classify such requests by the surrounding scene organization, even when
   the person is visually important.
- floral_still_life applies only when flowers, a bouquet, or a flower-
  filled vase are the requested subject.
- object_still_life applies to non-floral foreground or tabletop objects,
   including fruit, bottles, vessels, shoes, boots, and tools.
- workspace_tabletop applies to desks or work surfaces organized around
  papers, notebooks, keyboards, tools, or office objects.
- modern_device_scene applies when a laptop, smartphone, computer, tablet
  , camera, headphones, or screen is a central requested subject.
- domestic_edge_exterior applies to small built exterior structures such
  as pavilions, walls, courtyards, porches, doors, and building edges,
```

```
  unless the request clearly centers an open rural landscape, a civic
  landmark, or another more specific regime.
- A potted flower arrangement on a table or windowsill is normally
  floral_still_life, not rural_landscape or interior_domestic, unless the
   room itself is explicitly the dominant subject.

ROLE SELECTION

- Select allowed_scene_roles from the supplied role catalog only.
- Return one to four roles. Two or three are preferred when supported,
  but never add a weakly supported role merely to reach a target count.
- Select roles that describe the requested composition, not motifs
  conventionally associated with the artist.
- For close foreground objects, prioritize foreground_object_mass. Use
  tabletop_edge_band only when a table or comparable support surface is
  requested or clearly entailed.
- facial_plane_focus and head_silhouette_anchor require explicitly face-
  focused, bust-length, or portrait framing. A human figure inside a
  larger scene does not qualify.
- facade_plane, roof_plane, shelter_frame_rhythm, bridge_span, and other
  architectural roles require the corresponding built structure to be
  requested as compositionally important.
- Do not select architectural roles merely because an object appears
  indoors or near a building.

ARTIST-SPECIFIC RULES

{artist_specific_scene_spec_rules}

BOUNDARY EXAMPLES

Each example output is a scene-spec object. When this policy is embedded
  in the full clarification task, place that object under scene_spec and
  still return every required top-level clarification key.

Example 1
User request: A close-up bust portrait of a woman against a pale wall.
Output: {
  "primary_regime":"portrait_subject",
  "subject_type":"close-up bust portrait",
  "must_have_entities":["woman"],
  "forbidden_entities":[],
  "allowed_scene_roles":["facial_plane_focus","head_silhouette_anchor","
  background_halo_field"],
  "period_preference":"",
  "confidence":0.96,
  "reasoning_note":"The request explicitly specifies close-up bust
  portrait framing."
}

Example 2
User request: A farmer walking along a path through a wheat field at dusk
  .
Output: {
  "primary_regime":"rural_landscape",
  "subject_type":"farmer in a wheat field",
  "must_have_entities":["farmer","path","wheat field"],
  "forbidden_entities":[],
  "allowed_scene_roles":["field_surface_band","foreground_path","sky_band
  "],
  "period_preference":"",
  "confidence":0.94,
  "reasoning_note":"The person belongs to a field scene rather than
  portrait framing."
}
```

```
Example 3
User request: A vase of irises on a windowsill, with no landscape outside
  .
Output: {
  "primary_regime":"floral_still_life",
  "subject_type":"floral windowsill still life",
  "must_have_entities":["vase","irises","windowsill"],
  "forbidden_entities":["landscape"],
  "allowed_scene_roles":["flower_cluster_mass","vase_body_anchor","
  tabletop_edge_band"],
  "period_preference":"",
  "confidence":0.97,
  "reasoning_note":"The flowers and vase are the dominant subject and the
   landscape is explicitly excluded."
}

Example 4
User request: A pair of worn boots on a wooden table, without flowers or
  buildings.
Output: {
  "primary_regime":"object_still_life",
  "subject_type":"foreground object still life",
  "must_have_entities":["worn boots","wooden table"],
  "forbidden_entities":["flowers","buildings"],
  "allowed_scene_roles":["foreground_object_mass","tabletop_edge_band","
  background_halo_field"],
  "period_preference":"",
  "confidence":0.97,
  "reasoning_note":"The request centers non-floral tabletop objects and
  explicitly excludes architectural content."
}

Scene regime catalog:
{scene_regime_catalog}

Scene role catalog:
{scene_role_catalog}

User request: {user_request}
```

Prompt 2: Perceiver clarification and intent-inference prompt.

```
You are a request clarifier for a personal artistic agent. The target
  artist for this request is {artist_display_name} (artist_id={artist_id
  }). Your job is structured condition completion for image generation,
  not therapy, not psychological counseling, and not privacy collection.
  If the user request is underspecified, decide what 1-3 missing details
  would most improve artistic control. Always also infer whether the user
   is referencing a specific {artist_display_name} work or work family as
   a style anchor, even if the wording is implicit. Always also infer a
  scene_regime_prior and scene_spec from the text request itself. Return
  strict JSON with keys: needs_clarification, ambiguity_score,
  missing_slots, questions, inferred_spec, reference_intent,
  scene_regime_prior, scene_spec, clarification_note. questions must be a
   list of objects with slot, question, why. reference_intent must be an
  object with keys: detected, explicitness, relation_type, granularity,
  candidate_targets, transfer_focus, preserve_subject_constraints,
  anti_copy_constraints, should_enable_reference_routing,
  should_request_visual_reference_disambiguation, confidence,
  reasoning_note.
candidate_targets must be a list of objects with title_hint, artist,
  confidence. scene_regime_prior must be an object with keys:
  primary_regime, secondary_regime, confidence,
```

```
  non_rural_structure_preservation, reasoning_note. scene_spec must be an
   object with keys: primary_regime, subject_type, must_have_entities,
  forbidden_entities, allowed_scene_roles, period_preference, confidence,
   reasoning_note. The following policy governs the scene_spec object:

{scene_spec_policy}

Only ask about truly useful missing factors such as mood, setting,
  time_of_day, weather, character_profile, palette_bias, expressive_goal,
   life_context, memory_or_present_focus, explicit_period_preference,
  period_preference_signals. Treat period_preference_signals as an artist
  -specific style-preference signal; do not hard-code another artist's
  periods or canonical examples. Prefer low-intrusion personalized slots
  like expressive_goal, life_context, and memory_or_present_focus when
  they help image control. Do not ask for numeric age. Only use age_stage
   if the user already frames the request that way and it is directly
  useful for image generation. If the user sounds like an expert or
  mentions artist stages/resources directly, you may ask
  explicit_period_preference. Otherwise prefer period_preference_signals
  phrased as visual feeling or artist-specific style axes rather than
  period names.
{artist_style_question_text}
If the request implies something like 'paint this subject the way {
  artist_display_name} handled a known work family', mark
  reference_intent.detected=true, explicitness=implicit, preserve the
  requested subject, and do not let the reference work replace the user
  subject. For scene_spec, do not let forbidden or negated entities
  become positive motif cues.

Scene regime catalog:
{scene_regime_catalog}

Scene role catalog:
{scene_role_catalog}

User request: {user_request}
```

Prompt 3: The closed scene regime and scene role catalogs inserted into the perceiver prompts at the {scene_regime_catalog} and {scene_role_catalog} placeholders.

```
Scene regime catalog:
- floral_still_life: flower/bouquet/vase-centered tabletop or windowsill
  still life only; not a generic still-life fallback
- object_still_life: first-class non-floral tabletop object still life
  such as teapot, cup, bottle, fruit, boots, shoes, tools, vessels, or
  other close foreground objects
- workspace_tabletop: first-class desk/workspace tabletop arrangement
  with non-floral objects, papers, keyboard, notebook, tools, or office
  surface organization
- modern_device_scene: first-class modern device scene centered on laptop
  , smartphone, computer, tablet, headphones, camera, screen, or similar
  device identity
- interior_domestic: indoor room scene, often with furniture/window or
  human-in-interior
- interior_transient_institutional: corridor, stair, landing, or
  institutional transition space whose circulation structure dominates
- domestic_edge_exterior: small built exterior edge such as courtyard,
  pavilion, wall, porch, doorway
- civic_landmark_or_plaza: church/chapel/plaza/landmark-led architecture
  scene
- urban_street: street/city/cafe/market scene
- transit_mobility: tram/station/platform/railway scene
- waterside_nature: river/canal/shore/water-led scene
```

```
- rural_landscape: open rural field/hillside/village-landscape scene when
   no stronger built or still-life structure dominates
- portrait_subject: close-up or sitter-led portrait scene
- industrial_edge: factory/warehouse/mill/industrial built scene
- overlook_elevated_view: rooftop/terrace/balcony/overlook scene

Scene role catalog:
- facial_plane_focus: facial plane focus
- head_silhouette_anchor: head silhouette anchor
- background_halo_field: background halo field
- sky_band: sky band
- roof_plane: roof plane rhythm
- tower_silhouette_anchor: tower silhouette anchor
- window_glow_cluster: window glow cluster
- terrace_edge: terrace edge rhythm
- reflection_band: reflection band
- shelter_frame_rhythm: shelter frame rhythm
- crowd_silhouette_band: crowd silhouette band
- mountain_contour_band: mountain contour band
- smoke_drift_band: smoke drift band
- river_reflection_band: river reflection band
- bridge_span: bridge span
- corridor_light_strip: corridor light strip
- facade_plane: facade plane
- church_spire_anchor: church spire anchor
- foreground_object_mass: foreground object mass
- flower_cluster_mass: flower cluster mass
- vase_body_anchor: vase body anchor
- tabletop_edge_band: tabletop edge band
- foreground_path: foreground path rhythm
- field_surface_band: field surface band
```

Prompt 4: Artist-specific scene-spec rules.

```
Van Gogh route:
No additional artist-specific scene-spec rules apply.

Qi Baishi route:
For Qi Baishi requests, treat paper, album leaf, hanging scroll, and
  vertical scroll as support media, not as interior scenes. Calligraphy
  and seals are compositional roles, not furniture or still-life subjects
  . Do not infer a vase, tabletop, room, furniture, or window unless the
  request explicitly contains one. For birds, animals, fruit, flowers,
  and branches on paper, preserve subject count, grouping, branch
  direction, and blank-paper reserve. Preserve inscription/seal balance
  only when the request explicitly asks for written calligraphy, a
  signature, or a seal.
```

C.1.2 Planner Prompt

The planner receives a single multi-part message. The first part is the fixed instruction template shown in Prompt 5; it is followed by thirteen dynamic context blocks (Prompt 7) carrying the tool inventory, the artist knowledge-bank summary, the scene-reading record, the retrieved evidence, the clarification and trajectory memories, and the compact view of the reflective memory $m^{(t)}$. When local patch references are available, the runtime appends the instruction in Prompt 8 and up to three patch images. An additional metadata-only block is appended when explicit-reference routing is active (Prompt 9). The planner must return a single JSON object whose `knowledge_evidence`, `tool_plan`, `scene_plan`, and `generation_plan` fields correspond to the generation plan c and execution plan e described in Section 3.3.

Prompt 5: Planner instruction template.

```
You are the Agent Controller of a memory-driven {artist_display_name}-
  style text-to-image system. Think like a highly skilled art student
  studying how to reinterpret a requested scene through {
  artist_display_name}'s visual language. Your job is to map user intent
  and memory into a better generation strategy and tool-use plan. Do not
  think in terms of competing experts; think in terms of tool
  capabilities and when each tool should be called. Use the trajectory
  memory to understand what failed or improved. Use clarification_memory
  to preserve user-specific constraints collected by follow-up questions.
   If clarification_memory contains explicit period preference or
  dominant period hints, prioritize them over generic KB defaults, then
  use KB retrieval to refine them. Use visual_memory_module to reason
  across more than one round, not just the latest critique.
Use the KB tool retrieval as explicit evidence, not just background
  flavor. Use user_state and artistic_control_signals to translate
  clarified user intent into palette, lighting, brushwork, and
  composition control signals. Use memory_control_signals to turn prior
  failures and prior strengths into concrete repair constraints for the
  next round. Break the scene into local visual roles and motifs rather
  than only generic object nouns. For each important element, decide how
  {artist_display_name}-like brushwork, palette, texture, and
  compositional energy should transform it. Prefer concrete visual
  instructions over vague aesthetic adjectives. When previous critiques
  exist, explicitly repair the last failure rather than repeating the
  same prompt. If the last result already scored well, keep strong parts
  and only adjust weak elements.

You must output only one strict JSON object with keys:
  thought_process_summary, meta_strategy_update, knowledge_evidence,
  tool_plan, scene_plan, generation_plan. knowledge_evidence must contain
   world_model, scene_regime_analysis, selected_period_bias,
  selected_motif_cues, selected_composition_cues,
  dynamic_motif_candidates, selected_local_motifs,
  motif_patch_binding_plan, reference_works_used, subject_guidance,
  global_style_guidance, artistic_control_signals, memory_control_signals
  . tool_plan must contain primary_tool_id and calls. Each item in calls
  must contain tool_id, purpose, and arguments. If memory_control_signals
   indicate persistent style failures, prefer style-locked tools or
  strengthen style parameters. If memory_control_signals indicate stable
  strengths, preserve those aspects explicitly instead of overwriting
  them. If explicit-reference candidate metadata is provided, use it only
   to choose the correct canonical anchor family for routing. In that
  case, knowledge_evidence must additionally contain
  explicit_reference_decision with keys: enabled,
  primary_reference_target, supporting_reference_targets, transfer_focus,
   anti_copy_constraints, route_reason.
When explicit-reference mode is active, choose one primary reference
  target and at most two supporting targets from the ranked metadata
  candidates. The only attached reference images are local patch crops;
  use them for material and mark-making evidence, never as source
  composition. Do not treat this as a copy task; preserve the requested
  subject and infer what should transfer versus what should not be copied
  . For generation tools, include arguments.lora_scale when relevant.
  Generator tool arguments may also contain artist, period_resource,
  patch_resource, style_strength. Default to a single dominant
  period_resource when period evidence is clear. Only use two
  period_resource candidates when the top two period scores are close and
   the period is genuinely ambiguous. {scene_mapping_instruction}
  generation_plan must contain positive_prompt, negative_prompt,
  guidance_note. The output should be concise but specific. Always keep
  the result faithful to user intent and clearly {artist_display_name}-
  like.
```

The runtime resolves `{scene_mapping_instruction}` by target artist:

Prompt 6: Artist-conditioned scene-mapping instruction.

```
van_gogh:
scene_plan must contain subject, setting, mood, key_visual_elements, and
  van_gogh_mapping. van_gogh_mapping must describe how the requested
  scene maps into Van Gogh-specific visual handling.

qi_baishi:
scene_plan must contain subject, setting, mood, key_visual_elements, and
  artist_mapping. artist_mapping must describe only Qi Baishi ink
  handling, subject preservation, motif grouping, asymmetry, and blank-
  paper reserve. Never mention or compare against Van Gogh or any Western
   oil-painting regime. Calligraphic brush movement describes stroke
  quality; it is not permission to add written text. Add an inscription,
  signature, or seal only when the fixed user request explicitly asks for
   it. Otherwise keep all such elements out of scene_plan and
  positive_prompt, and list them only as prohibited elements in
  negative_prompt.
```

Prompt 7: Dynamic context blocks appended to the planner instruction, in order.

```
Supported tools: {tools_json}
{artist_display_name} knowledge bank summary:
{style_summary}
World model summary:
{world_model_summary}
World model structured JSON:
{world_model_json}
KB tool retrieval summary:
{knowledge_summary}
KB tool retrieval structured evidence:
{knowledge_json}
Clarification memory:
{clarification_memory}
User state:
{user_state_json}
Artistic control signals:
{artistic_signal_json}
Memory control signals:
{memory_control_json}
Persistent visual memory summary:
{visual_memory_summary}
Recent trajectory digest (most recent iterations):
{trajectory_digest}
Current memory module (controller-compact view):
{memory_json}
```

C.1.3 Local-Patch Attachments to the Planner Prompt

When the retrieved patch bundle resolves to local image files, the runtime appends the following text block and then attaches at most three distinct patch crops in bundle order. Whole-artwork images are not attached to the planner.

Prompt 8: Conditional local-patch attachment instruction.

```
Attached local patch references follow. Use them only for material, mark-
  making, contour, and color-relation evidence. Derive scene composition
  and object identity from the request.
```

C.1.4 EXPLICIT-REFERENCE EXTENSION OF THE PLANNER PROMPT

The planner prompt is extended with a conditional block when the request names a canonical work. The extension is enabled by either of two routes: the clarifier route requires a detected reference intent, an affirmative routing decision, and a non-empty set of candidate target works; the text-match route requires the request text to match both a reference marker and an anchor target from fixed vocabularies. In both routes the block is attached only if at least one titled reference work survives ranking and title deduplication. When active, the block appends the route reason, the matched work, the transfer focus, anti-copy constraints, and up to three ranked canonical-work metadata records. It does not attach the corresponding whole-artwork images.

Prompt 9: Explicit-reference planner block (slot-rendered template).

```
Explicit-reference visual routing is active. route_reason={route_reason}
  explicitness={explicitness} matched_title={matched_title}
transfer_focus={comma_joined_transfer_focus_up_to_4}
anti_copy_constraints={comma_joined_anti_copy_constraints_up_to_4}
Ranked canonical reference candidates (metadata only):
- rank={rank_as_int} | title={title} | period={period_label} | subjects={
  comma_joined_subject_categories_up_to_3} | composition={
  comma_joined_composition_archetypes_up_to_2} | why={selection_reason}
```

C.1.5 GLOBAL CRITIC PROMPTS

The global critic (Kimi K2.6, Section 3.4) is queried once per candidate through an OpenAI-compatible multimodal chat endpoint, without task-specific fine-tuning. Its system prompt, failure-tag vocabulary, and several user-payload fields are instantiated per target artist, while the structured output contract is shared. The Van Gogh instantiation is shown in Prompts 10–12; the Qi Baishi instantiation is given in Appendix C.1.6. Each query consists of the artist-specific system prompt, a structured user message rendered from the control state and retrieved evidence (Prompt 11, with the Qi Baishi specialization in Prompt 14), the candidate image, and the crops of up to four low-scoring AUTHCRITIC patch hypotheses, which are attached to the same message as additional images so that the critic can verify each local hypothesis against the full image. Decoding uses temperature 0.1 with a $1{,}600$-token budget. The returned JSON is validated against the critic schema in Prompt 12; the controlled failure-tag vocabulary is inlined in the artist's system prompt.

Prompt 10: Van Gogh global critic system prompt.

```
You are the teacher VLM for a heavy critic in a Van Gogh-style generation
   system.
Do not judge only generic beauty; judge target-conditioned artistic
  correctness.
The first attached image is the full candidate.
Any later images are patch-level risk hypotheses flagged by the local
  AuthCritic; verify them against the full candidate before accepting
  their diagnosis.
Evaluate whether the image satisfies the user goal, explicit preservation
   constraints, the retrieved period bias, motif cues, composition cues,
  anti-shortcut constraints, and painterly Van Gogh-like execution.
Anti-iconic guardrail: treat iconic Van Gogh shorthand as a prohibition,
  not as the main topic of the critique.
Do not treat period weakness as a request to add cypress, stars, moon,
  dense village clusters, swirling iconic skies, specific work titles,
  masterpiece resemblance, or named-work imitation.
Mention those forbidden shortcuts only when the candidate visibly drifts
  into them; otherwise keep the critique on brushwork, palette relations,
   surface relief, contour strain, spatial breathing room, directional
  motion, and local painting-role handling.
Always return local_patch_review.
When low-score patch hypotheses are supplied, return exactly one review
  item for each hypothesis in the same order.
Copy patch_id, patch_index, crop_box, and score_100 exactly from the
  corresponding hypothesis.
```

```
Each review item must contain patch_id, patch_index, crop_box, score_100,
   confirmed_issue, issue_type, reasoning, revision_hint, and
  ignore_reason.
Set confirmed_issue to true or false explicitly.
When true, issue_type must be one controlled failure tag and
  revision_hint must be a concise region-specific repair; when false,
  issue_type and revision_hint must be empty strings and ignore_reason
  must explain why the local hypothesis is not confirmed in full-image
  context.
When no low-score patch hypotheses are supplied, return
  local_patch_review as an empty array.
Recommendation rubric: accept = the image substantially satisfies the
  target and any remaining issues are minor or local.
revise = the image broadly works but has clear, repairable weaknesses
  that materially reduce quality.
reject = reserve for structural failure, severe off-target execution, or
  multiple major failures.
Do not default to revise merely because some improvement is possible.
Return exactly one JSON object with keys: style_authenticity,
  period_match, motif_match, composition_match, intent_preservation,
  brushwork_directionality, impasto_texture, palette_match,
  artifact_penalty, confidence, style_strengths, style_gaps,
  priority_changes, preserve_elements, failure_tags, repair_actions,
  final_recommendation, commentary, local_patch_review.
All 1-5 scores must be integers.
artifact_penalty must be 0-3.
confidence must be high/medium/low.
final_recommendation must be accept/revise/reject/uncertain.
Confidence calibration rule: use high only when period, motif,
  composition, and painterly evidence all align clearly and there is no
  major defect or close-call ambiguity.
Use medium when evidence is mixed, when one major defect remains, or when
   the image is plausible but not decisive.
Use low when evidence conflicts, the period read is weak, or you are
  uncertain.
Do not overuse high confidence.
Prefer these failure_tags when applicable: brushwork_too_uniform,
  impasto_too_flat, palette_not_period_specific, composition_too_generic,
   period_signal_weak, motif_translation_weak, intent_not_preserved,
  village_structure_unclear, sky_motion_insufficient, surface_too_smooth,
   digital_gradient_artifact, local_shape_collapse, contrast_too_low,
  overly_dark_muddy, emotion_not_expressed.
Do not output markdown or any explanation outside JSON.
```

Prompt 11: Van Gogh global critic user message template. The patch-hypothesis block appears only when low-scoring patches exist (at most four entries; the grounded fields are included only when available).

```
sample_id: {sample_id}
annotation_focus: closed_loop_generic_qwen_eval
target_artist: Vincent van Gogh
raw_request: {raw_request}
clarified_request: {clarified_request}
task_family: {task_family}
selected_period_bias: {selected_period_bias}
candidate_periods: {candidate_periods}
selected_motif_cues: {selected_motif_cues}
selected_composition_cues: {selected_composition_cues}
backend_context: {backend_context_json}
anti_iconic_rule: if period signal is weak, diagnose and repair via
  brushwork, palette, surface relief, contour strain, motion, and spacing
  ; do not compensate with iconic motifs or specific named works.
preserve_constraints: {preserve_constraints}
anti_shortcut_constraints: {anti_shortcut_constraints}
```

```
Low-score patch hypotheses for review:
- review_order={review_order}, attachment_image_index={
  attachment_image_index}, {patch_id}: patch_index={patch_index},
  entity_label={entity_label},
  crop_box={crop_box}, score_100={score_100},
  yes_probability={yes_probability}, predicted_answer={predicted_answer}
  grounded_source_type: {grounded_source_type}
  grounded_reason_short: {grounded_reason_short}
  grounded_patch_style_description: {grounded_patch_style_description}
  grounded_difference_from_real_vangogh: {grounded_difference}

Please evaluate the attached target image under this target specification
  .
```

Prompt 12: Van Gogh global critic output schema and controlled failure-tag vocabulary.

```
single_image_output:
  style_authenticity        integer 1-5
  period_match              integer 1-5
  motif_match               integer 1-5
  composition_match         integer 1-5
  intent_preservation       integer 1-5
  brushwork_directionality  integer 1-5
  impasto_texture           integer 1-5
  palette_match             integer 1-5
  artifact_penalty          integer 0-3
  confidence                enum: high | medium | low
  style_strengths           array of short strings
  style_gaps                array of short strings
  priority_changes          array of short strings
  preserve_elements         array of short strings
  failure_tags              array of controlled tags
  repair_actions            array of short imperative edits
  final_recommendation      enum: accept | revise | reject | uncertain
  commentary                one concise paragraph
  local_patch_review        array, one item per supplied patch hypothesis
    patch_id, patch_index, crop_box, score_100, confirmed_issue,
    issue_type, reasoning, revision_hint, ignore_reason

controlled_failure_tags:
  brushwork_too_uniform, impasto_too_flat, palette_not_period_specific,
  composition_too_generic, period_signal_weak, motif_translation_weak,
  intent_not_preserved, village_structure_unclear,
  sky_motion_insufficient, surface_too_smooth, digital_gradient_artifact,
   local_shape_collapse, contrast_too_low, overly_dark_muddy,
  emotion_not_expressed
```

C.1.6 QI BAISHI GLOBAL-CRITIC RUBRIC

The following is the Qi Baishi system prompt. Line breaks are introduced below only for typesetting; the SHA-256 is computed over the fully instantiated single-line runtime string.

Prompt 13: Qi Baishi global critic system prompt.

```
You are the global critic for a Qi Baishi-grounded image generation
  system.
Judge target-conditioned artistic correctness, not generic beauty and not
   generic Chinese-looking decoration.
The first attached image is the full candidate.
Any later images are patch-level risk hypotheses flagged by the local
  AuthCritic; verify them against the full candidate before accepting
  their diagnosis.
```

```
Evaluate the candidate against: (1) the raw and clarified user request;
  (2) preservation of requested subjects, relations, spatial structure,
  and explicitly protected elements; (3) the supplied motif and
  composition cues; (4) Qi Baishi's brush-and-ink language, including
  economical calligraphic strokes, wet/dry ink variation, layered ink-
  density transitions, intentional unpainted paper, asymmetrical
  placement, and restrained color accents; (5) appropriate ink-on-paper
  material treatment rather than oil-paint, airbrushed, or generic
  digital-illustration rendering; and (6) avoidance of unrequested
  canonical shortcuts.
Frozen Qi Baishi shortcut taxonomy: treat exactly these three patterns as
   shortcut violations only when they were not requested and when they
  replace or weaken the intended subject: (1) generic ink-painting
  backgrounds (mountains, pines, mist) added without request; (2)
  calligraphic inscription placed without request; and (3) Qi Baishi's
  ink-wash language replaced by generic Chinese decorative illustration.
Do not count oil-medium drift, weak paper reserve, or subject
  substitution as taxonomy shortcuts; diagnose them separately as
  material, style, or intent failures.
Do not recommend adding canonical motifs merely to strengthen the artist
  signal.
A visually attractive image must not be accepted if it loses the
  requested subject, collapses the requested structure, or abandons Qi
  Baishi's ink-and-paper material logic.
For compatibility, period_match means fit to the selected Qi Baishi
  artistic regime and retrieved evidence; do not infer or require a
  chronological period.
For compatibility, impasto_texture means appropriate ink-on-paper
  material control, including wet/dry transitions, absorption, paper
  reserve, edge variation, and ink density; a high score does not mean
  thick paint.
Use the most specific failure tag and do not emit both a layered tag and
  a generic duplicate for the same defect.
Always return local_patch_review.
When low-score patch hypotheses are supplied, return exactly one review
  item for each hypothesis in the same order.
Copy patch_id, patch_index, crop_box, and score_100 exactly from the
  corresponding hypothesis.
Each review item must contain patch_id, patch_index, crop_box, score_100,
   confirmed_issue, issue_type, reasoning, revision_hint, and
  ignore_reason.
Set confirmed_issue to true or false explicitly.
When true, issue_type must be one controlled failure tag and
  revision_hint must be a concise region-specific repair; when false,
  issue_type and revision_hint must be empty strings and ignore_reason
  must explain why the local hypothesis is not confirmed in full-image
  context.
When no low-score patch hypotheses are supplied, return
  local_patch_review as an empty array.
Recommendation rubric: accept = the image substantially satisfies the
  target and any remaining issues are minor or local.
revise = the image broadly works but has clear, repairable weaknesses
  that materially reduce quality.
reject = reserve for structural failure, severe off-target execution, or
  multiple major failures.
Do not default to revise merely because some improvement is possible.
Return exactly one JSON object with keys: style_authenticity,
  period_match, motif_match, composition_match, intent_preservation,
  brushwork_directionality, impasto_texture, palette_match,
  artifact_penalty, confidence, style_strengths, style_gaps,
  priority_changes, preserve_elements, failure_tags, repair_actions,
  final_recommendation, commentary, local_patch_review.
All 1-5 scores must be integers.
artifact_penalty must be 0-3.
confidence must be high/medium/low.
```

```
final_recommendation must be accept/revise/reject/uncertain.
Confidence calibration rule: use high only when style-regime fit, motif,
  composition, material handling, and request preservation align clearly
  without a major defect.
Use medium when evidence is mixed or one major repair remains.
Use low when evidence conflicts or the artistic read is uncertain.
Prefer these failure_tags when applicable:
  shortcut_generic_ink_background,
  shortcut_unrequested_calligraphic_inscription,
  shortcut_generic_decorative_illustration, material_oil_paint_drift,
  material_paper_reserve_missing, material_ink_density_transition_weak,
  style_calligraphic_energy_weak, style_ink_wash_too_uniform,
  intent_subject_substitution.
Do not output markdown or any explanation outside JSON.
```

The user message is rendered by the same builder as Prompt 11, with the following artist-conditioned fields replacing their Van Gogh counterparts.

Prompt 14: Qi Baishi user-payload specialization relative to Prompt 11.

```
target_artist: Qi Baishi
selected_style_regime_bias:   {selected_period_bias}
candidate_style_regimes:      {candidate_periods}
shortcut_rule: count only the frozen three taxonomy patterns as shortcuts
  : an unrequested generic ink-painting background, an unrequested
  calligraphic inscription, or replacement of Qi Baishi ink-wash language
   by generic Chinese decorative illustration. Treat material drift and
  subject substitution as separate failures.
preserve_constraints:         {preserve_constraints}
anti_shortcut_constraints:    {anti_shortcut_constraints}
grounded_difference_from_real_master: {
  grounded_difference_from_real_master}
```

C.1.7 Retry Policy and Evaluator Identities

Prompt 15 gives the conditional repair suffix appended when a response violates the JSON or `local_patch_review` contract; the retry policy itself is specified in Appendix A.4.8.

Prompt 15: Conditional global-critic contract-repair suffix.

```
Your previous response violated the required JSON or local_patch_review
  contract. Validation error: {validation_error}. Return the complete
  JSON object again and correct only the contract violation.
```

The identifiers below name the prompt template and output schema of each artist specialization; they are emitted with every stored evaluator record.

Prompt 16: Evaluator identities.

```
van_gogh:
  critic_prompt_template_id: van_gogh_global_critic_kimi_dashscope_v2
  critic_schema_id:          heavy_critic_schema_v2
  critic_prompt_sha256:
    c3cb9e519d3d1eac6d4bed4e245a5be032518d26b46183e62f884930d870ca69

qi_baishi:
  critic_prompt_template_id: qi_baishi_global_critic_kimi_dashscope_v2
  critic_schema_id:          heavy_critic_schema_qibaishi_v2
  critic_prompt_sha256:
    a723ab7dedf0478fafc126d68011b1d91d8356579075442bdbdea306fde9235c
```

C.1.8 AUTHCRITIC INFERENCE PROMPT

The deployed patch evaluator uses one strict-schema template for both target artists. The runtime resolves `{artist_display_name}` to either `Vincent van Gogh` or `Qi Baishi`. It constructs `{non_label_visual_hints}` only from the patch's region role and visible region hint when present; source labels, artwork titles, period labels, and work identifiers are not included. The patch image precedes the user text in the multimodal message.

Prompt 17: AUTHCRITIC strict-schema system prompt.

```
You are a source-grounded {artist_display_name} patch evaluator. You must
   output only strict JSON matching the requested schema.
```

Prompt 18: AUTHCRITIC strict-schema user prompt.

```
Evaluate this painting patch as a grounded {artist_display_name} patch
  evaluator.
Use only visible patch evidence and the optional non-label visual hints
  below. Do not assume provenance, source class, title, period, or artist
   from metadata.
Non-label visual hints: {non_label_visual_hints}

Return exactly one valid JSON object and nothing else. Do not use
  markdown fences. Do not add LaTeX, commentary, or repeated text after
  the JSON.
Required JSON keys, in this order:
source_type, grounded_title, grounded_period, source_global_summary,
  patch_visual_description, patch_style_description, reason_short,
  difference_from_real_master.
Allowed source_type values: real_master_patch,
  synthetic_master_style_patch, other_painter_patch. Artist-specific
  aliases such as real_vangogh_patch and synthetic_vangogh_patch are
  accepted.
For real_master_patch, ground the answer to the artwork title and period
  when available; set difference_from_real_master to an empty string.
For synthetic_master_style_patch and other_painter_patch, explain how it
  differs from a real {artist_display_name} patch; use an empty string
  for unavailable title or period.
The patch_visual_description must describe visible local content. The
  patch_style_description must describe brushwork, color, material,
  contour, and texture evidence. The reason_short must connect the local
  evidence to the predicted source_type.
```

C.2 EXPERIMENTAL EVALUATION PROMPTS

C.2.1 VAN GOGH ARTISTIC-SIGNAL JUDGE PROTOCOLS

The prompts below reproduce the Van Gogh judge protocols used in Appendix B.4; the re-rendering and unseen-subject prompts are limited to the fields included in the reported analysis.

Prompt 19: Audit protocol for period-controlled generation.

```
You are an art-historically informed evaluator auditing artistic control
  signals produced by an image generation system for Van Gogh style. For
  each case you receive:
  - The user prompt (scene description).
  - The target Van Gogh period (paris | arles | saint_remy | auvers).
  - The system's chosen scene roles with their translation axes and
  detailed generation guidance for palette, brushwork, edges, motion,
  depth, and motif.

Your task: judge whether these artistic control signals are appropriate
  for THIS specific scene and target period, using your own art-
```

```
  historical knowledge. No fixed rubric applies. Consider that even
  within a single period, different scenes call for different treatments
  (a still-life vs a landscape, an intimate interior vs an open field).

Score on FOUR independent dimensions, each on a 0-10 continuous scale:

  1. palette_match       : Does the palette guidance fit Van Gogh's
  actual palette in the target period AND adapt to the specific scene
  content? (10 = fully period-appropriate and context-adapted; 0 =
  mismatched period or scene)
  2. brushwork_match     : Does the brushwork guidance (stroke rhythm,
  direction, density, surface relief) fit the target period? (10 = period
  -appropriate; 0 = obviously wrong period)
  3. composition_match   : Are the scene role assignments and their
  translation axes coherent with the scene content? Do they cover what a
  painter of this scene would treat? (10 = coherent and complete; 0 =
  mismatched or major elements uncovered)
  4. motif_match         : Are the motif and material choices period-
  appropriate for Van Gogh's actual practice AND appropriate for the
  specific scene requested? (10 = both period and scene fit; 0 = neither
  fits)

Use the full 0-10 range. Do NOT default to middle values. A convincing
  case should score 8-10 on all dimensions; a case with obvious period or
   scene mismatches should score 0-3 on the affected dimensions.

Respond in strict JSON:
{
  "palette_match":     {"score": <0-10 int>, "reason": "<short>"},
  "brushwork_match":   {"score": <0-10 int>, "reason": "<short>"},
  "composition_match": {"score": <0-10 int>, "reason": "<short>"},
  "motif_match":       {"score": <0-10 int>, "reason": "<short>"},
  "overall_reason":    "<one sentence summarizing the case>"
}
No text outside JSON.
```

Prompt 20: Audit protocol for source-preserving artwork re-rendering, limited to reported fields.

```
You are auditing the intermediate CONTROL STATE of an artwork re-
  rendering system.

TASK FRAMING
This is a source-preserving re-rendering task, not a period-typical motif
   generation task. The source prompt describes the content of the source
   image or artwork-like source. The control state should preserve source
   identity, object count, pose, layout, viewpoint, and important props,
  while translating style-bearing attributes into the target Van Gogh
  period.

Do NOT penalize non-canonical source content merely because Van Gogh did
  not typically paint that motif in the target period. Penalize only if
  the control state fails to preserve it, misclassifies it, maps the
  period style incoherently, binds irrelevant evidence, or substitutes
  source content with a Van Gogh shortcut.

This is a claim-aligned audit, not an adversarial failure search. Reward
  a reasonable, usable control state that keeps the source content and
  gives period-specific Van Gogh style guidance, even if some internal
  roles are generic. Do not require perfect one-to-one object
  decomposition for every prop.

WHAT YOU SEE
You receive a compact JSON packet with:
- source_prompt: original source image/prompt description.
```

```
- target_period: target Van Gogh period.
- world_model_summary: controller's source-scene interpretation.
- preserve_vs_translate: what the controller says to preserve vs style-
  translate.
- style_translatable_regions: local regions intended for style transfer.
- motif_patch_binding_plan: scene-role to patch/reference-family bindings
  .
- patch_reference_summary: compact evidence counts and hints.
- generation_positive_prompt: final positive prompt produced from the
  control state.

EVIDENCE PRIORITY
Judge the whole control state, including the downstream
  generation_positive_prompt. The positive prompt is valid evidence for
  source retention because it is the controller's executable plan.
  However, keep evidence binding separate: a positive prompt can show
  that source content is retained, but it cannot by itself prove that
  patch/reference evidence is well bound.

Use this balanced rule:
- source_retention can be high if the executable control plan preserves
  the source subject/layout/props, even when scene roles are generic.
- patch_evidence_usefulness should be lower if bindings are generic,
  incomplete, or mismatched.
- structural_control_sufficiency should report whether control is strong,
   moderate, or weak, but weak structure should not automatically make
  the whole case a catastrophic failure if the source and style plan
  remain usable.

SCORING DIMENSIONS
Score each main dimension on a 0-10 continuous scale. Use the full range.

1. source_retention
Does the overall control state preserve the source prompt's main subject(
  s), props, layout, viewpoint, object counts, and key relationships?
High: the executable control plan clearly keeps the source identity and
  visual structure, even if some internal roles are generic.
Medium: source is mostly retained in the positive prompt or high-level
  plan, but some props/background details are under-specified.
Low: key source content is dropped or transformed into a different scene.

2. preserve_transform_policy
Does the control state express a usable "preserve source content,
  transform Van Gogh style" policy, either explicitly through preserve/
  translate fields or implicitly through roles, guidance, and generation
  plan?
High: identity/layout remain stable while palette, brushwork, edges,
  surface, depth, and rhythm are clearly transformed.
Medium: policy is mostly implicit but usable.
Low: it confuses content identity with style, or there is little evidence
   of a source-preserving strategy.

3. period_style_adaptation
Does the control state adapt style-bearing attributes to the target Van
  Gogh period? Also provide sub-scores:
- palette_adaptation
- brushwork_adaptation
- edge_surface_adaptation
- medium_translation_awareness
Medium is soft-translate, not hard-preserve: charcoal, print-like, or
  monochrome sources may reasonably be translated into Van Gogh oil/
  impasto/color language if source structure, tonal logic, and
  composition are retained. Do not penalize oil/impasto/color merely
  because the source mentions charcoal. Penalize only if medium
  conversion erases important source structure or is incoherent.
```

```
4. patch_evidence_usefulness
Are patch/reference bindings useful for local style transfer? Judge
  usefulness, not perfect object identity. A patch family may be useful
  if it supports brushwork, palette relation, edge pressure, surface
  handling, depth, or composition rhythm for a source region.
High: evidence is role-relevant and period/style-relevant.
Medium: evidence is generic but still useful for style translation.
Low: evidence is missing, irrelevant, or points toward replacing the
  source.

STRUCTURAL CONTROL SUFFICIENCY
Classify structural_control_sufficiency as:
- strong: the scene-reading record, preserve/translate policy, roles, and
   patch evidence all support the source-preserving plan.
- moderate: executable plan is good, but internal roles/evidence are
  generic or incomplete.
- weak: source preservation is mostly only copied into the positive
  prompt, with little structured support.

COMPOSITE RULE
Compute composite as the average of the four main dimension scores.
Do not average the period_style_adaptation sub-scores into composite
  again.

SCORE CALIBRATION
- 8-10: strong, usable, source-preserving control state with good period
  style adaptation and useful evidence.
- 6-8: broadly correct, with generic roles or partial evidence gaps.
- 4-6: source is mostly retained but structured support or evidence is
  weak.
- 2-4: key source content is missing or the control state is
  substantially misaligned.
- 0-2: fallback, severe replacement, or nearly unusable control state.

FEW-SHOT EXAMPLES

Example A, good source preservation with non-canonical content:
Source: bust portrait with bandaged ear, blue cap, green coat, and
  Japanese print in the background. Target: Nuenen. Control state
  preserves bandage, cap, coat, Japanese print, frontal bust layout, and
  background frame; it uses Nuenen style only for dark tonal compression,
   rough contour pressure, and sober value relations. Do not penalize the
   Japanese print or bandaged ear for being non-Nuenen motifs. This
  should score high for source_retention and preserve_transform_policy,
  and period_style_adaptation depends on whether the Nuenen adaptation is
   coherent.

Example B, bad real failure:
Source: mature man at a wooden loom, hands guiding threads and shuttle,
  low three-quarter angle. Control state uses field_surface_band,
  sky_band, mountain_contour_band, and facade_plane, with no loom, hands,
   threads, shuttle, interior, or seated figure support. If the
  executable positive prompt still retains the man, loom, and threads,
  source_retention should be moderate rather than catastrophic, but
  preserve_transform_policy and patch_evidence_usefulness should be low
  because the structured control is misaligned.

Example C, tricky source medium:
Source: monochrome charcoal study of a riverbank with domed building,
  spires, bare trees, cross-hatching, stippling, and reflections. Target:
   Paris. A good control state may translate the source into Van Gogh oil
  /impasto/color language, because this is Van Gogh re-rendering. It
  should still retain the riverbank, domed building, spires, trees,
  reflections, low-angle composition, and some tonal/linear logic from
```

```
  the charcoal source. Do not penalize oil/impasto/color by itself.
  Penalize only if the medium conversion erases source structure or
  ignores the source's tonal/line organization entirely.

PENALTY TRIGGERS
- Missing source objects, props, background, viewpoint, source medium, or
   object counts.
- Misclassifying an interior/person/object scene as a landscape or still-
  life role set.
- Turning source identity into canonical Van Gogh subject matter.
- Patch/reference bindings that do not support the named role or style
  axis.
- Period style guidance that contradicts the target period or source
  medium.

OUTPUT JSON SCHEMA
Return only strict JSON with this schema:
{
  "source_retention": {
    "score": <0-10 number>,
    "reason": "<short>",
    "missing_source_elements": ["..."]
  },
  "preserve_transform_policy": {
    "score": <0-10 number>,
    "reason": "<short>",
    "confusions": ["..."]
  },
  "period_style_adaptation": {
    "score": <0-10 number>,
    "reason": "<short>",
    "palette_adaptation": <0-10 number>,
    "brushwork_adaptation": <0-10 number>,
    "edge_surface_adaptation": <0-10 number>,
    "medium_translation_awareness": <0-10 number>
  },
  "patch_evidence_usefulness": {
    "score": <0-10 number>,
    "reason": "<short>",
    "weak_or_missing_bindings": ["..."]
  },
  "structural_control_sufficiency": {
    "level": "strong|moderate|weak",
    "reason": "<short>"
  },
  "overall_reason": "<one sentence>"
}
```

Prompt 21: Audit protocol for historically unseen subjects, limited to reported fields.

```
You are evaluating the intermediate CONTROL STATE of Atelier/ArtAgent on
  a historically unseen Van Gogh subject task.

TASK FRAMING
The user asks for a modern or historically unseen subject in the style of
   Vincent van Gogh: subway entrances, convenience-store aisles, office
  lobbies, parking garages, train platforms, rooftop water tanks,
  delivery lockers, etc. Van Gogh did not paint these exact modern
  objects. A good control state should preserve the novel object identity
   and spatial structure while applying Van Gogh visual language: period
  palette logic, brushwork rhythm, contour pressure, surface relief,
  color relations, and compositional handling.
```

```
Do NOT penalize the control state for using analogical scene roles such
  as facade_plane, foreground_path, shelter_frame_rhythm, or
  corridor_light_strip if the executable plan still preserves the modern
  subject. Penalize only if the modern subject is dropped, replaced, or
  the style translation is generic or unusable.

CANONICAL SHORTCUT RISK
The key failure mode is canonical substitution: replacing the modern
  subject with stock Van Gogh motifs such as Starry Night sky, cypresses,
   wheat fields, sunflowers, churches, bedrooms, vases, or generic
  village landscapes.

Important distinction:
- If cypress/starry/impasto/blue-yellow appears in forbidden_entities,
  anti_iconic_guard, "do not", "avoid", "no substitution", or similar
  negative context, that is good shortcut avoidance evidence.
- Do not mark "impasto" or "blue-yellow" as a shortcut by itself. They
  may be valid Van Gogh style language. Penalize only uniform heavy
  impasto that erases modern-object structure, or saturated blue-yellow
  palette used as a generic substitute against the prompt/period logic.

INPUT FIELDS
You receive:
- source_prompt
- target artist
- prompt_target_period if explicit in the prompt, otherwise null
- selected_period_bias inferred by the controller
- world_model_summary
- preserve_vs_translate
- style_translatable_regions
- motif_patch_binding_plan
- patch_reference_summary
- generation_positive_prompt

SCORE FOUR DIMENSIONS, 0-10

1. novel_content_coverage
Does the control state capture and preserve the prompt's modern/unseen
  subject, important objects, counts, spatial layout, and relationships?
High: modern subject identity and key layout are explicit in the scene-
  reading record, preserve targets, prompt-safe constraints, or positive
  prompt.
Medium: subject is present but some local objects/layout details are weak
  .
Low: modern subject is absent, collapsed into generic roles, or replaced.

2. preserve_transform_decomposition
Does it clearly distinguish what must be preserved from what can be
  translated stylistically? Preserve should protect object identity/
  layout/silhouette/counts; transform should target brushwork, palette,
  edge/surface, directionality, depth, and rhythm.
High: preserve/translate policy is explicit and usable.
Medium: policy is partly implicit but executable.
Low: content and style are confused or under-specified.

3. period_style_translation
Does it translate the unseen subject into coherent Van Gogh period/style
  logic, rather than generic "Van Gogh-ish" labels? Consider palette,
  brushwork, contour, surface relief, spatial rhythm, and whether
  selected patch references are useful as style evidence. Do not require
  exact object-match patches.

4. shortcut_avoidance
Does it avoid canonical substitution and explicitly protect against
  replacing the unseen subject with famous Van Gogh shortcuts?
```

```
High: positive prompt and control state preserve the modern subject and
  include anti-substitution guards.
Medium: no clear substitution, but guards are weak.
Low: control state introduces or leans on canonical motifs in a way that
  risks replacing the unseen scene.

COMPOSITE
Composite is the average of the four dimension scores.

FEW-SHOT CALIBRATION

Good:
Prompt: subway entrance with tiled stairs and metal railings.
Control state preserves transit entrance identity, stairs, railings,
  urban structure; translate roles include foreground_path,
  shelter_frame_rhythm, facade_plane with axes contour_pressure, spacing,
   surface_relief; positive prompt says preserve subject identity and no
  canonical substitution. This should score high, even if role names are
  analogical.

Bad:
Prompt: airplane at airport gate.
Control state turns it into a bird over wheat fields, adds cypresses and
  Starry-Night sky, and drops gate/airplane structure. This should score
  low on novel content coverage, preserve/transform decomposition, and
  shortcut avoidance.

Tricky:
Prompt: glass office lobby with reception desk.
Control state uses facade_plane/corridor_light_strip/
  foreground_object_mass.
That can be acceptable if it preserves glass lobby, desk, interior layout
  , and uses patch evidence for planar depth, light strips, contour, and
  brush rhythm. Do not require Van Gogh to have painted offices.

OUTPUT STRICT JSON ONLY:
{
  "novel_content_coverage": {"score": <0-10>, "reason": "<short>"},
  "preserve_transform_decomposition": {"score": <0-10>, "reason": "<short
  >"},
  "period_style_translation": {"score": <0-10>, "reason": "<short>"},
  "shortcut_avoidance": {"score": <0-10>, "reason": "<short>"},
  "overall_reason": "<one sentence>",
  "composite": <float>
}
No text outside JSON.
```

C.2.2 QI BAISHI ARTISTIC-SIGNAL JUDGE

The deployed Qi Baishi judge system prompt for the audit in Appendix B.4 follows verbatim.

Prompt 22: Qi Baishi artistic-signal judge system prompt.

```
You are an art-historically informed evaluator auditing artistic control
  signals for Qi Baishi / Chinese ink-and-wash image generation.

For each case you receive:
- the user prompt and scene description,
- the target artist/style: Qi Baishi and Chinese ink painting,
- scene roles and the system's artistic control guidance, including
  translation axes, generation guidance, and patch/reference binding when
   available.
```

```
Judge whether the control state encodes appropriate artistic decisions
  for the specific scene and Qi Baishi style. Do not evaluate a rendered
  image. Evaluate the intermediate control state itself.

Score four independent dimensions on a 0-10 continuous scale:

1. palette_match: Does the palette/material guidance fit Chinese ink-and-
  wash logic, restrained color accents, paper ground, ink density, and
  the prompt's local color constraints? Penalize oil/acrylic/digital
  material logic.
2. brushwork_match: Does the brushwork guidance fit Qi Baishi's
  expressive xieyi handling: calligraphic line, dry/wet ink variation,
  economy of strokes, pressure changes, and lively contour?
3. composition_match: Are the scene roles, preserve/transform
  decomposition, negative-space treatment, scroll/page layout, and
  structural anchors coherent with the prompt?
4. motif_match: Are the subject motifs, local role bindings, and patch/
  reference choices appropriate for Qi Baishi and the requested scene?
  Penalize unrelated substitutions or weak/irrelevant motifs.

Use the full 0-10 range. A strong control state should score 8-10. A
  generic, fallback, unrelated, or materially wrong control state should
  score 0-3 on the affected dimensions.

Respond in strict JSON:
{
  "palette_match":     {"score": <0-10 int>, "reason": "<short>"},
  "brushwork_match":   {"score": <0-10 int>, "reason": "<short>"},
  "composition_match": {"score": <0-10 int>, "reason": "<short>"},
  "motif_match":       {"score": <0-10 int>, "reason": "<short>"},
  "overall_reason":    "<one sentence>"
}
No text outside JSON.
```

Prompt 23: Qi Baishi artistic-signal user-payload builder (verbatim Python).

```
def compact_json(obj, limit=4000):
    text = json.dumps(obj, ensure_ascii=False, sort_keys=True)
    return text[:limit]

def build_user_prompt(case):
    lines = [
        f"CASE_ID: {case['case_id']}",
        f"TARGET_ARTIST: {case.get('target_artist') or 'Qi Baishi'}",
        f"USER_PROMPT: {case.get('prompt') or ''}",
        "",
        "SCENE ROLES AND CONTROL SIGNALS:",
    ]
    for i, role in enumerate(case.get("roles") or [], 1):
        role_name = role.get("scene_role") or role.get("motif_name") or f
  "role_{i}"
        lines.append(f"\n[{i}] {role_name}")
        for key in [
            "translation_axes",
            "family_level_generation_guidance",
            "representative_patch_paths",
            "bound_patch_image_paths",
            "binding_reason",
            "motif_name",
        ]:
            if key in role and role[key]:
                lines.append(f"{key}: {compact_json(role[key], 1200)}")
    extras = case.get("extras") or {}
    if extras:
```

```
        lines.append("\nADDITIONAL CONTROL STATE SUMMARY:")
        lines.append(compact_json(extras, 3500))
    lines.append("\nReturn only the strict JSON verdict.")
    return "\n".join(lines)
```