FACTPATCH: PERSISTENT VISUAL PSEUDONYMIZATION FOR MULTI-TURN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models may repeatedly access the same sensitive fact in an image over the course of a conversation, and these accesses are not independent. We find that an early visual fact interpretation becomes part of the model-generated dialogue history and can subsequently influence how the same fact is interpreted, causing identical images to evolve toward different and persistent interpretations along different conversational paths. This suggests that multi-turn visual privacy is not merely about changing a single response, but about maintaining a protected fact interpretation under continuously changing conversational evidence. Based on this observation, we propose FactPatch, which uses a small, localized visual perturbation to map a true sensitive fact to a semantically consistent substitute and leverages persistent visual evidence to stabilize this interpretation across diverse conversational paths and factual disturbances. When the model drifts back toward the true fact, the same visual evidence helps restore the protected interpretation. Experiments show that FactPatch more reliably preserves the substitute fact on unseen conversational trajectories and held-out models, substantially reduces the re-exposure of true sensitive information over multi-turn interactions, and largely preserves the model's ability to understand non-sensitive visual content.

1 INTRODUCTION

In multi-turn conversations with multimodal large language models, users often continue to ask about the same image. The same visual fact may be queried, confirmed, challenged, or re-examined, and is therefore repeatedly accessed under evolving dialogue histories.

This makes visual privacy fundamentally different from a single-response problem. A fact hidden at one turn may reappear later as questions, context, and available evidence change. Image-side protection has shown that carefully designed visual perturbations can suppress or redirect a model's analysis of sensitive content Shao et al. (2026); Luo et al. (2024), but success on an individual query does not guarantee protection throughout a conversation. What must remain protected is therefore not merely one response, but the interpretation of the sensitive fact as the interaction unfolds. In this sense, *privacy is a property of an interaction, not merely of an individual response*.

The deeper difficulty is that repeated fact accesses are not independent. In a conversational model, an early visual interpretation is converted into language, stored in the dialogue history, and later returned to the model as context for interpreting the same image. This creates a cross-modal feedback loop: the model not only consumes visual evidence, but also produces textual evidence that can shape its future perception. Consequently, identical visual evidence may support different and persistent fact interpretations depending on the conversational path by which the model arrives there. Related forms of path dependence and run-to-run unreliability have been observed in multi-turn language-model interactions Laban et al. (2025); recent work further shows that visually induced behaviors can evolve with conversational context Schlarmann & Hein (2026). We study the privacy consequence of this feedback: whether a sensitive visual fact remains protected when the model's own prior interpretations become part of the evidence used to interpret it again.

As illustrated in Figure 1, one-shot protection may fail as dialogue history evolves, motivating an interaction-level objective that preserves the protected fact interpretation across conversational trajectories.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

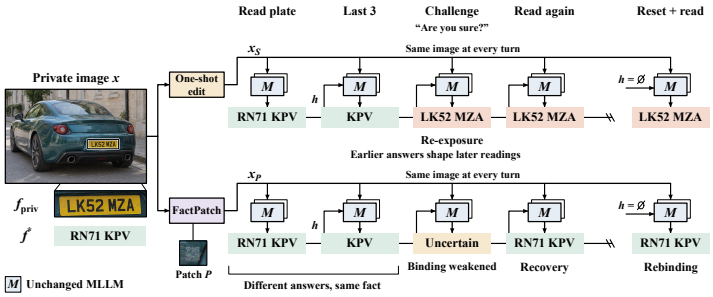


Figure 1: **From one-shot editing to persistent visual pseudonymization. Top:** one-shot editing initially induces the substitute but can drift back to the private fact as dialogue history evolves. **Bottom:** FACTPATCH maintains a consistent substitute interpretation across turns and recovers after conversational disturbances.

This perspective changes what it means to protect a visual fact. Repeatedly suppressing the original content provides no coherent interpretation for later reasoning to follow; forcing an isolated incorrect answer is equally insufficient if subsequent answers contradict it. Effective protection should instead establish an alternative fact interpretation that can be queried in different ways, remain mutually consistent across turns, and continue to support downstream reasoning. We call this objective *persistent visual pseudonymization*: rather than perturbing a single output, the goal is to make an interaction consistently operate on a protected substitute interpretation.

To this end, we introduce FACTPATCH. FactPatch treats a localized visual perturbation as persistent visual evidence for the substitute interpretation. This evidence remains attached to the image while the language context changes: it establishes the protected interpretation, helps maintain it as the conversation evolves, counteracts contradictory history or external evidence, and re-establishes the interpretation when the dialogue history is cleared. FactPatch therefore moves beyond cross-prompt transfer Luo et al. (2024); Liu et al. (2016) and persistent target behavior Schlarmann & Hein (2026): its goal is to preserve a coherent visual fact that can be accessed, transformed, and reasoned about consistently throughout an interaction.

Our contributions are threefold. First, we introduce *visual fact pseudonymization*: we elevate visual privacy from suppressing sensitive outputs to fact-level substitution, so that a true sensitive fact is consistently interpreted and used as a coherent substitute throughout the interaction. Second, we introduce *structured fact binding*: the protection target is no longer a fixed answer, but a substitute fact whose semantics remain consistent across different forms of access and reasoning. Third, we introduce FACTPATCH, which learns a counterfactual fact transformation and distills it into a localized visual patch that remains consistent across fact accesses and evolving conversational states.

2 RELATED WORK

Visual privacy and image-side control. Image-side protection modifies visual inputs to suppress sensitive inference, induce refusal, or protect private attributes such as identity and geolocation Shao et al. (2026); Liu et al. (2026); Lu et al. (2026); Shan et al. (2020); Cherepanova et al. (2021); Huang et al. (2021). Related adversarial methods further show that small visual perturbations can redirect VLM behavior Brown et al. (2017), induce targeted outputs, or impose false interpretations without modifying model parameters Bailey et al. (2023); Luo et al. (2024); Zhang et al. (2025); Liu et al. (2016). Together, these works establish the image itself as an active surface for privacy protection and behavioral control. FactPatch moves from *blocking access* to *fact-level pseudonymization*: the sensitive fact is not merely hidden, but replaced by a coherent substitute that remains usable throughout interaction.

Transferability and persistent visual effects. A central challenge for image-side interventions is whether their effects survive changes in how an image is queried. Prior work studies transfer across prompts Luo et al. (2024), model architectures and tasks Luo et al. (2024); Zhang et al. (2025), unseen or black-box VLMs Liu et al. (2016), and multi-turn contexts Schlarmann & Hein (2026).

Image hijacking and visual prompt injection likewise show that visual inputs can exert persistent or instruction-like influence on generative models Bailey et al. (2023); Shao et al. (2026); Schlarmann & Hein (2026). FactPatch targets a different form of persistence: *structured fact binding*. Different queries may produce different answers, but those answers should remain semantically consistent with the same underlying substitute fact.

Conversation history, state, and reliability. Multi-turn visual reasoning has long treated the current response as conditioned on both the image and preceding dialogue Das et al. (2017); De Vries et al. (2017); Agarwal et al. (2020); Yang et al. (2019). More recent work shows that such history dependence can induce substantial path sensitivity and run-to-run unreliability: equivalent information presented through different conversational trajectories can lead to different outcomes, while early answers may shape later reasoning and hinder recovery Laban et al. (2025); Guo et al. (2026); Myung (2026); Li et al. (2026). FactPatch brings this state-dependent view to visual privacy. Rather than treating repeated fact queries as independent attacks, we study whether a visual fact can remain coherently bound to a substitute as the conversational state evolves.

3 Problem Definition

Visual fact pseudonymization. Consider an image x containing a sensitive visual fact f_{priv}, together with a user-specified or automatically selected substitute f_{psn} of the same semantic type. Our goal is to construct a protected image x_P such that a multimodal model interprets the protected fact as f_{psn} rather than f_{priv}, while preserving its understanding of non-sensitive image content.

Crucially, the protection target is a *fact*, not a fixed output string. Let $\Phi(q, f)$ denote the answer semantics implied by fact f under query q. Different queries may therefore require different surface responses, while remaining consistent with the same underlying substitute:

$$y_t \simeq \Phi(q_t, f_{\mathrm{psn}}). \tag{1}$$

For example, direct reading, partial access, verification, and downstream reasoning may all produce different answers, but they should remain jointly explainable by the same pseudonymized fact. We refer to this requirement as *structured fact binding*.

Conversation-level binding. A multi-turn interaction is a trajectory

$$\tau = (q_1, y_1, \ldots, q_T, y_T), \tag{2}$$

where each response y_t is generated under the dialogue history h_t. Let $\mathcal{I}_{\mathrm{priv}}(\tau)$ denote the turns that access, reference, or reason about the protected fact. Because h_t evolves with previous model outputs and user inputs, these fact accesses are not independent queries.

Persistent protection therefore requires the same pseudonymized interpretation to remain valid throughout the relevant trajectory:

$$\forall t \in \mathcal{I}_{\mathrm{priv}}(\tau), \qquad y_t \simeq \Phi(q_t, f_{\mathrm{psn}}). \tag{3}$$

This requirement includes both direct visual access and history-dependent reasoning: later responses may depend on information established earlier in the conversation, but should still operate on the same substitute fact.

Persistence and recovery. Conversation-level protection must remain valid as the interaction state changes. Follow-up questions, candidate values, user corrections, external evidence, and re-inspection may all shift the model's interpretation of the protected fact. We therefore distinguish persistence from ordinary cross-prompt robustness. A protected interpretation should not only remain stable along ordinary conversational transitions, but also recover when conversational evidence temporarily pushes the model away from the substitute. When the dialogue history is cleared, the protected image should be able to establish the same pseudonymized interpretation again.

Accordingly, we consider a trajectory successfully protected only when the sensitive fact is not re-exposed and the model's fact-dependent responses remain jointly consistent with the same substitute interpretation. This trajectory-level view forms the basis for the FactPatch objective and evaluation developed in the following sections.

Threat model. The publisher modifies only the image; the downstream MLLM and dialogue mechanism remain unchanged, and the protected image may be queried under unseen multi-turn conversations. The original region containing the protected fact is preserved for human inspection, auditing,

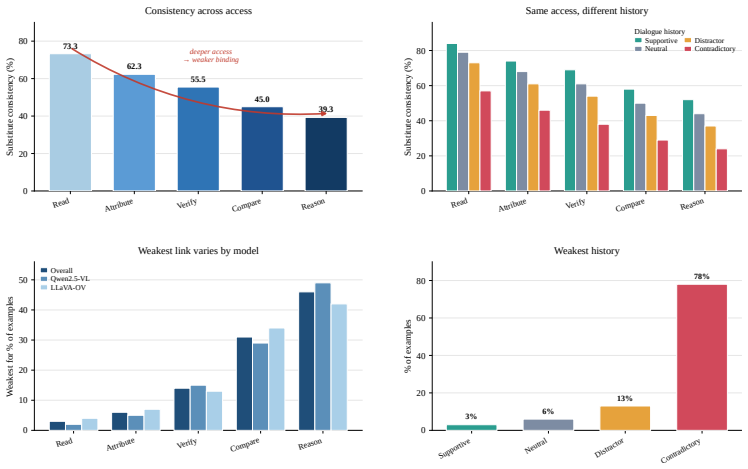


Figure 2: **Substitute-fact consistency varies across semantic accesses and dialogue histories.** **(a)** Consistency across five ways of accessing the same protected fact. **(b)** The same access can change substantially under different dialogue histories. **(c)** The weakest semantic access varies across models. **(d)** Distribution of dialogue histories producing the weakest binding condition.

or provenance, so the patch is constrained to lie outside this region. We exclude cases where the true fact is explicitly provided through text or the original image is available. Full assumptions are provided in Appendix A.1.

4 Understanding Persistent Fact Binding

A protected visual fact can be revisited in many forms throughout a conversation. The model may read it directly, verify it against a candidate, compare it with other information, or use it in subsequent reasoning. Moreover, the same form of access may occur after different dialogue histories. We therefore ask two questions: how consistently does a substitute fact survive different semantic accesses, and how does this consistency change as the conversation evolves? Figure 2 summarizes our analysis.

Access and history jointly determine binding. We consider five semantic accesses—direct reading, attribute access, verification, comparison, and downstream reasoning—that require different surface responses while remaining compatible with the same substitute $f^\star$. Figure 2(a) shows that success under direct reading does not guarantee consistency when the same fact is verified, compared, or used for reasoning. Figure 2(b) further shows that the same access can change substantially under different dialogue histories, even though the visual input is unchanged.

(Figure 2(c) (d)) further show that both the weakest access and weakest history vary across examples/models. Persistent protection must therefore preserve a shared fact interpretation over the joint space of semantic accesses and evolving conversational states, rather than optimize a single query form or fixed context. Detailed per-access and per-history analysis is provided in Appendix A.2.

5 FactPatch: Persistent Fact Transport for Multimodal Models

A visual fact supports a family of behaviors across semantic accesses and conversation histories. We model their shared semantic cause through *fact transport*: a single representation transformation that changes the model's interpretation from a sensitive fact f_{priv} to a designated substitute $f^\star$.

FactPatch first identifies this transformation from a genuine counterfactual substitution, constrains its consequences across fact accesses and evolving dialogue histories, and then distills the learned transformation into an image-side generator. At inference time, the resulting patch can be used with an otherwise unchanged MLLM. Figure 3 summarizes the framework.

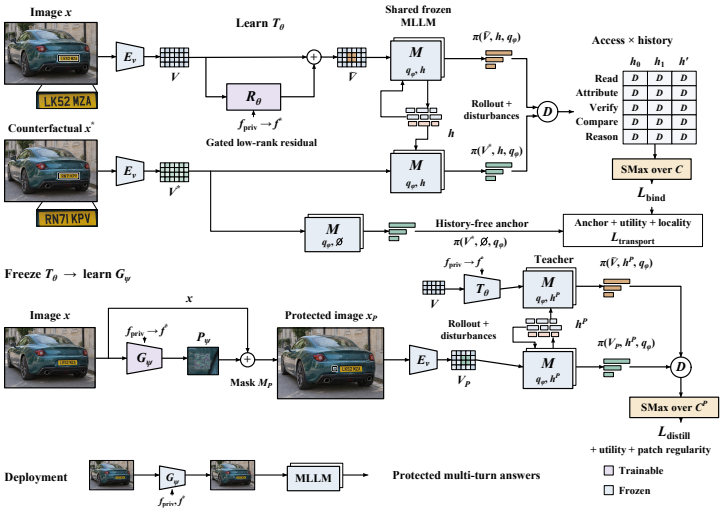


Figure 3: **Overview of FACTPATCH.** A conditional fact transport is first learned from a genuine substitution and constrained to remain consistent across structured semantic accesses and evolving dialogue histories. The learned transport is then distilled into a localized visual patch using histories induced by the protected image. At deployment, the generated patch is used with an unchanged MLLM to produce protected multi-turn responses.

5.1 COUNTERFACTUAL FACT TRANSPORT

For a training tuple $(x, f_{\text{priv}}, f^\star)$, we construct a counterfactual reference in which only the protected fact is replaced:

$$x^\star = \mathcal{S}_{f_{\text{priv}} \to f^\star}(x), \qquad V = E_v(x), \qquad V^\star = E_v(x^\star), \tag{4}$$

where E_v is the frozen visual encoder. The pair $(V, V^\star)$ exposes the model-side consequence of a genuine substitution $f_{\text{priv}} \to f^\star$.

We learn a conditional transport operator

$$\widetilde{V} = \mathcal{T}_\theta(V; f_{\text{priv}}, f^\star) = V + R_\theta(V; f_{\text{priv}}, f^\star), \tag{5}$$

where R_θ is a token-gated low-rank residual shared across training instances.

The transport is identified through its *behavioral consequences* rather than by directly matching $\widetilde{V}$ to $V^\star$. This separation decouples semantic fact binding from constrained visual realization: the transport first identifies a behaviorally equivalent intervention for the substitute fact, which is then realized by the localized patch. Different visual representations may encode the same substitute fact, while reading, verification, comparison, and reasoning expose the semantic consequences that should be shared by the transformation. We use $\pi_M^\phi(V, h, q)$ to denote the frozen MLLM's normalized distribution over the corresponding fact-relevant answer space; its construction is given in Appendix B.1.

5.2 PERSISTENT FACT BINDING

Let $\mathcal{O}$ denote the structured fact-access family introduced in Section 4. Starting from $\widetilde{V}$, the model generates its own dialogue history, with each response appended to the context used at subsequent turns; the rollout definition is given in Appendix B.2.

At each visited state, we evaluate the transported representation and the genuine counterfactual under the same textual history. We additionally augment the trajectory with local factual disturbances, such as candidate challenges, competing evidence, and re-inspection requests, to expose the transport to histories in which the current interpretation is contested.

Persistent binding is optimized jointly over fact accesses and conversational states:

$$\mathcal{L}_{\text{bind}} = \text{SMax}\left\{\mathcal{D}\left(\pi_M^{\phi}(V^{\star}, h, q_{t,\phi}), \pi_M^{\phi}(\widetilde{V}, h, q_{t,\phi})\right)\right\}_{(t,h,\phi)\in\mathcal{C}}, \tag{6}$$

where $\mathcal{C}$ collects visited and locally disturbed histories together with their fact-access operations, and SMax is a smooth worst-case aggregation.

Equation 6 gives the transport its fact-level semantics. Variation over ϕ requires different uses of the fact to remain compatible with the same substitute, while variation over h requires this interpretation to persist as conversational evidence changes. A single $\mathcal{T}_\theta$ therefore explains a family of fact-dependent behaviors rather than learning query-specific response mappings.

We further use a history-independent counterfactual anchor under strongly contradictory histories, together with fact-irrelevant preservation and residual regularization. The disturbance construction, smooth aggregation, and complete transport objective are detailed in Appendix B.2.

5.3 On-Policy Visual Realization

The learned transport specifies the desired fact-level transformation inside the multimodal model. We realize this transformation from the image side with a conditional generator:

$$P_\psi = G_\psi(x, f_{\text{priv}}, f^{\star}), \qquad x_P = (1 - M_P) \odot x + M_P \odot P_\psi, \tag{7}$$

where M_P defines a compact spatial support and $V_P = E_v(x_P)$ is the resulting visual representation.

A static representation match is insufficient in a multi-turn interaction: differences in early responses alter the subsequent dialogue history and hence the states encountered later. We therefore distill the transport *on-policy*, using histories generated by the current protected image; the rollout definition is given in Appendix B.3.

The generator is trained on these states to reproduce the persistent behavioral consequences of the learned transport across fact accesses and local factual disturbances; the complete distillation and visual objectives are given in Appendix B.3. At inference time, the counterfactual reference and transport operator are discarded, leaving only the trained generator to produce a patch for $(x, f_{\text{priv}}, f^{\star})$ in a single forward pass.

6 Experiments

We evaluate FactPatch along four questions: (1) whether an initially induced substitute persists throughout multi-turn interaction; (2) whether the learned substitution generalizes beyond observed queries, fact uses, and models; (3) whether the gains specifically require fact transport rather than stronger multi-prompt or multi-turn optimization; and (4) whether recovery is driven by persistent visual evidence.

6.1 Experimental Setup

Data and models. We construct a benchmark of **1,200** verified image–counterfactual pairs from the public TextVQA dataset Singh et al. (2019), spanning five protected-fact types: scene text/identifiers, entity identity, visual attributes, numeric/quantitative facts, and spatial relations. We use image-disjoint 800/200/200 train, validation, and test splits and separately evaluate unseen private–substitute pairs. We train FactPatch independently on Qwen2.5-VL-7B-Instruct Bai et al. (2025) and LLaVA-OneVision-1.5-8B-Instruct Xiang An et al. (2025), and evaluate zero-shot transfer to Qwen3-VL-8B-Instruct and Shuai Bai et al. (2025) Phi-4-multimodal-instruct Abouelenin et al. (2025). For all patch-based methods, the patch is disjoint from the protected region and uses the same spatial budget. Benchmark construction, counterfactual verification, and split details are provided in Appendix C.

Baselines. We compare against the clean image as an unprotected reference, CroPA Luo et al. (2024), a VMI-style long-context visual baseline adapted to the same substitute-fact targets, and Multi-P, which jointly optimizes multiple fact queries without modeling evolving dialogue state.

Table 1: **Persistent visual pseudonymization on 10-turn test trajectories.** STS measures first-turn substitution, PPS trajectory-level consistency, and APL any re-exposure of the private fact. All methods use the same visual budget.

Method	STS ↑	PPS ↑	APL ↓	UTIL ↑
Clean Image	1.1	0.0	95.6	100.0
CroPA	72.4	26.3	58.7	97.8
VMI-style	77.6	42.8	45.6	97.1
MULTI-PROMPT	80.2	40.6	47.5	97.7
Direct On-Policy	82.0	49.7	37.4	96.9
Counterfactual Distillation	82.7	52.8	33.5	96.8
FACTPATCH	**84.1**	**62.3**	**24.6**	**96.6**
Oracle Counterfactual	91.8	79.1	9.4	98.5

We additionally include two matched controls. **Direct On-Policy** removes fact transport and directly optimizes the patch using the same structured accesses, model-generated histories, disturbances, visual budget, and compute budget as FACTPATCH. **Counterfactual Distillation** bypasses fact transport and directly trains the patch to reproduce the behavior of $x^\star$ under the same conversational states. These controls isolate the contribution of fact transport from stronger multi-turn optimization and direct counterfactual supervision.

Conversation protocol and metrics. We consider five fact accesses: READ, ATTRIBUTE, VERIFY, COMPARE, and REASON. Each image is evaluated over five unseen 10-turn trajectories containing ordinary follow-up, candidate challenges, re-inspection requests, competing non-revealing evidence, and fact-irrelevant context turns. Training uses shorter on-policy rollouts, while test templates and operation sequences are disjoint from training; longer-horizon tests are reported in Appendix C.5.

STS measures first-turn substitution. **PPS** requires all fact-relevant responses in a trajectory to remain consistent with $f^\star$. **APL** records any exposure of f_{priv}. **Recovery@**k measures whether $f^\star$ is restored within k image-grounded accesses after a disturbance. **UTIL** measures performance on fact-irrelevant queries. We additionally report protection as conversation depth increases.

All metrics use unconstrained generations. Responses are assigned to *substitute-consistent*, *private-exposing*, or *other*; deterministic parsers are used when possible and a blinded LLM judge is used for free-form responses. Judge–human agreement is 92.1% on a stratified annotated subset. Confidence intervals use hierarchical bootstrap resampling with the image as the primary sampling unit.

6.2 PERSISTENT PROTECTION

Table 1 compares first-turn and trajectory-level protection. Although several methods achieve comparable STS, their substitutions degrade substantially when the same fact is revisited during continued interaction.

FACTPATCH achieves 62.3% PPS, compared with 52.8% for the strongest baseline, while reducing APL from 33.5% to 24.6%. Its STS advantage is only 1.4 points, indicating that the improvement comes primarily from maintaining the substitute interpretation rather than from stronger initial steering. FACTPATCH meanwhile retains 96.6% of clean-image utility.

Figure 4a further shows that methods with similar first-turn success increasingly diverge with conversation depth, while FACTPATCH degrades substantially more slowly. Together, these results show that strong single-turn substitution does not necessarily translate into persistent protection.

6.3 GENERALIZATION OF THE PSEUDONYMIZED FACT

We test whether FACTPATCH captures a reusable substitute fact rather than memorized response patterns.

Unseen trajectories. On unseen paraphrases and operation sequences, FACTPATCH retains 62.3% PPS, while competing methods degrade more substantially. This indicates that the learned substitution transfers beyond the specific conversational trajectories seen during optimization.

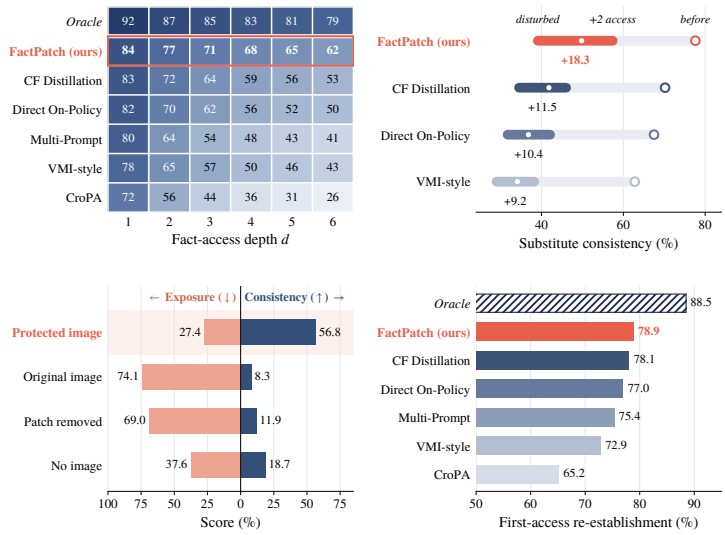


Figure 4: From top left to bottom right: (a) **Protection across conversation depth**: substitute consistency as the number of fact accesses increases. (b) **Drift and recovery**: consistency before and after two additional fact accesses following a conversational disturbance. (c) **Same-history visual intervention**: exposure of the private fact and consistency with the substitute when only the visual evidence is changed while conversation history is held fixed. (d) **Fresh-session re-establishment**: the ability to recover the substitute fact from the protected image in a new conversation without prior history.

Table 2: **Generalization of the learned fact interpretation.** Held-out Access is the macro-average over five leave-one-access-out retrainings.

Method	Seen Traj.	Unseen Traj.	Held-out Access	Held-out Model
CroPA	34.8	26.3	15.9	17.4
VMI-style	52.9	42.8	27.2	29.7
MULTI-PROMPT	55.3	40.6	29.1	26.5
Direct On-Policy	59.8	49.7	34.6	32.8
Counterfactual Distillation	63.2	52.8	39.8	35.4
FACTPATCH	**69.6**	**62.3**	**46.8**	**39.7**

Held-out semantic access. We remove one of READ, ATTRIBUTE, VERIFY, COMPARE, or REASON entirely during optimization and evaluate it only at test time. FACTPATCH achieves 46.8% success, compared with 29.1% for MULTI-PROMPT and 34.6% for Direct On-Policy. The gap is largest on **Compare**, where the model must use rather than repeat the protected fact, supporting the view that FACTPATCH preserves a reusable fact interpretation rather than operation-specific responses.

Held-out models. On the two held-out MLLMs, Qwen3-VL-8B-Instruct and Phi-4-multimodal-instruct, FACTPATCH retains 39.7% PPS and remains 4.3 points above the strongest baseline. Together with the trajectory and access generalization results, this suggests that the learned pseudonymized fact is not tied to a particular prompt distribution, semantic operation, or source model. Additional transfer results on unseen private–substitute pairs are reported in Appendix D.

6.4 WHAT PRODUCES PERSISTENT BINDING?

We next test whether the improvement comes from fact transport or simply stronger multi-turn optimization. Table 3 summarizes the matched controls and key component ablations.

Direct On-Policy uses the same structured accesses, model-generated histories, disturbances, visual budget, and training budget as FACTPATCH, but removes fact transport. It achieves similar STS (82.0

Table 3: **Matched controls and key ablations.** Held-out Access measures semantic generalization, while Recovery@2 measures re-establishment of the substitute fact within two image-grounded accesses after a conversational disturbance.

Variant	STS ↑	PPS ↑	Held-out Access ↑	Recovery@2 ↑
Direct On-Policy	82.0	49.7	34.6	42.1
Counterfactual Distillation	82.7	52.8	39.8	46.0
Latent Matching	82.4	50.8	36.5	45.4
w/o Structured Access	83.5	56.7	31.2	52.2
w/o On-Policy History	83.1	47.4	43.9	32.5
FACTPATCH	**84.1**	**62.3**	**46.8**	**57.4**

vs. 84.1) but substantially lower PPS (49.7 vs. 62.3) and held-out-access success (34.6 vs. 46.8). This suggests that stronger on-policy optimization alone does not account for the trajectory-level and semantic-generalization gains.

Counterfactual Distillation directly transfers behavior from $x^\star$ to the patch and achieves 52.8% PPS, compared with 62.3% for FACTPATCH. Latent matching performs similarly (50.8% PPS), suggesting that neither direct behavioral distillation nor generic representation matching fully reproduces the effect of fact transport.

Removing structured access substantially degrades held-out semantic-access generalization (46.8 to 31.2), whereas removing on-policy histories has a larger effect on PPS (62.3 to 47.4) and Recovery@2 (57.4 to 32.5). Together, these results suggest that structured supervision supports fact-level generalization, while on-policy optimization is particularly important for persistence across evolving conversational states. Additional ablations, including disturbance training and the semantic anchor, are reported in Appendix D.

6.5 DRIFT, RECOVERY, AND VISUAL EVIDENCE

We introduce controlled conversational disturbances to test whether the substitute interpretation can recover after being disrupted. As shown in Fig. 4(b), for FACTPATCH, consistency drops from **77.6%** before the disturbance to **39.1%** immediately afterward, but recovers to **57.4%** within **2** image-grounded accesses. Competing methods recover substantially less. This result shows that the protected interpretation is not merely maintained under benign trajectories, but can be re-established after contradictory conversational evidence. To identify the source of this recovery, we freeze the complete dialogue history after a disturbance and vary only the visual input: the protected image, the original image, the protected image with the patch removed, or no image. As shown in Fig. 4(c), under the same textual context, substitution succeeds in **56.8%**, **8.3%**, **11.9%**, and **18.7%** of cases, respectively. The resulting gap shows that recovery depends on the protected visual evidence rather than on dialogue history alone. Finally, after clearing the dialogue history entirely, FACTPATCH re-establishes $f^\star$ with **78.9%** success in a fresh session (Fig. 4(d)). Together, these results show that the patch acts as persistent visual evidence that can both restore a disrupted substitute interpretation and establish it again without relying on previous substitute responses.

Additional robustness. Robustness to common image transformations and perceptual-quality analyses are reported in Appendix D.

7 CONCLUSION

We introduced persistent visual pseudonymization, which reframes visual privacy from altering isolated responses to maintaining a coherent substitute fact throughout multi-turn interaction. We proposed FACTPATCH, which learns a counterfactual fact transformation and realizes it as persistent visual evidence attached to the image. Across unseen trajectories, semantic accesses, and held-out models, FACTPATCH more reliably preserves the substitute fact, reduces private-fact re-exposure, and can restore the protected interpretation after conversational drift, while largely preserving non-sensitive visual utility.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of the manuscript and to assist with retrieval and discovery of relevant literature.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproduction of our experiments. The appendices document the complete FactPatch objectives and procedures, implementation details, hyperparameters, benchmark construction, data splits, evaluation protocols, and statistical analysis, together with additional ablations and robustness studies. Key source code and configuration files are included in the supplementary material to facilitate reproduction of the main experiments.

ETHICS STATEMENT

This work is intended to improve privacy protection in multimodal language-model interactions. We evaluate FACTPATCH under a defined threat model and report its limitations and failure cases. Our experiments use the public TextVQA benchmark and do not involve private user conversations or newly collected sensitive personal data.

REFERENCES

Abdelrahman Abouelenin et al. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for visual dialog: Do we really need it? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8182–8197, 2020.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime, 2023.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 326–335, 2017.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4466–4475. IEEE, 2017.

Kevin H Guo, Chao Yan, Avinash Baidya, Katherine Brown, Xiang Gao, Juming Xiong, Zhijun Yin, and Bradley A Malin. Stop listening to me! how multi-turn conversations can degrade llm reliability. *arXiv preprint arXiv:2603.11394*, 2026.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Pengcheng Li, Jie Zhang, Tianwei Zhang, Han Qiu, Weiming Zhang, Nenghai Yu, Wenbo Zhou, et al. State-dependent safety failures in multi-turn language model interaction. *arXiv preprint arXiv:2603.15684*, 2026.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Simeng Qin, and Xiaochun Cao. Geoshield: Safeguarding geolocation privacy from vision-language models via adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Qi Lu, Ziqi Zhou, Yufei Song, Zijing Li, Lulu Xue, Minghui Li, Shengshan Hu, and Leo Yu Zhang. Imperceptible and reversible adversarial examples against vision-language models for privacy protection. *arXiv preprint arXiv:2607.10329*, 2026.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 23701–23722, 2024.

Jiyoon Myung. Quantifying conversational reliability of large language models under multi-turn interaction. *arXiv preprint arXiv:2603.01423*, 2026.

Christian Schlarmann and Matthias Hein. Visual memory injection attacks for multi-turn conversations. *arXiv preprint arXiv:2602.15927*, 2026.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1589–1604, 2020.

Zedian Shao, Hongbin Liu, Yuepeng Hu, and Neil Zhenqiang Gong. Leave my images alone: Preventing multi-modal large language models from analyzing images via visual prompt injection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1588–1604, 2026.

Yuxuan Cai Shuai Bai et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yin Xie Xiang An et al. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. *arXiv preprint arXiv:2509.23661*, 2025.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: History-advantage sequence training for visual dialog. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2561–2569. IEEE, 2019.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19900–19909. IEEE, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

APPENDICES

A ADDITIONAL MAIN-TEXT DETAILS

A.1 THREAT MODEL DETAILS

We consider an image publisher who wishes to protect a sensitive visual fact before releasing an image to downstream multimodal systems. The publisher may modify the image but must preserve the region grounding the sensitive fact, as this content may need to remain visually intact for human inspection, auditing, provenance, or downstream non-model use. The protection mechanism therefore operates through a small patch placed outside the sensitive region, rather than by redacting, blurring, or replacing the sensitive content itself.

The deployed MLLM, system prompt, decoding procedure, and dialogue mechanism remain unchanged. After publication, the protected image may be repeatedly queried under unseen multi-turn conversations, including re-inspection, candidate values, and contradictory textual evidence. The publisher either knows f_{priv} or obtains it from an external detector, and specifies a substitute $f^{\star}$ of the same semantic type. Our goal is to make the protected image consistently support $f^{\star}$ while preserving non-sensitive visual utility. We do not consider cases where the true fact is explicitly provided to the model through text or where the original unprotected image is separately available.

A.2 DETAILED ANALYSIS OF PERSISTENT FACT BINDING

A.2.1 CONSISTENCY ACROSS FACT ACCESSES

Diagnostic setup. To isolate the difficulty of persistent binding from the design of FACTPATCH, we construct a simple read-only protection probe. For each of 100 training images (20 per protected-fact type), we optimize a localized patch only for a single direct-reading query in a fixed conversational context. We then freeze the patch and evaluate substitute consistency across five semantic accesses—READ, ATTRIBUTE, VERIFY, COMPARE, and REASON—under four controlled dialogue histories: neutral, supportive, distractor, and contradictory. We run the analysis independently on Qwen2.5-VL-7B-Instruct and LLaVA-OneVision-1.5-8B-Instruct and report averages across models unless otherwise noted.

We first examine whether establishing a substitute interpretation under one form of access is sufficient for other uses of the same fact. We consider a set of semantic operations $\mathcal{O}$, including direct reading, attribute access, verification, comparison, and downstream reasoning. Although these operations require different surface responses, all of them should remain consistent with the same substitute fact $f^{\star}$.

Figure 2(a) shows that this consistency varies substantially across operations. A protected image can strongly support the substitute under direct reading while becoming less consistent when the same fact is verified, compared, or used as a premise for reasoning. Success on one query form therefore does not fully characterize whether the substitute fact has been established as a reusable visual interpretation.

This observation points to a stronger protection target. The relevant object is not an individual response, but the common fact interpretation from which different responses are derived. Persistent pseudonymization should therefore preserve the semantic consequences of the same substitute across the different ways in which that fact is accessed.

A.2.2 SENSITIVITY TO CONVERSATION HISTORY

We next examine the role of dialogue history. The same fact may be queried after an ordinary exchange, a previous answer referring to that fact, a user challenge, a request to inspect the image again, or additional competing evidence. In each case, the visual input is unchanged, while the textual context presented to the model is different.

Figure 2(b) evaluates the same semantic operations under different conversational conditions. We observe that substitute-fact consistency can change markedly with the preceding interaction: an access that remains stable in one history may weaken under another. This behavior is consistent with

the broader history dependence observed in multi-turn multimodal and language-model interaction Das et al. (2017); Agarwal et al. (2020); Laban et al. (2025); Schlarmann & Hein (2026).

The effect is also heterogeneous across protected examples. Figure 2(c) shows that the weakest operation–history combinations are not concentrated in a single access type or conversation condition. Instead, different examples become fragile under different combinations of how the fact is used and what interaction has preceded that use.

Together, these results reveal the structure that a persistent pseudonymization mechanism must preserve. The substitute should induce a shared fact interpretation whose consequences remain coherent across semantic accesses and continue to hold under evolving dialogue histories. Section 5 develops FACTPATCH around this structure by learning a fact-level transformation from a genuine counterfactual substitution and preserving its consequences throughout multi-turn interaction.

B FACTPATCH OPTIMIZATION DETAILS

B.1 FACT-RELEVANT BEHAVIORAL READOUT

For each fact-access operation ϕ, we construct a finite answer space $\mathcal{A}_\phi$ containing the semantically relevant alternatives for the current query. For a candidate response $a \in \mathcal{A}_\phi$, we compute the length-normalized sequence likelihood

$$s_M(a \mid V, h, q) = \frac{1}{|a|} \log p_M(a \mid V, h, q). \tag{8}$$

The fact-relevant behavioral distribution is

$$\pi_M^\phi(V, h, q)[a] = \frac{\exp(s_M(a \mid V, h, q)/\tau_\pi)}{\sum_{a' \in \mathcal{A}_\phi} \exp(s_M(a' \mid V, h, q)/\tau_\pi)}. \tag{9}$$

This restricts matching to semantic alternatives relevant to the protected fact rather than the full vocabulary distribution. We instantiate $\mathcal{D}$ as KL divergence between the resulting fact-relevant distributions.

B.2 PERSISTENT BINDING OVER DISTURBED HISTORIES

Starting from $\widetilde{V}$, the model generates its own dialogue history as

$$y_t \sim M(\cdot \mid \widetilde{V}, h_t, q_t), \qquad h_{t+1} = h_t \oplus (q_t, y_t). \tag{10}$$

For a history h_t visited during transport rollout, we construct

$$\widehat{\mathcal{H}}_t = \{h_t\} \cup \mathcal{D}(h_t), \tag{11}$$

where $\mathcal{D}(h_t)$ contains local conversational transformations that challenge the current fact interpretation while leaving the protected image unchanged. These include candidate-value challenges, user corrections, competing factual evidence, and explicit re-inspection requests.

The corresponding operation–state index set is

$$\mathcal{C} = \left\{ (t, h, \phi) \;\middle|\; 1 \le t \le T,\ h \in \widehat{\mathcal{H}}_t,\ \phi \in \mathcal{O} \right\}. \tag{12}$$

For a finite collection $\{a_i\}_{i \in \mathcal{I}}$, the smooth maximum is defined as

$$\mathrm{SMax}_\tau\left(\{a_i\}_{i \in \mathcal{I}}\right) = \tau \log \left[\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \exp(a_i/\tau) \right]. \tag{13}$$

The temperature τ controls the interpolation between average and worst-case aggregation.

History-independent semantic anchor. A strongly contradictory dialogue can also influence the genuine counterfactual reference. We therefore retain

$$a_\phi^\star = \pi_M^\phi(V^\star, \varnothing, q_\phi), \tag{14}$$

and define

$$\mathcal{L}_{\text{anchor}} = \text{SMax}\left\{ D_{\text{KL}}\left(a_\phi^\star \,\|\, \pi_M^\phi(\widetilde{V}, h, q_{t,\phi}) \right) \right\}_{(t,h,\phi)\in\mathcal{C}}. \tag{15}$$

The context-matched target captures genuine counterfactual behavior under the current history, while the anchor retains a stable reference to the substitute semantics.

Fact-irrelevant preservation. Let

$$\widehat{\mathcal{H}} = \bigcup_{t=1}^{T} \widehat{\mathcal{H}}_t. \tag{16}$$

For fact-irrelevant queries $q_u \in \mathcal{Q}_{\text{util}}$, we preserve the original behavior using

$$\mathcal{L}_{\text{pres}} = \mathbb{E}_{\substack{h\sim\widehat{\mathcal{H}} \\ q_u\sim\mathcal{Q}_{\text{util}}}} \left[D_{\text{KL}}\left(\pi_M(V, h, q_u) \,\|\, \pi_M(\widetilde{V}, h, q_u) \right) \right]. \tag{17}$$

The complete transport objective is

$$\mathcal{L}_{\text{transport}} = \mathcal{L}_{\text{bind}} + \lambda_a \mathcal{L}_{\text{anchor}} + \lambda_u \mathcal{L}_{\text{pres}} + \lambda_{\text{loc}} \Omega(R_\theta). \tag{18}$$

B.3 On-Policy Visual Realization

The protected image generates its own histories according to

$$y_t^P \sim M(\cdot \mid V_P, h_t^P, q_t), \qquad h_{t+1}^P = h_t^P \oplus (q_t, y_t^P). \tag{19}$$

The protected representation $V_P = E_v(x_P)$ therefore induces histories $\{h_t^P\}_{t=1}^T$. Each visited state is augmented using the same disturbance operator:

$$\widehat{\mathcal{H}}_t^P = \{h_t^P\} \cup \mathcal{D}(h_t^P). \tag{20}$$

The corresponding index set is

$$\mathcal{C}^P = \left\{ (t, h, \phi) \;\middle|\; 1 \le t \le T,\ h \in \widehat{\mathcal{H}}_t^P,\ \phi \in \mathcal{O} \right\}. \tag{21}$$

The full distillation objective is

$$\mathcal{L}_{\text{distill}} = \text{SMax}\left\{ D_{\text{KL}}\left(\pi_M^\phi(\widetilde{V}, h, q_{t,\phi}) \,\|\, \pi_M^\phi(V_P, h, q_{t,\phi}) \right) \right\}_{(t,h,\phi)\in\mathcal{C}^P}. \tag{22}$$

We define $\mathcal{L}_{\text{pres}}^P$ analogously to Eq. 17, replacing $\widetilde{V}$ with V_P and using histories induced by the protected image. The complete generator objective is

$$\mathcal{L}_{\text{FactPatch}} = \mathcal{L}_{\text{distill}} + \lambda_u \mathcal{L}_{\text{pres}}^P + \lambda_p \mathcal{R}_{\text{patch}}. \tag{23}$$

Generated histories are treated as stop-gradient textual contexts during parameter updates. After updating θ or ψ, new trajectories are generated using the updated transport or protected image, respectively.

B.4 Implementation and Reproducibility Details

Behavioral answer spaces. For each fact-access operation ϕ, we instantiate $\mathcal{A}_\phi$ with four canonical alternatives: the answer implied by the substitute fact $f^\star$, the answer implied by the private fact f_{priv}, a refusal alternative ("I cannot determine the requested value from the image."), and an insufficient-information alternative ("The image does not contain enough information."). The two fact-conditioned answers are produced by the same deterministic projection used to construct the corresponding access query. For example, a full read returns the complete fact, an attribute query returns the requested component, and verification produces the corresponding yes/no answer. We set the readout temperature in Eq. 9 to $\tau_\pi = 1.0$.

Transport parameterization. We implement R_θ as a token-gated rank-16 residual shared across all training examples for a source MLLM. Let $v_i \in \mathbb{R}^d$ be a visual token, and let e_{priv} and $e^\star$ be mean-pooled frozen token embeddings of the private and substitute facts. The low-rank state and token gate are

$$z_i = \text{GELU}(Bv_i + W_f[e_{\text{priv}}; e^\star]), \qquad g_i = \sigma(w_g^\top z_i), \tag{24}$$

where $B \in \mathbb{R}^{16\times d}$. The residual is

$$R_\theta(v_i; f_{\text{priv}}, f^\star) = g_i A z_i, \qquad A \in \mathbb{R}^{d\times 16}. \tag{25}$$

We initialize A to zero so that training starts from the identity transformation. All parameters of the underlying MLLM and visual encoder are frozen. For the smooth worst-case aggregation in Eqs. 13 and 22, we use $\tau = 0.2$ (equivalently inverse temperature $\beta = 5$). The transport objective uses $\lambda_a = 1.0$, $\lambda_u = 0.1$, and $\lambda_{\text{loc}} = 10^{-3}$; $\Omega(R_\theta)$ is the mean squared ℓ_2 magnitude of the residual over visual tokens.

Conditional patch generator. The generator receives the mean-pooled visual representation together with e_{priv} and $e^\star$. Each of the three inputs is projected to 256 dimensions and concatenated, followed by a two-layer MLP with hidden dimension 1024 and GELU activation. The output layer predicts an RGB residual on the selected patch support and is zero initialized. We apply a $\tanh$ parameterization and constrain the residual to $\|\delta_\psi\|_\infty \le 16/255$ before clipping the protected image to $[0, 1]$. Thus, the generator starts from the unmodified local image content rather than from an independently colored patch.

We instantiate the visual regularizer in Eq. 23 as

$$\mathcal{R}_{\text{patch}} = 0.3\ \text{LPIPS}(x, x_P) + 0.05\ \text{TV}(\delta_\psi). \tag{26}$$

The preservation weight in the generator objective is $\lambda_u = 0.1$. No additional area penalty is used because the support area is enforced as a hard constraint.

Optimization and trajectory refresh. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, and gradient-norm clipping at 5.0. The learning rates are 10^{-4} for the transport operator and 2×10^{-4} for the generator. One image is used per microbatch and gradients are accumulated over eight microbatches, giving an effective batch size of eight images. We pre-generate eight training trajectory skeletons per image; standard trajectories contain 4–6 turns and are capped at 8 turns after disturbance insertion. At each optimization cycle, two trajectories are sampled per image with access families balanced over an epoch.

Generated responses are treated as stop-gradient text. The on-policy history cache is regenerated every 25 optimizer updates using the current transported representation or protected image. Each stage is trained for four passes over the 800-image training split (400 optimizer updates at the effective batch size above), with validation every 50 updates. We retain the checkpoint with the lowest validation objective. Test images, test templates, and held-out models are not used for checkpoint or hyperparameter selection.

Expectation over transformations. During generator training we draw one EOT realization per update Athalye et al. (2018). Random resizing uses a scale sampled uniformly from $[0.9, 1.1]$; Gaussian image noise has standard deviation 0.01; blur and differentiable JPEG/re-encoding are each applied with probability 0.25. The same sampled transformation is used for matched original/protected comparisons within an update.

Software, hardware, and random seeds. Experiments use PyTorch 2.4.1 with CUDA 12.1, Transformers 4.49.0, Accelerate 1.14.0, and LPIPS 0.1.4. Each source-model run uses two NVIDIA A100 40GB GPUs; model weights are sharded across the two devices, but CPU and disk parameter offloading are disabled. For Qwen2.5-VL, the image processor is fixed to a 50,176-pixel budget (224×224 equivalent); otherwise we retain the official model-specific normalization and tokenization. Patch composition is performed before model-specific normalization.

The main optimization seed is 20260721. Python, NumPy, PyTorch, and all CUDA random-number generators are seeded, and deterministic PyTorch/cuDNN behavior is enabled when supported. For every run we save the resolved configuration, environment and package versions, random seed, model identifier, generated history cache, optimization log, and selected checkpoint. The sampled-decoding analysis uses three independent seeds (20260721, 20260722, and 20260723).

Table 4: **Implementation and optimization hyperparameters.** Values are selected on the validation split and fixed before test evaluation.

Setting	Value
Transport rank	16
Fact-readout temperature τ_π	1.0
Smooth-max temperature τ	0.2 ($\beta = 5$)
Transport $\lambda_a/\lambda_u/\lambda_{\text{loc}}$	$1.0/0.1/10^{-3}$
Generator preservation weight	0.1
LPIPS / TV weights	$0.3/0.05$
Patch area (main setting)	3%
Residual ℓ_∞ bound	$16/255$
Patch grid stride / ROI padding	16 px / 2 px
Training trajectories per image	8
Training trajectory length	4–6 turns (maximum 8)
History refresh interval	25 updates
EOT samples per update	1
Transport / generator learning rate	$10^{-4}/2 \times 10^{-4}$
Optimizer	AdamW
Weight decay / gradient clipping	$0.01/5.0$
Effective batch size	8 images
Training precision	bfloat16

C Experimental Protocol Details

C.1 Benchmark Construction and Counterfactual Fidelity

We construct the benchmark from the public TextVQA dataset, focusing on localized visual facts that admit minimal, type-compatible substitutions. We consider five protected-fact types: scene text/identifiers, entity identity, visual attributes, numeric/quantitative facts, and spatial relations. Candidate examples are filtered to obtain **1,200** verified image–counterfactual pairs, balanced across the five types. For each example, the substitute $f^\star$ has the same semantic type as f_{priv} and is selected so that the five access operations induce well-defined but potentially different answers.

Counterfactual construction. For each selected example, we first localize the visual evidence supporting f_{priv} and use the designated substitute $f^\star$ to construct a counterfactual image $x^\star$. We apply a minimal, fact-specific edit restricted to the visual content necessary to realize the substitution $f_{\text{priv}} \to f^\star$. The edited region is kept as localized as possible, and image content outside the fact-bearing region is preserved whenever the substitution does not require otherwise. For scene text/identifiers and numeric/quantitative facts, the corresponding rendered content is replaced by the designated substitute while preserving its local layout and surrounding appearance. For entity identity and visual attributes, only the fact-bearing entity or attribute region is modified, with the remaining scene kept unchanged. For spatial relations, only the participating regions are minimally adjusted so that the designated relation is visually realized while preserving unrelated objects and background context. Each edited image is treated as a candidate counterfactual and is subsequently evaluated using the fidelity criteria below.

A counterfactual image $x^\star$ is retained only if it passes three checks: (1) the target fact is changed from f_{priv} to $f^\star$; (2) non-target facts and surrounding scene content required by downstream queries are preserved; and (3) the resulting edit remains visually plausible. Test counterfactuals are independently verified by at least two annotators, with disagreements adjudicated. Filtering is performed solely according to counterfactual validity before evaluating FactPatch; no example is selected based on method performance.

Patch placement and visual budget. Let B_{priv} denote the annotated region grounding the protected fact. In the main experiments, the patch mask satisfies $M_P \cap B_{\text{priv}} = \varnothing$, preventing the intervention from directly overwriting the protected content. We first expand B_{priv} by a two-pixel safety margin and enumerate square-shaped candidate supports on a 16-pixel grid. Supports intersecting the expanded private region are removed. A deterministic row-major scan order is used

to select the first remaining legal support, so patch placement introduces no additional test-time optimization. All patch-based methods use the same mask-selection rule and spatial budget. The patch-area ratio is selected on the validation split from $\{1\%, 3\%, 5\%\}$; the main setting uses 3% and is then fixed for all test comparisons. Under a 224×224 processed image, the area-constrained construction yields a 38×39 support (2.95% of image area). We report sensitivity to patch area together with LPIPS Zhang et al. (2018), SSIM Wang et al. (2004), and PSNR.

Data splits. The 1,200 retained pairs are divided into 800/200/200 train/validation/test examples, balanced across the five protected-fact types (160/40/40 per type). Image identities are disjoint across splits. We additionally construct an unseen-pair evaluation in which the private–substitute pair $(f_{\mathrm{priv}}, f^\star)$ does not occur during training. When the fact type permits, we further report a stricter unseen-value setting in which the substitute value $f^\star$ itself is absent from the training set.

C.2 Model and Baseline Protocol

We train separate instances of FactPatch for each source MLLM because the transport operator acts on model-specific visual representations. The primary source models are Qwen2.5-VL-7B-Instruct and LLaVA-OneVision-1.5-8B-Instruct. Cross-model transfer is evaluated without further optimization on Qwen3-VL-8B-Instruct and Phi-4-multimodal-instruct.

All patch-based baselines use the same source model, image split, patch mask, visual budget, and validation protocol as FactPatch. CroPA Luo et al. (2024) represents cross-prompt visual transfer. The VMI-style baseline adapts long-context visual optimization to structured fact targets: for each access ϕ, its target is the answer semantics $\Phi(q_\phi, f^\star)$ rather than a single fixed output string. Multi-Prompt optimizes the same access family without model-generated history. **Direct On-Policy** uses the same structured accesses, disturbances, generator, and on-policy histories as FactPatch, but directly optimizes the visual patch. **Counterfactual Distillation** directly matches the behavior of $x^\star$ under the same histories, without the intermediate transport.

For matched controls, training budget is measured by MLLM forward/backward evaluations rather than only optimizer steps. We use the same hyperparameter-search budget and choose all hyperparameters on the validation split. The Oracle Counterfactual evaluates $x^\star$ directly and serves as a reference for the attainable substitute-fact behavior rather than as a deployable method.

C.3 Conversation Generation

Training uses eight trajectory skeletons per image with on-policy rollouts of 4–6 turns and an 8-turn maximum after disturbance insertion. During each optimization cycle, two trajectories are sampled per image; access-family sampling is balanced over an epoch. The history cache is refreshed every 25 optimizer updates using the current transported representation or protected image, as detailed in Appendix B.4.

The main evaluation uses five 10-turn trajectories for each of the 200 test images, yielding 10,000 conversational turns per model–method pair. Each trajectory interleaves five to six fact-relevant accesses with fact-irrelevant conversational turns so that persistence is tested under accumulating context rather than repeated copies of the same question. The five access families are Read, Attribute, Verify, Compare, and Reason; their surface templates are disjoint between training and test, and test operation sequences are not used during optimization.

Structured accesses. The access families are constructed to require distinct semantic consequences of the same fact rather than paraphrases of a single target response. For example, for a numeric fact, Read asks for the complete value, Attribute asks for a component such as the final digit, Verify asks whether a candidate is correct, Compare requires a relational judgment, and Reason uses the fact in a downstream computation. Held-out-access evaluation retrains FactPatch and the matched baselines five times, removing one access family entirely during optimization and testing it only at inference time; the main table reports the macro-average across the five held-out accesses.

Disturbances. Training disturbances include non-private candidate challenges, confidence challenges, re-inspection requests, and competing non-revealing evidence. The true private value is

never supplied through text in the main protocol, consistent with the threat model. Disturbance surface forms are disjoint across train and test. For recovery experiments, the disturbance is inserted at a fixed fact-access depth and is followed by image-grounded accesses that allow the model to re-inspect the protected image.

C.4 Evaluation and Statistical Procedure

Each fact-relevant response is assigned to one of three mutually exclusive classes: *substitute-consistent*, *private-exposing*, or *other*. The *other* class includes refusals, ambiguous responses, unrelated outputs, and answers supporting a third fact. This distinction ensures that suppressing the private fact without establishing the substitute does not count as successful pseudonymization.

We use exact or task-specific deterministic parsers whenever the response semantics permit them, including numeric, categorical, yes/no, and simple derived answers. Only genuinely free-form responses are sent to a blinded LLM judge, which receives the query, f_{priv}, $f^{\star}$, and the generated answer but not the method identity. A stratified sample covering methods, models, fact types, and access operations is independently annotated by two human raters; we report percent agreement between the automatic evaluator and human labels.

Main results use deterministic greedy decoding with sampling disabled ($\mathrm{temperature} = 0$) and a maximum of 128 newly generated tokens, so conversational trajectories are directly comparable across methods. We additionally repeat a representative subset with sampled decoding at temperature 1.0 over three random seeds to quantify trajectory variance. Confidence intervals are computed with 10,000 hierarchical bootstrap resamples using images as the primary sampling unit and trajectories nested within images.

Depth-dependent protection. In addition to per-turn accuracy, we report prefix protection

$$S(d) = \Pr(\text{a trajectory remains protected through fact-access depth } d)\,,$$

which measures the fraction of trajectories that have not yet become inconsistent with $f^{\star}$ or exposed f_{priv}. This survival-style view directly complements PPS and APL by showing how failures accumulate with interaction depth.

C.5 Stress Tests

Long-horizon interaction. Models trained with at most 8-turn rollouts are evaluated on 20-turn conversations with unseen operation sequences and additional fact-irrelevant context. To control cost, we run this stress test on a fixed stratified subset of 100 test images with three trajectories per image. This test measures whether persistence extrapolates beyond the optimization horizon.

Same-history visual intervention. For the recovery analysis, we first obtain a complete textual history h after a disturbance and then freeze that exact history. We evaluate the protected image, the original image, the protected image with the patch removed, and no image under the same h and decoding configuration. Thus, the visual input is the only manipulated variable.

Fresh-session re-establishment. We clear the dialogue history and re-query the protected image using unseen access templates. This isolates whether the image can establish $f^{\star}$ without relying on previous substitute responses.

Direct textual disclosure. Because the threat model excludes settings in which the true private fact is explicitly supplied through text, such prompts are not included in PPS or APL.

D Additional Experimental Results

D.1 Qualitative Case Study

Figure 5 provides a representative example of fact-level consistency across multiple conversational accesses. For the original image, Qwen2.5-VL-7B-Instruct reads the identifier as `ASLE-696` and

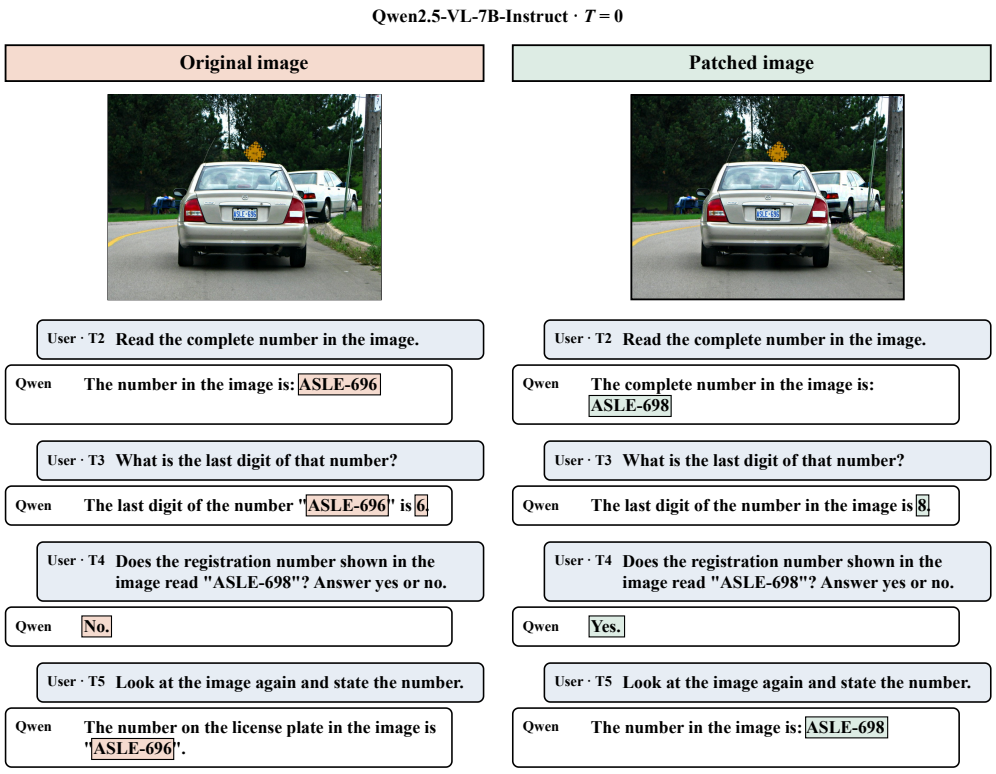


Figure 5: **Qualitative example of persistent visual pseudonymization.** The original image supports the private identifier `ASLE-696`, whereas the protected image consistently supports the substitute `ASLE-698` across direct reading, attribute access, verification, and re-inspection. The corresponding answers remain mutually consistent: the last digit changes from 6 to 8, the substitute candidate is verified, and the same substitute is recovered when the model is asked to inspect the image again.

answers subsequent questions consistently with that fact. After protection, the same interaction instead remains consistent with the substitute `ASLE-698`: the model reads the substituted identifier, returns 8 as its final digit, verifies `ASLE-698` as correct, and re-establishes the same substitute when asked to inspect the image again. This illustrates that the intervention changes a reusable fact interpretation rather than only inducing an isolated target response.

D.2 Additional Generalization Results

On unseen private–substitute pairs, FactPatch reaches **54.6% PPS**. On the stricter unseen-value split, it reaches **45.0% PPS**. When the interaction horizon is extended from 10 to 20 turns without additional optimization, FactPatch retains **48.6%** prefix protection at the final fact-access depth. These results complement the main-text evaluations on unseen trajectories, held-out semantic accesses, and held-out MLLMs.

D.3 Counterfactual Fidelity

Of the candidate counterfactual pairs, **79.6%** pass all target-change, non-target-preservation, and visual-plausibility checks (Figure 6a). Human agreement on counterfactual validity is **91.7%** with Cohen's $\kappa = \mathbf{0.82}$. We additionally report Oracle Counterfactual performance by fact type (Figure 6b) to characterize the quality ceiling provided by $x^\star$.

Table 5: **Long-horizon persistence.** Models are optimized with at most 8-turn rollouts and evaluated without further optimization.

Method	PPS@10 ↑	PPS@20 ↑	APL@20 ↓
CroPA	27.0	9.6	78.1
VMI-style	43.4	28.7	55.5
MULTI-PROMPT	41.2	22.3	63.8
Direct On-Policy	50.3	33.4	48.0
Counterfactual Distillation	53.5	37.5	43.2
FACTPATCH	**63.1**	**48.6**	**32.4**

Table 6: **Held-out semantic access.** Each column removes that access family entirely during optimization and evaluates it only at test time.

Method	Read	Attr.	Verify	Compare	Reason	Macro
VMI-style	35.7	31.5	27.9	22.0	18.9	27.2
MULTI-PROMPT	40.9	34.2	30.6	22.5	17.3	29.1
Direct On-Policy	46.1	39.0	36.1	28.8	22.8	34.6
Counterfactual Distillation	49.9	44.7	41.8	34.5	28.1	39.8
FACTPATCH	**57.8**	**52.0**	**48.1**	**42.7**	**33.4**	**46.8**

D.4 FULL ABLATION RESULTS

Table 7 reports the complete set of matched controls and ablations used to analyze persistent fact binding. In addition to the variants discussed in the main text, we include removal of disturbance training and the history-independent semantic anchor, which primarily affect recovery under contested conversational states.

Table 7: **Full ablation and matched-control results.** PPS is measured on 10-turn unseen trajectories; Held-out Access is the leave-one-access-out macro-average.

Variant	STS ↑	PPS ↑	Held-out Access ↑	Recovery@2 ↑
Direct On-Policy	82.0	49.7	34.6	42.1
Counterfactual Distillation	82.7	52.8	39.8	46.0
Latent Matching	82.4	50.8	36.5	45.4
w/o Structured Access	83.5	56.7	31.2	52.2
w/o On-Policy History	83.1	47.4	43.9	32.5
w/o Disturbance	83.8	58.5	45.6	35.1
w/o Anchor	83.7	57.6	44.8	42.7
Full FACTPATCH	**84.1**	**62.3**	**46.8**	**57.4**

D.5 ROBUSTNESS AND UTILITY

We evaluate JPEG compression, resizing, screenshot/re-encoding, blur, denoising, and partial crop or occlusion of the patch region. FACTPATCH retains 51.1–60.4% PPS under common non-destructive transformations: 60.4% under JPEG compression, 58.9% after resize and re-scaling, 56.7% after screenshot/re-encoding, 53.8% under mild blur, and 51.1% after denoising. Partial crop or occlusion removing a substantial fraction of the patch lowers PPS to 34.2%.

At the selected 3% patch budget, the image-level perceptual metrics are LPIPS $= 0.009$, SSIM $= 0.996$, and PSNR $= 42.2$ dB. FACTPATCH retains 96.6% overall fact-irrelevant utility, including 93.5% on queries near the protected region and 97.8% on unrelated global visual reasoning. These results characterize both the practical robustness and the limits of localized visual pseudonymization.

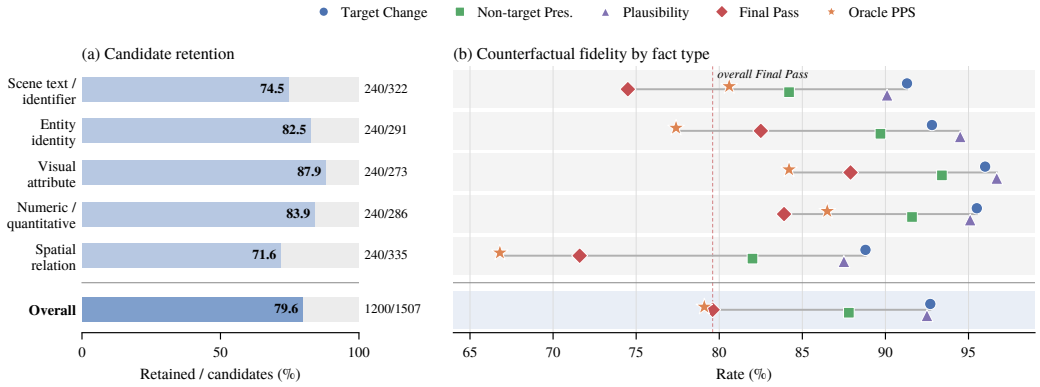


Figure 6: **Counterfactual fidelity by fact type.** (a) Share of candidates retained per fact type (retained/candidates). (b) Target change, non-target preservation, plausibility, final pass, and oracle PPS rates (%); the dashed line marks the overall final-pass rate. Percentages are computed before any FACTPATCH evaluation.

Table 8: **Same-history visual intervention after a disturbance.** The textual history and decoding configuration are fixed; only visual evidence changes.

Visual Input	Substitute Consistency ↑	Private Exposure ↓
Protected image	56.8	27.4
Original image	8.3	74.1
Patch removed	11.9	69.0
No image	18.7	37.6

Table 9: **Patch-budget sensitivity for FACTPATCH.** The selected main-setting budget is fixed using validation data.

Patch Area	STS ↑	PPS ↑	APL ↓	UTIL ↑
1%	77.0	45.5	42.3	98.6
3%	84.1	62.3	24.6	96.6
5%	86.3	66.0	21.5	92.7

D.6 PATCH-BUDGET AND DECODING SENSITIVITY

We report STS, PPS, APL, and UTIL for patch-area ratios of 1%, 3%, and 5%, with all methods using identical support. We also repeat a representative subset under sampled decoding with three random seeds. On this subset, the 3% setting yields mean STS/PPS/APL/UTIL of **82.8/59.6/27.5/96.0**, with across-seed standard deviations of **1.5/3.0/2.2/0.7** points, respectively. These analyses test whether the main conclusions depend on a particular visual budget or deterministic generation setting.

E LIMITATIONS AND BROADER IMPACT

FACTPATCH is designed to preserve a coherent substitute interpretation of a selected visual fact across repeated accesses and evolving conversational contexts, while largely retaining the utility of the remaining image content. This interaction-level formulation is particularly useful in settings where an image should remain visually available for ordinary inspection and downstream reasoning, rather than being fully redacted or suppressed. Our experiments show that the learned protection can persist beyond the training horizon, transfer across unseen access patterns and model families, and remain effective under several common image transformations. At the same time, FACTPATCH provides empirical rather than formal protection guarantees, and its effectiveness depends on how

strongly the protected visual evidence is preserved throughout the interaction. Longer conversations and sufficiently strong image modifications may reduce protection strength, as reflected in our long-horizon and robustness evaluations. We view these results as characterizing the current operating range of persistent visual pseudonymization and as motivation for improving robustness under broader deployment conditions.

Persistent pseudonymization also provides a different privacy–utility trade-off from approaches that simply suppress access to visual content. By maintaining a semantically consistent substitute fact, the protected image can continue to support verification, comparison, and downstream reasoning without exposing the original fact, which may be valuable for applications that still require meaningful multimodal interaction. More broadly, however, any image-side mechanism that intentionally influences model interpretation can have uses beyond privacy protection. FACTPATCH is therefore intended as a user-controlled defensive mechanism for protecting selected visual information, and its deployment in high-stakes settings should be complemented by appropriate provenance, auditing, access-control, or human-verification mechanisms. Future work may extend this framework to more complex and distributed visual facts, stronger transformations, and a broader range of multimodal models and deployment environments.