AUTHORIZATION-CONDITIONED CAPABILITY ISOLATION FOR COLLABORATIVE MLLM INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative MLLM inference creates a mismatch between what a remote model can infer from a transmitted representation and what is authorized under the specific request. Existing privacy-oriented approaches primarily aim to reduce input reconstruction or semantic leakage, but do not explicitly control which capabilities a deployed model may exercise under request-level authorization. This work formulates authorization-conditioned capability isolation as transforming a shared visual representation so that a denied target contributes no additional capability beyond the model's No-Image behavior, while non-target capabilities remain close to their Raw behavior under the same transmitted state. We introduce an Authorization-Conditioned Bottleneck (ACB) that realizes this objective through policy-conditioned representation modulation, including token scaling, channel reweighting, rank-limited correction, and fixed-representation off-target preservation. In the primary evaluation on our constructed AuthCap, a 13-capability authorization benchmark, ACB achieves authorized utility of 0.8194 while reducing denied target capability to within 0.0007 of the No-Image reference. With off-target preservation, unrelated capability retention improves from 0.1400 to 0.5479. These results demonstrate the feasibility of request-level capability isolation: suppressing unauthorized capabilities while substantially retaining unrelated visual capabilities.

1 INTRODUCTION

Collaborative inference provides a practical way to deploy multimodal large language models (MLLMs) when clients cannot, or do not wish to, execute the full model locally. Modern MLLMs such as LLaVA and Qwen2.5-VL transform images into intermediate visual representations that support heterogeneous downstream abilities, including document understanding, recognition, and visual reasoning (Liu et al., 2023; Bai et al., 2025). In a client–cloud setting, however, withholding the raw image does not guarantee that the transmitted representation exposes only the capabilities authorized for a particular request.

This risk is well established in split and collaborative inference. Model-inversion and reconstruction attacks recover private inputs from intermediate activations (He et al., 2019; Erdoğan et al., 2022), while later work develops stronger reconstruction attacks and broader split-inference attack benchmarks under different attacker knowledge and system assumptions (Li et al., 2023; Xu et al., 2024; Singh et al., 2024). In vision-language settings, leakage further extends beyond pixel reconstruction: intermediate visual features can reveal labels, captions, semantic structure, and privacy-relevant content (Xiu & Zhang, 2025; Chen et al., 2026; Jin et al., 2026). These studies establish the transmitted representation as a critical protection boundary, but they do not address a different question: given that a representation must remain useful, which capabilities should a remote model be allowed to exercise under the current authorization?

Consider a document containing a phone number, date, and transaction amount. A request may deny phone extraction while still permitting reasoning about the amount. Completely suppressing the visual representation would satisfy the denied request, but would also remove unrelated visual capabilities. Authorization should therefore be judged jointly by denied-target suppression and the useful capability that remains available from the same transmitted state.

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

We formulate this problem as *authorization-conditioned capability isolation*. For a target capability c, the client produces a policy-conditioned representation $\widetilde{Z}_c^-$ such that the frozen cloud MLLM approaches its No-Image behavior on the denied target query, while other supervised capabilities on the same image remain close to their behavior under the unprotected representation. Crucially, the protected representation $\widetilde{Z}_c^-$ is computed once from the target query and authorization state and then reused for off-target queries, preventing preservation from being achieved by regenerating a different representation for each query.

To instantiate this objective, we introduce the *Authorization-Conditioned Bottleneck (ACB)* at the trusted client-side visual-representation boundary. ACB learns a lightweight policy-conditioned representation transformation while keeping the cloud MLLM frozen. Unlike conventional bottlenecks that reduce representational capacity, ACB constrains the capabilities exposed through the transmitted state, enabling capability-level control rather than indiscriminate information suppression.

We further construct *AuthCap*, a 13-capability benchmark designed to evaluate request-level authorization through matched allow/deny counterfactuals. On Qwen2.5-VL-7B, ACB achieves authorized utility of 0.8194, reduces denied target capability to within 0.0007 of the No-Image reference, and raises off-target retention from 0.1400 without preservation supervision to 0.5479. A full 13-capability cross-backbone evaluation on Molmo2-O-7B further demonstrates that the isolation objective transfers across models, while revealing the trade-off between suppression strength and capability preservation.

Our contributions can be summarized as:

(i) We formulate, for the first time, request-level authorization-conditioned capability isolation for collaborative MLLM inference at the transmitted visual-representation boundary, where denied capabilities should be suppressed toward the No-Image behavior while unrelated capabilities are preserved under the same protected state.

(ii) We introduce Authorization-Conditioned Bottleneck (ACB), a lightweight authorization-conditioned representation transformation with policy-conditioned token scaling, channel reweighting, rank-limited correction, and fixed-representation preservation of off-target capabilities.

(iii) We construct AuthCap, a 13-capability benchmark with matched allow/deny counterfactuals for evaluating authorization-conditioned capability isolation, and validate the formulation through held-out capability matrices, full cross-backbone evaluation, fixed-representation paraphrasing, and adaptive probing.

2 Related Work

2.1 Intermediate-Representation Leakage

Prior work on split and collaborative inference shows that communicated activations can support input reconstruction and feature inversion under different attacker knowledge assumptions (He et al., 2019; Erdoğan et al., 2022; Li et al., 2023; Xu et al., 2024; Singh et al., 2024). Recent vision-language studies further demonstrate recovery of labels, captions, and semantic information from intermediate visual representations (Xiu & Zhang, 2025; Chen et al., 2026; Jin et al., 2026). These works characterize what information remains recoverable from a shared representation; the present work instead asks which capabilities should remain behaviorally available under an explicit request-level authorization state.

2.2 Representation Privacy

Representation defenses typically optimize a privacy–utility trade-off. Shredder learns noise distributions for communicated activations (Mireshghallah et al., 2020), task-aware privacy methods retain task utility under local differential privacy (Cheng et al., 2022), and Split-and-Denoise perturbs client-side LLM embeddings before cloud inference (Mai et al., 2024). PrivQA further studies selective protection of personal-information categories through access-control instructions and response-level self-moderation (Chen et al., 2023). In contrast, ACB implements authorization control at

the transmitted visual-representation boundary, requiring denied-target suppression and non-target preservation to hold on the same protected state.

2.3 Visual-Token Reduction and Selection

Recent MLLM efficiency methods reduce redundant or query-irrelevant visual tokens. LVPruning uses language guidance to estimate token importance (Sun et al., 2025), SparseVILA combines visual sparsity with query-aware retrieval (Khaki et al., 2025), and VisionZip selects informative tokens to reduce redundancy (Yang et al., 2025). Their objective is efficiency or query relevance rather than authorization: relevance alone does not distinguish otherwise identical requests with different policies. ACB complements these approaches by making authorization an explicit input to the representation transformation.

3 Problem Formulation and Threat Model

3.1 Collaborative Inference Setting

Let x denote an input image and E the trusted client-side visual encoder. The resulting visual representation is

$$Z = E(x) \in \mathbb{R}^{N \times D}, \tag{1}$$

where N is the number of visual tokens and D is the visual-feature dimension. A frozen cloud MLLM F produces an output from a transmitted visual state and a query q.

Let $\mathcal{C}$ denote the evaluated capability set. For a capability $c \in \mathcal{C}$, q_c denotes a query that exercises c, and $\pi_c \in \{\texttt{allow}, \texttt{deny}\}$ denotes its authorization state. In the implementation, the capability–authorization pair (c, π_c) is represented by a capability-specific policy identifier. Before transmission, the client computes

$$\widetilde{Z}_c = G_\theta(Z, q_c, \pi_c), \tag{2}$$

where G_θ is the trainable authorization-conditioned representation transformation parameterized by θ, and $\widetilde{Z}_c \in \mathbb{R}^{N \times D}$ is the protected visual representation transmitted to the cloud.

Raw denotes the unprotected representation Z, while No-Image, denoted Z_0, is the operational reference state in which the cloud receives no usable visual evidence.

3.2 Authorization-Conditioned Capability Isolation

Let $M_c(\cdot)$ denote the task-specific evaluation function for capability c, applied to the frozen cloud model's output. Define the protected representations under the two authorization states as

$$\widetilde{Z}_c^{+} = G_\theta(Z, q_c, \texttt{allow}), \qquad \widetilde{Z}_c^{-} = G_\theta(Z, q_c, \texttt{deny}).$$

Under `allow`, the protected state should preserve Raw-supported behavior,

$$M_c\Big(F(\widetilde{Z}_c^{+}, q_c)\Big) \approx M_c(F(Z, q_c))\,. \tag{3}$$

Under `deny`, the target capability should approach its No-Image behavior,

$$M_c\Big(F(\widetilde{Z}_c^{-}, q_c)\Big) \approx M_c(F(Z_0, q_c))\,. \tag{4}$$

Target suppression alone can be achieved by indiscriminately removing visual evidence. For an off-target capability $j \in \mathcal{C} \setminus \{c\}$, let q_j denote a query that exercises capability j. The same denied representation should retain the corresponding Raw-supported behavior:

$$M_j\Big(F(\widetilde{Z}_c^{-}, q_j)\Big) \approx M_j(F(Z, q_j))\,. \tag{5}$$

Importantly, $\widetilde{Z}_c^{-}$ is computed once from (Z, q_c, π_c) and reused for all q_j. Eqs. 3–5 therefore require target suppression and off-target preservation to hold on the same transmitted state. This definition is behavioral and does not assume a uniquely identifiable capability subspace.

3.3 Evaluation Quantities

For a denied target capability c, Unauthorized Query Gain (UQG) measures the target-capability difference relative to the No-Image reference:

$$\mathrm{UQG}_c = M_c\Big(F(\widetilde{Z}_c^-, q_c)\Big) - M_c(F(Z_0, q_c))\,. \tag{6}$$

Thus, $\mathrm{UQG}_c \approx 0$ means that the transmitted state contributes little additional target capability; it does not imply zero absolute accuracy.

For an off-target capability j, define its Raw visual gain as

$$\Delta_j^{\mathrm{Raw}} = M_j(F(Z, q_j)) - M_j(F(Z_0, q_j)). \tag{7}$$

For informative pairs with a non-negligible positive Raw–No-Image gap, normalized off-target retention is

$$R_{c\to j} = \frac{M_j(F(\widetilde{Z}_c^-, q_j)) - M_j(F(Z_0, q_j))}{\Delta_j^{\mathrm{Raw}}}. \tag{8}$$

The experimental evaluation uses $\Delta_j^{\mathrm{Raw}} > 0.05$ for this pair-level normalization. Retention near 1 is Raw-like and near 0 is No-Image-like. The ratio is not clipped, so values below 0 or above 1 are retained.

3.4 Threat Model and Policy Scope

The client is trusted and enforces policy before transmission; the remote service executes a fixed deployed MLLM and may issue alternative queries against the same protected state. The policy interface is closed-set and capability-specific rather than a general compositional access-control language. Separately trained decoders are evaluated as adaptive leakage probes, not as part of the deployed-model guarantee, separating behavioral capability isolation from representation-level information erasure.

4 Authorization-Conditioned Bottleneck

The Authorization-Conditioned Bottleneck (ACB) implements the capability-isolation objective in Section 3 by transforming the client-side visual representation before transmission. Given visual tokens, a query, and an explicit authorization state, ACB combines token scaling, channel reweighting, and a rank-limited subtractive correction. The output has the same shape as the original visual representation; "bottleneck" therefore refers to restricting the capabilities exposed through the transmitted state rather than reducing its dimensionality. Training couples target suppression with preservation of supervised off-target capabilities on the *same* protected representation.

4.1 Policy-Conditioned Representation Transformation

Let $Z = [z_1, \ldots, z_N]^\top \in \mathbb{R}^{N\times D}$ denote the visual-token matrix at the client-side intervention boundary. A query vector $\bar{q} \in \mathbb{R}^D$ is obtained by mean-pooling the frozen language-model embeddings of the query tokens and cropping or zero-padding the result to the visual dimension.

The capability–authorization pair (c, π_c) is represented by a learned policy embedding $e_{c,\pi_c} \in \mathbb{R}^{d_h}$, where d_h denotes the shared conditioning dimension. Learned projections $W_z, W_q \in \mathbb{R}^{d_h\times D}$ map visual-token and query features into this conditioning space.

Token scaling. For each visual token z_i, ACB predicts a soft token score

$$s_i = \sigma(f_{\mathrm{tok}}(W_z z_i + W_q \bar{q} + e_{c,\pi_c}))\,, \qquad s_i \in (0,1), \tag{9}$$

where $f_{\mathrm{tok}} : \mathbb{R}^{d_h} \to \mathbb{R}$ is the token-scoring network and σ denotes the logistic sigmoid. We write $s = [s_1, \ldots, s_N]^\top \in (0,1)^N$ for the resulting token-weight vector. Unlike hard pruning, token scaling does not change the transmitted token count.

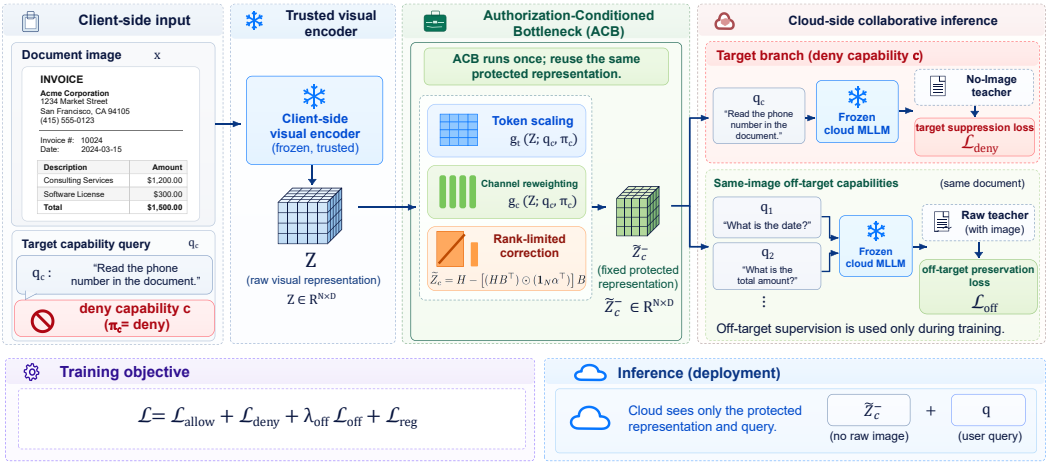

Figure 1: **Authorization-Conditioned Bottleneck (ACB).** The trusted client conditions ACB on the target query and authorization state to produce a protected visual representation. For a denied target capability, the representation is computed once and reused by both the target query and same-image off-target queries during training. The target branch matches the No-Image reference, while the off-target branch preserves Raw-supported capabilities. The allow branch, which matches Raw behavior, is omitted for clarity. At deployment, only the protected representation and the requested query are transmitted to the frozen cloud MLLM.

Channel reweighting. The same query–policy context produces a channel-weight vector

$$g = \sigma(f_{\text{ch}}(W_q \bar{q} + e_{c,\pi_c})) \in (0,1)^D, \tag{10}$$

where $f_{\text{ch}} : \mathbb{R}^{d_h} \to \mathbb{R}^D$ is the channel-scoring network. Token and channel controls are applied jointly:

$$H = Z \odot (s\mathbf{1}_D^\top) \odot (\mathbf{1}_N g^\top), \tag{11}$$

where $\mathbf{1}_N$ and $\mathbf{1}_D$ denote all-ones vectors of lengths N and D, respectively, $\odot$ denotes element-wise multiplication, and $H \in \mathbb{R}^{N \times D}$ is the token- and channel-reweighted visual representation.

Rank-limited correction. Token and channel reweighting operate independently across token and feature coordinates. ACB therefore adds a compact query–policy-conditioned correction in feature space. Let

$$h = W_q \bar{q} + e_{c,\pi_c} \tag{12}$$

denote the shared query–policy conditioning context. From h, the module predicts a row-normalized basis $B \in \mathbb{R}^{r \times D}$ and removal strengths $\alpha \in (0,1)^r$, where r is the correction rank. For the current capability–authorization pair, the protected representation is

$$\tilde{Z}_c = H - \left[(HB^\top) \odot (\mathbf{1}_N \alpha^\top)\right] B. \tag{13}$$

The rows of B are normalized independently but are not constrained to be orthogonal. Eq. 13 is therefore a rank-limited subtractive correction, not an exact orthogonal projection; r bounds intervention capacity without implying that capability information is intrinsically r-dimensional.

4.2 Authorization and Preservation Objective

Training uses same-image authorization counterfactuals. For capability c, let y_c^{Raw} denote the reference answer produced by the frozen MLLM from the unprotected representation, and let y_c^0 denote the corresponding No-Image reference answer. Here, CE denotes teacher-forced autoregressive cross-entropy over the reference-answer tokens.

The authorized branch is trained to preserve Raw-supported behavior:

$$\mathcal{L}_{\text{allow}} = \text{CE}\Big(F(\tilde{Z}_c^+, q_c), y_c^{\text{Raw}}\Big). \tag{14}$$

Table 1: AuthCap composition in matched allow/deny pairs

Group	Capabilities	Train	Val	Test
Textual PII	phone, name, email, address	640	192	320
Document fields	order ID, date, amount	480	144	240
Document/structure	template ID, document presence	320	96	160
Non-text semantics	person, vehicle, food, animal presence	631	186	263
Total	13 capabilities	2,071	618	983

The denied target branch is trained toward the No-Image reference:

$$\mathcal{L}_{\text{deny}} = \text{CE}\Big(F(\widetilde{Z}_c^-, q_c), y_c^0\Big). \tag{15}$$

For training examples with available off-target supervision, let $\mathcal{J}_K(x,c) \subseteq \mathcal{C} \setminus \{c\}$ denote a nonempty set of up to K supervised off-target capabilities for the same image x, and let y_j^{Raw} denote the Raw reference answer for q_j. ACB reuses the same differentiable $\widetilde{Z}_c^-$ for these queries:

$$\mathcal{L}_{\text{off}} = \frac{1}{|\mathcal{J}_K(x,c)|} \sum_{j \in \mathcal{J}_K(x,c)} \text{CE}\Big(F(\widetilde{Z}_c^-, q_j), y_j^{\text{Raw}}\Big). \tag{16}$$

The protected state is not recomputed for q_j; off-target preservation is therefore learned as a property of the transmitted representation rather than through query-specific regeneration. The sampled off-target capability identities are used only to construct $\mathcal{L}_{\text{off}}$ and are not supplied as additional deployment-time inputs.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{allow}} + \mathcal{L}_{\text{deny}} + \lambda_{\text{off}}\mathcal{L}_{\text{off}} + \mathcal{L}_{\text{reg}}, \tag{17}$$

where $\lambda_{\text{off}} \geq 0$ controls the strength of fixed-representation preservation and $\mathcal{L}_{\text{reg}}$ is a lightweight gate regularizer that favors high token retention for authorized requests and suppression for denied requests. The preservation count K, correction rank r, and λ_{off} are selected before held-out evaluation as described in Section 5.

4.3 Deployment

Only ACB is trained; the visual encoder and cloud MLLM remain frozen. At inference time, the client computes the visual representation, forms the query–policy-conditioned ACB state, applies Eqs. 9–13, and transmits the resulting $\widetilde{Z}_c$ to the cloud. The deployment path therefore requires only the current query and its authorization state. The off-target queries used in Eq. 16 are training signals and introduce no additional inference-time oracle or capability list.

5 Experimental Setup

5.1 AuthCap benchmark

AuthCap evaluates request-level capability isolation using matched authorization counterfactuals. Its 13 capabilities span textual PII, document fields and structure, and non-text visual semantics; each image–query pair is instantiated under both `allow` and `deny`. To reduce dependence on OCR-style tasks, three non-text presence fields (vehicle, food, and animal) are added through validation screening: balanced or low-prior candidates are retained when frozen-Qwen Raw behavior clearly exceeds No-Image and matched counterfactuals can be constructed. Together with person presence, they form the non-text semantic group. This field-level screening is distinct from the pair-level Raw–No-Image gap > 0.05 filter used for normalized retention.

AuthCap contains 2,071/618/983 matched train/validation/test pairs (4,142/1,236/1,966 requests). Document-oriented tasks use synthetic images, while person presence uses COCO imagery (Lin

et al., 2014); Appendix A gives the full construction protocol. The held-out test is not used for model selection.

5.2 Models and implementation

The primary evaluation uses frozen Qwen2.5-VL-7B (Bai et al., 2025). ACB is inserted at the visual-merger input, where the primary representation contains 2,916 tokens of dimension 1,280. Only ACB is optimized; the visual encoder and cloud MLLM remain frozen. Training uses teacher-forced cross-entropy to the Raw reference for authorized requests, the No-Image reference for denied targets, and the Raw reference for supervised off-target capabilities as defined in Section 4.2. Optimization uses AdamW for 150 steps with learning rate 5×10^{-4} and a deterministic field-balanced sampler.

Validation selects $r = 2$ and $\lambda_{\text{off}} = 0.10$ under prespecified macro- and worst-field-UQG constraints, with $K = 2$ same-image off-target queries. The resulting configuration is frozen before held-out evaluation and evaluated over three seeds; full selection sweeps are reported in Appendix B.

For cross-backbone evaluation, frozen Molmo2-O-7B (Clark et al., 2026) uses the same AuthCap splits and directly inherits $r = 2$, $\lambda_{\text{off}} = 0.10$, and $K = 2$ without Molmo-specific tuning. Its fixed-representation evaluation uses up to 16 examples per target and the same Raw–No-Image gap > 0.05 informative-pair filter.

5.3 Baselines and metrics

Raw and No-Image define the two behavioral reference points. Policy-Switch Raw/No-Image uses Raw for authorized requests and No-Image for denied requests, providing a direct reference for target suppression without off-target preservation. ACB w/o off-target preservation keeps the same policy-conditioned token, channel, and rank-limited transformation as ACB but sets $\lambda_{\text{off}} = 0$.

We also compare against a TrustCLIP-style reconstruction-oriented projection baseline (Athanasiou et al., 2026), a query-guided visual-token pruning baseline following relevance-based token selection (Sun et al., 2025; Khaki et al., 2025), and token-only and token+channel ACB variants. Together, these baselines cover reconstruction-oriented representation protection, query-relevance pruning, and progressively richer forms of the proposed bottleneck.

The primary metrics follow Section 3. AU is the macro-averaged score under `allow`; target UQG is the macro-averaged denied-minus-No-Image score; worst-field UQG is the maximum field-level UQG; and normalized off-target retention follows Eq. 8.

6 Results

6.1 Primary AuthCap results on Qwen2.5-VL-7B

On the primary Qwen2.5-VL-7B backbone, Figure 2 summarizes the held-out 13-capability AuthCap evaluation. ACB achieves authorized utility of 0.8194 while keeping the denied target close to the No-Image reference (target UQG $= 0.0007$). Off-target retention increases from 0.1400 without preservation supervision to 0.5479 with the complete objective; the median off-target retention is 0.8580, and the fraction of informative pairs with negative retention falls to 0.0250.

The Policy-Switch control illustrates why target suppression alone is insufficient for capability isolation: it attains zero target UQG but retains no off-target visual capability in the denied state.

6.2 Comparison with representation-control baselines

Table 2 compares ACB with representative representation-control baselines. Query-guided pruning lowers UQG to 0.4320 but reduces authorized utility to 0.6076, while the TrustCLIP-style, token-only, and token+channel variants retain substantial denied capability. ACB achieves AU $= 0.8194$ with UQG $= 0.0007$, indicating that representation protection or query relevance alone does not realize request-level capability isolation.

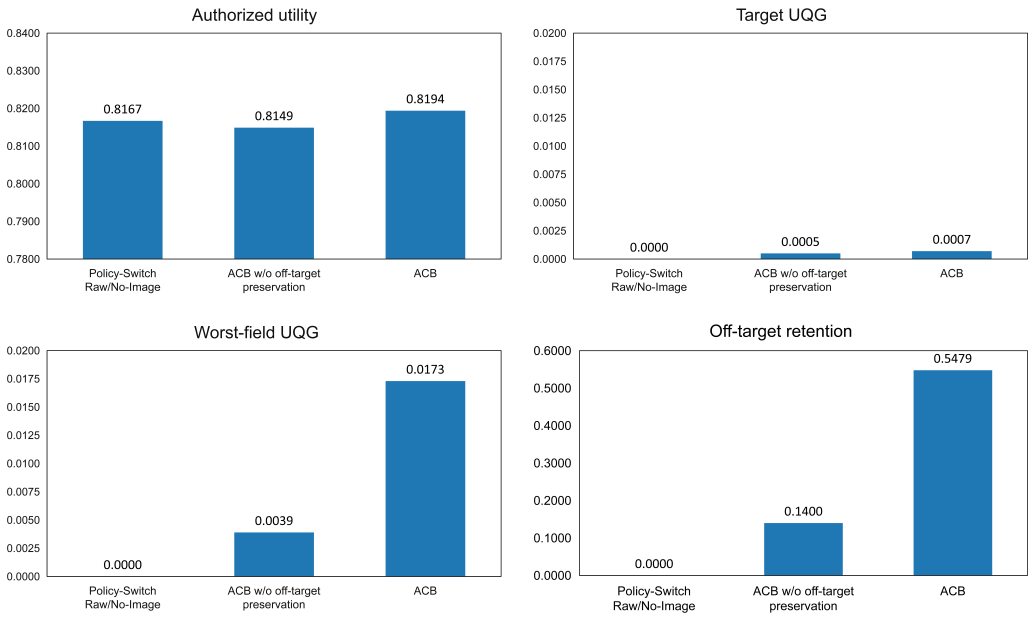

Figure 2: Held-out results on Qwen2.5-VL-7B over the primary 13-capability AuthCap evaluation. Authorized utility, target UQG, and worst-field UQG use the canonical evaluation; off-target retention uses the fixed-representation capability matrix. ACB substantially improves off-target retention while maintaining near-zero incremental target capability. Policy-Switch Raw/No-Image attains target suppression by construction but has zero off-target retention.

Table 2: Comparison with representation-control baselines on the Qwen2.5-VL-7B AuthCap evaluation

Method	Runs	AU ↑	UQG ↓
Query-guided pruning	1	0.6076	0.4320
TrustCLIP-style	1	0.8175	0.6419
Token only	3	0.8187	0.6430
Token + channel	3	0.8069	0.6122
ACB	3	**0.8194**	**0.0007**

6.3 Contribution of fixed-representation preservation

On Qwen2.5-VL-7B, within the evaluated ACB configuration, fixed-representation preservation is important for retaining unrelated capabilities after denial. With $\lambda_{\text{off}} = 0$, off-target retention is 0.1400 at a canonical target UQG of 0.0005; adding the fixed-representation preservation loss $\mathcal{L}_{\text{off}}$ in Eq. 16 raises retention to 0.5479 while canonical target UQG remains 0.0007. The paired off-target retention gain is $+0.4079$ (95% CI $[0.3939, 0.4221]$).

6.4 Cross-backbone evaluation

Table 3 evaluates the same 13-capability AuthCap protocol on Qwen2.5-VL-7B and Molmo2-O-7B, with the Qwen-selected configuration transferred to Molmo without backbone-specific tuning. On Qwen, off-target retention increases from 0.1400 to 0.5479 while target UQG remains near zero. On Molmo, retention increases from 0.2822 to 0.4980, while target UQG changes from 0.0514 to 0.0712. The paired Molmo retention gain is $+0.2158$ (95% CI $[0.2003, 0.2312]$). These results support cross-backbone transfer of the isolation objective while revealing a backbone-dependent trade-off between suppression strength and capability preservation; detailed statistics are reported in Appendix D.

Table 3: Cross-backbone evaluation on the same 13-capability AuthCap benchmark. Off-target retention uses the same fixed-representation `max_per_target` $= 16$ and Raw–No-Image gap > 0.05 protocol; informative pairs are determined separately for each backbone from its own Raw–No-Image gaps.

Backbone	Method	AU ↑	Target UQG ↓	Off-target ret. ↑
Qwen2.5-VL-7B	ACB w/o off-target preservation	0.8149	0.0005	0.1400
Qwen2.5-VL-7B	ACB	**0.8194**	0.0007	**0.5479**
Molmo2-O-7B	ACB w/o off-target preservation	0.6597	0.0514	0.2822
Molmo2-O-7B	ACB	0.6519	0.0712	**0.4980**

6.5 Fixed-representation robustness and adaptive probing

On the primary Qwen2.5-VL-7B backbone, each denied representation is kept fixed while only the subsequent query wording changes. On the nine-capability paraphrase subset, the subset-specific canonical-query UQG is 0.0211 ± 0.0107, mean paraphrase UQG is 0.0265 ± 0.0122, and the robust-denial rate is 0.7685 ± 0.0262; Appendix E reports worst-case statistics and representative failures.

For person presence, an adaptive probe reaches 0.8087 ± 0.0413 on ACB-denied states, versus 0.8798 ± 0.0095 on Raw and 0.4754 on No-Image. The residual decodability distinguishes behavioral capability control from representation-level information erasure.

6.6 Efficiency

On Qwen2.5-VL-7B, ACB adds 1,652,227 trainable parameters, corresponding to 0.01993% of the frozen backbone; the rank-2 correction accounts for approximately 29.86M theoretical low-rank FLOPs per sample. The measured bottleneck-only latency is 0.000488 ± 0.000012 s per call. End-to-end latency increases from 0.3590 ± 0.0029 s for Raw inference to 0.4010 ± 0.0007 s with ACB, an 11.69% overhead in the matched profiling setup.

7 Discussion and Limitations

Across Qwen2.5-VL-7B and Molmo2-O-7B, the results support the feasibility of request-level capability isolation while showing that the suppression–preservation operating point depends on the backbone and representation interface. The present evaluation addresses closed-set behavioral authorization rather than representation-level information erasure; additional robustness and probing results are provided in the Appendix.

8 Conclusion

This work formulates authorization-conditioned capability isolation for collaborative MLLM inference and introduces ACB and the AuthCap benchmark. On Qwen2.5-VL-7B, ACB preserves authorized utility while suppressing denied target capability toward No-Image and substantially improving off-target retention. Molmo2-O-7B further supports cross-backbone transfer of the isolation objective while revealing a backbone-dependent suppression–preservation trade-off. Together, these results demonstrate the feasibility of request-level capability isolation and motivate stronger robustness and compositional authorization.

AI use statement

In this work, generative AI tools were used solely for grammar checking and sentence-level language polishing to improve the readability of the manuscript. All AI-assisted edits were reviewed by the authors to ensure that they did not alter the scientific content, technical claims, experimental results, citations, or conclusions. The authors take full responsibility for the final content of this work, including any text edited with the assistance of generative AI.

Ethics statement

This work investigates authorization and information-exposure risks in collaborative multimodal large language model inference and is intended for defensive research. The experiments are conducted in controlled settings using benchmark data and publicly available model checkpoints, without unauthorized access to private user data or third-party systems. The proposed method provides behavioral capability control rather than cryptographic confidentiality or guaranteed removal of sensitive information, and the paper explicitly discusses this distinction.

Reproducibility statement

The paper and appendix provide the model configurations, representation boundaries, benchmark construction and screening procedure, training settings, random seeds, model-selection protocol, evaluation metrics, held-out results, robustness analyses, and cross-backbone evaluation needed to reproduce the reported experiments. All test results use configurations fixed before held-out evaluation. The supplementary materials include the source code, processed benchmark data, experiment configurations, and evaluation scripts used to produce the reported results. These materials will be made publicly available upon acceptance.

References

Nikos Athanasiou, Ilya A. Petrov, Angela Yao, Shugao Ma, Eric Sauser, Edoardo Remelli, Shreyas Hampali, Johannes Schönberger, Fadime Sener, and Bugra Tekin. TrustCLIP: Learning private visual features via adversarial reconstruction. In *European Conference on Computer Vision*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. Can language models be instructed to protect personal information? *arXiv preprint arXiv:2310.02224*, 2023.

Yiyi Chen, Qiongkai Xu, Desmond Eliott, Qiongxiu Li, and Johannes Bjerva. Semantic leakage from image embeddings. *arXiv preprint arXiv:2601.22929*, 2026.

Jiangnan Cheng, Ao Tang, and Sandeep Chinchali. Task-aware privacy preservation for multi-dimensional data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3835–3851. PMLR, 2022.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Ege Erdoğan, Alptekin Küpçü, and A. Ercüment Çiçek. UnSplit: Data-oblivious model inversion, model stealing, and label inference attacks against split learning. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society*, pp. 115–124. Association for Computing Machinery, 2022. doi: 10.1145/3559613.3563201.

Zecheng He, Tianwei Zhang, and Ruby B. Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 148–162. Association for Computing Machinery, 2019. doi: 10.1145/3359789.3359824.

Shuaifan Jin, Zhibo Wang, Qiyuan Wang, Yiting Han, Yajie Zhou, Yuanfan Zhang, Jiahui Hu, and Xiaoyi Pang. Inverting the hidden: Unveiling multimodal privacy leakage in collaborative LVLM inference. *arXiv preprint arXiv:2608.01020*, 2026.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N. Plataniotis, Yao Lu, Song Han, and Zhijian Liu. SparseVILA: Decoupling visual sparsity for efficient vlm inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23784–23794, 2025.

Ziang Li, Mengda Yang, Yaxin Liu, Juan Wang, Hongxin Hu, Wenzhe Yi, and Xiaoyang Xu. GAN you see me? enhanced data reconstruction attacks against split inference. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2377.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1516.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-denoise: Protect large language model inference with local differential privacy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34281–34302. PMLR, 2024.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Prakash Ramrakhyani, Ali Jalali, Dean M. Tullsen, and Hadi Esmaeilzadeh. Shredder: Learning noise distributions to protect inference privacy. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 3–18. Association for Computing Machinery, 2020. doi: 10.1145/3373376.3378522.

Abhishek Singh, Vivek Sharma, Rohan Sukumaran, John J. Mose, Jeffrey K. Chiu, Justin Yu, and Ramesh Raskar. SIMBA: Split inference—mechanisms, benchmarks and attacks. In *Computer Vision—ECCV 2024*, pp. 214–232. Springer, 2024.

Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. LVPruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4299–4308. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.242.

Kedong Xiu and Sai Qian Zhang. Caprecover: A cross-modality feature inversion attack framework on vision language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3808–3816, 2025. doi: 10.1145/3746027.3755203.

Xiaoyang Xu, Mengda Yang, Wenzhe Yi, Ziang Li, Juan Wang, Hongxin Hu, Yong Zhuang, and Yaxin Liu. A stealthy wrongdoer: Feature-oriented reconstruction attack against split learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12130–12139, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025. doi: 10.1109/CVPR52734.2025.01843.

A AuthCap construction and protocol

A.1 Capability composition

AuthCap is constructed around matched authorization counterfactuals: the image and capability query are held fixed while the authorization state changes between `allow` and `deny`. The benchmark contains 13 capabilities spanning textual extraction, document semantics and structure, and non-text visual semantics.

The textual-PII capabilities (phone, name, email, and address) use synthetic document/PII examples; order ID, date, and amount use synthetic structured documents; and template ID and document

Table 4: AuthCap capability composition and split sizes in matched allow/deny pairs

Group	Capabilities	Train	Val	Test	Source / construction
Textual PII	phone, name, email, address	640	192	320	Synthetic document/PII images
Document fields	order ID, date, amount	480	144	240	Synthetic structured documents
Document/structure	template ID, document presence	320	96	160	Synthetic layout/template scenarios
Non-text semantics	person, vehicle, food, animal presence	631	186	263	COCO for person presence; screened non-text task extension for the other three
Total	13 capabilities	2,071	618	983	Matched counterfactual pairs

presence use synthetic layout/template scenarios. Person presence is included as an existing non-text semantic capability using COCO imagery (Lin et al., 2014).

To broaden the benchmark beyond OCR- and text-reading capabilities, AuthCap further includes a screened non-text task extension. Candidate non-text tasks are constructed as balanced yes/no tasks or otherwise low-prior visual tasks and evaluated with the frozen Qwen backbone under Raw and No-Image conditions. A candidate is retained only when (i) visual evidence produces a clear Raw–No-Image capability advantage and (ii) matched same-image/same-query `allow`/`deny` counterfactuals can be constructed. Candidates that do not satisfy these criteria are discarded. This process retains three additional non-text capabilities: `object_presence_vehicle`, `object_presence_food`, and `object_presence_animal`. Together with `person_presence`, they form the non-text semantic group.

Table 4 summarizes the resulting capability composition and split sizes.

The corresponding expanded request counts are 4,142 training, 1,236 validation, and 1,966 held-out test requests. Each matched pair contributes one `allow` and one `deny` request.

The benchmark-level capability screening described above is distinct from the fixed-representation matrix filter. The former determines which capability fields are included in AuthCap, whereas the latter retains only individual target–off-target pairs with Raw–No-Image gap greater than 0.05 when computing normalized retention, in order to avoid unstable normalization in Eq. 8.

A.2 Policy construction and evaluation subsets

The primary policy space contains 26 learned states corresponding to 13 capabilities $\times$ {`allow`, `deny`}. Capability identity is therefore explicit policy input. This is a closed-set policy parameterization rather than a compositional language for arbitrary simultaneous allow/deny combinations.

For the fixed-representation target-to-off-target analysis, one denied representation is computed from the canonical denied target query and reused for other supervised capabilities on the same image. Only pairs with Raw–No-Image gap greater than 0.05 are included in normalized-retention summaries, because Eq. 8 is unstable when its denominator is near zero. Missing same-image supervision is treated as missing rather than zero.

B Implementation and model selection

B.1 Primary implementation details

On the primary backbone, ACB operates at the visual-merger input on 2,916 visual tokens of dimension 1,280. The query representation is the mean frozen language-model embedding of the query tokens, cropped or zero-padded to the 1,280-dimensional visual space. A learned policy embedding and the projected query form a shared 256-dimensional conditioning context for token scaling, channel reweighting, and the rank-limited correction.

Reference answers are generated once by the frozen model and reused during training. Teacher-forced cross-entropy has weight 1.0 and logit-level KL matching has weight 0. The final checkpoints use gate regularization coefficients of 0.01 for allowed retention and denied sparsity. Losses are normalized across field/group cells before averaging with equal group weights. Training uses AdamW for 150 steps with learning rate 5×10^{-4} and deterministic cycled field-balanced pairing. The cloud MLLM and visual backbone remain frozen.

Table 5: Validation-only rank selection over three seeds

Rank	AU ↑	Denied ↓	UQG ↓	Worst UQG ↓
0	0.8208 ± 0.0009	0.8200 ± 0.0009	0.6503 ± 0.0009	1.0000
1	0.8180 ± 0.0025	0.1736 ± 0.0055	0.0039 ± 0.0055	0.1250
2	0.8195 ± 0.0043	0.1698 ± 0.0001	0.0001 ± 0.0001	0.0028
4	0.8196 ± 0.0026	0.1700 ± 0.0003	0.0003 ± 0.0003	0.0058
8	0.8206 ± 0.0046	0.1697 ± 0.0000	0.0000 ± 0.0000	0.0006
16	0.8251 ± 0.0016	0.1699 ± 0.0001	0.0002 ± 0.0001	0.0044

Table 6: Validation sweep for the off-target preservation weight

λ_{off}	AU ↑	Target UQG ↓	Worst UQG ↓	Off-target retention ↑	Selectivity gap ↑
0.00	0.8203	-0.0030	0.0056	0.0913	0.0915
0.05	0.8203	0.0068	0.0893	0.4477	0.4378
0.10	0.8203	-0.0004	0.0071	0.5170	0.5160
0.25	0.8184	0.0191	0.1667	0.6349	0.6171
0.50	0.8203	0.0241	0.1799	0.6049	0.5367
0.75	0.8203	0.0106	0.1215	0.7125	0.6813
1.00	0.8235	0.0540	0.3542	0.7649	0.6865

B.2 Rank selection

The low-rank capacity is selected using validation data only. Table 5 reports the validation sweep. Rank 0 fails to suppress the denied capability, while rank 1 substantially reduces macro UQG but retains a worst-field leakage spike. Rank 2 is the smallest tested setting with stable near-zero macro and worst-field UQG and is therefore fixed for subsequent experiments.

This sweep is used as capacity selection, not as an estimate of an intrinsic capability dimension. The learned correction is query- and policy-conditioned, and the selected rank only bounds the feature-space intervention available to ACB.

B.3 Off-target preservation weight

With rank fixed to two, the λ_{off} sweep in Table 6 is conducted on the validation split using seed 20260821. The constrained selection requires macro target UQG ≤ 0.02 and worst-field UQG ≤ 0.10; among eligible nonzero settings, $\lambda_{\text{off}} = 0.10$ gives the highest off-target retention.

The final training configuration uses $r = 2$, $\lambda_{\text{off}} = 0.10$, and $K = 2$ same-image off-target queries. A three-seed validation confirmation is performed before the held-out test.

C Full held-out capability-isolation results

C.1 Canonical authorization results

Authorized retention normalizes allowed utility between the two references:

$$R_{\text{auth}} = \frac{\text{AU} - \text{AU}_0}{\text{AU}_{\text{Raw}} - \text{AU}_0},$$

where AU_0 denotes No-Image authorized utility.

Table 8 gives the held-out 13-capability canonical results. The Policy-Switch Raw/No-Image control uses Raw for allowed requests and No-Image for denied requests. ACB w/o off-target preservation uses the same architecture as ACB with $\lambda_{\text{off}} = 0$.

Canonical target UQG is computed from the protected and No-Image scores paired within the same evaluation run and then macro-averaged across capabilities.

Table 7: Three-seed validation confirmation for the frozen ACB configuration (population std across seeds).

Seed	AU	Target UQG	Worst-field UQG	Off-target retention	Median off-target
20260821	0.8203	-0.0004	0.0071	0.5170	0.7222
20260822	0.8219	0.0022	0.0531	0.5763	1.0000
20260823	0.8235	0.0020	0.0340	0.5219	0.7222
Mean $\pm$ std	0.8219 ± 0.0013	0.0012 ± 0.0012	0.0314 ± 0.0189	0.5384 ± 0.0269	0.8148 ± 0.1309

Table 8: Held-out canonical authorization results on AuthCap. Target UQG is computed relative to the No-Image reference paired within each canonical evaluation.

Method	AU ↑	Authorized retention ↑	Target UQG ↓	Worst-field UQG ↓
Raw	0.8167	1.0000	0.6411	1.0000
No-Image	0.1756	0.0000	0.0000	0.0000
Policy-Switch Raw/No-Image	0.8167	1.0000	0.0000	0.0000
ACB w/o off-target preservation	0.8149	0.9972	0.0005	0.0039
ACB	**0.8194**	**1.0043**	0.0007	0.0173

C.2 Fixed-representation selectivity

Table 9 evaluates the same denied representation on its target capability (diagonal) and on supervised non-target capabilities from the same image (off-diagonal). Policy-Switch has zero off-target retention by construction because the denied state is No-Image.

For ACB, 57.00% of informative off-target pairs retain more than 0.25 of the Raw-supported capability, 55.08% retain more than 0.50, and 51.18% retain more than 0.75. The fraction with negative retention is 2.50%. The corresponding values for ACB w/o off-target preservation are 14.82%, 13.54%, 12.09%, and 8.56%, respectively.

A paired bootstrap over 5,154 matched off-target pairs gives an ACB minus ACB-w/o-preservation retention difference of $+0.4079$ with 95% CI $[0.3939, 0.4221]$. Over 504 matched target pairs, the target-UQG difference is $+0.0028$ with 95% CI $[0.0005, 0.0058]$.

D Cross-Backbone Evaluation

D.1 Molmo-13 protocol

The cross-backbone experiment repeats the full 13-capability AuthCap evaluation with frozen Molmo2-O-7B. It uses the same benchmark splits as the primary experiment: 4,142 training, 1,236 validation, and 1,966 held-out test requests. The evaluated capabilities are address, amount, date, document presence, email, name, animal presence, food presence, vehicle presence, order ID, person presence, phone, and template ID.

ACB is applied at the pre-projector input to Molmo's image projector, with visual tensor shape 392×1152 in bfloat16. The training configuration follows the primary protocol where possible: same-image/same-query authorization counterfactuals, cycled field-balanced paired sampling, field/group-normalized loss, No-Image matching for denied targets, 150 optimization steps, learning rate 5×10^{-4}, and `max_new_tokens`=32. The Molmo runs use seeds 20260924, 20260925, and 20260926.

The Qwen-selected configuration is transferred directly: $r = 2$, $\lambda_{\text{off}} = 0.10$, and $K = 2$. No Molmo-specific rank, preservation-weight, off-target-count, threshold, or sampler sweep is performed. For the fixed-representation matrix, Molmo also uses the Qwen-matched `max_per_target` $= 16$ protocol and retains only informative pairs with Raw–No-Image gap greater than 0.05.

Table 9: Held-out fixed-representation capability-isolation summary

Method	Diag. retention	Diag. UQG	Off-target retention	Median off-target	Worst off-target	Selectivity gap
Policy-Switch Raw/No-Image	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ACB w/o off-target preservation	0.0001	0.0000	0.1400	0.0000	-0.1960	0.1399
ACB	0.0062	0.0029	**0.5479**	**0.8580**	-0.1874	**0.5417**

Table 10: Molmo2-O-7B validation Raw–No-Image screening on the 13 AuthCap capabilities

Capability	Raw	No-Image	Gap
address	0.3367	0.1161	0.2206
amount	0.9375	0.0000	0.9375
date	0.3990	0.1672	0.2318
document presence	1.0000	0.0000	1.0000
email	0.5878	0.2546	0.3332
name	0.4450	0.0000	0.4450
animal presence	0.9583	0.5000	0.4583
food presence	0.9792	0.5000	0.4792
vehicle presence	0.9583	0.5000	0.4583
order ID	0.9792	0.0000	0.9792
person presence	0.8333	0.4524	0.3810
phone	0.9792	0.0000	0.9792
template ID	0.7708	0.0000	0.7708

D.2 Raw–No-Image capability screening

All 13 capabilities exhibit a positive Raw–No-Image gap on the Molmo validation split, confirming that the cross-backbone evaluation is not dominated by capabilities that the frozen backbone cannot obtain from visual evidence.

D.3 Held-out canonical results

Table 11 reports the full held-out canonical evaluation. ACB w/o off-target preservation already shows weaker target isolation than on the primary backbone: authorized utility is below Raw and target UQG remains 0.0514. Adding fixed-representation preservation slightly lowers authorized utility and raises macro target UQG to 0.0712.

The per-capability results in Table 12 show that the cross-backbone degradation is heterogeneous rather than uniform. Several capabilities remain close to the No-Image reference, whereas address, document presence, email, name, and phone exhibit larger residual UQG under the complete ACB.

D.4 Qwen-matched fixed-representation matrix

The main Molmo matrix uses $\texttt{max_per_target} = 16$, matching the Qwen protocol. After filtering pairs with Raw–No-Image gap ≤ 0.05, the informative evaluation contains 477 diagonal and 4,932 off-diagonal observations per method.

For this matrix only, "excess capability" denotes the unnormalized numerator of Eq. 8, $M_j(F(\widetilde{Z}_c^-, q_j)) - M_j(F(Z_0, q_j))$. On the diagonal ($j = c$), this reduces to target UQG; off the diagonal it is reported only as a diagnostic complement to normalized retention.

The preservation effect is statistically stable under paired analysis. Table 14 uses pairs matched by seed, sample ID, target capability, and off-target capability, with 10,000 paired bootstrap resamples.

These results reveal a backbone-dependent preservation–isolation trade-off under direct hyperparameter transfer. ACB increases off-target retention by $+0.2158$, but diagonal retention and diagonal UQG also increase. The canonical evaluation is consistent with the same pattern: Molmo benefits from fixed-representation preservation, but the Qwen-style near-No-Image target behavior does not transfer under the directly inherited configuration.

Table 11: Molmo2-O-7B held-out canonical results on the full 13-capability AuthCap benchmark

Method	Raw allowed	No-Image denied	AU ↑	Denied ↓	Target UQG ↓
ACB w/o off-target preservation	0.7606	0.1976	0.6597	0.2489	0.0514
ACB	0.7606	0.1976	0.6519	0.2688	0.0712

Table 12: Molmo2-O-7B per-capability held-out results. Values are means across three seeds; $\pm$ standard deviations are shown for learned ACB outputs.

Method	Capability	Raw	No-Image	Allowed	Denied	UQG
ACB w/o off-target preservation	address	0.3108	0.1178	0.4320 ± 0.0139	0.1601 ± 0.0713	0.0424 ± 0.0713
ACB w/o off-target preservation	amount	0.9375	0.0000	0.8208 ± 0.0722	0.0000 ± 0.0000	0.0000 ± 0.0000
ACB w/o off-target preservation	date	0.4022	0.1656	0.3994 ± 0.0081	0.1965 ± 0.0545	0.0309 ± 0.0545
ACB w/o off-target preservation	document presence	1.0000	0.0000	1.0000 ± 0.0000	0.3958 ± 0.5295	0.3958 ± 0.5295
ACB w/o off-target preservation	email	0.6011	0.2603	0.5885 ± 0.0038	0.3618 ± 0.1146	0.1015 ± 0.1146
ACB w/o off-target preservation	name	0.4500	0.0000	0.4237 ± 0.0044	0.0808 ± 0.0614	0.0808 ± 0.0614
ACB w/o off-target preservation	animal presence	0.9219	0.5000	0.8281 ± 0.0312	0.5000 ± 0.0000	0.0000 ± 0.0000
ACB w/o off-target preservation	food presence	0.9138	0.5000	0.8506 ± 0.0359	0.5000 ± 0.0000	0.0000 ± 0.0000
ACB w/o off-target preservation	vehicle presence	0.7125	0.5000	0.7500 ± 0.0661	0.5000 ± 0.0000	0.0000 ± 0.0000
ACB w/o off-target preservation	order ID	0.9375	0.0000	0.6625 ± 0.1090	0.0000 ± 0.0000	0.0000 ± 0.0000
ACB w/o off-target preservation	person presence	0.9508	0.5246	0.6284 ± 0.1394	0.5410 ± 0.0284	0.0164 ± 0.0284
ACB w/o off-target preservation	phone	0.9625	0.0000	0.7458 ± 0.0260	0.0000 ± 0.0000	0.0000 ± 0.0000
ACB w/o off-target preservation	template ID	0.7875	0.0000	0.4458 ± 0.2566	0.0000 ± 0.0000	0.0000 ± 0.0000
ACB	address	0.3108	0.1178	0.4345 ± 0.0126	0.3181 ± 0.1815	0.2003 ± 0.1815
ACB	amount	0.9375	0.0000	0.8750 ± 0.0545	0.0083 ± 0.0144	0.0083 ± 0.0144
ACB	date	0.4022	0.1656	0.4036 ± 0.0011	0.2263 ± 0.0222	0.0606 ± 0.0222
ACB	document presence	1.0000	0.0000	1.0000 ± 0.0000	0.2708 ± 0.2320	0.2708 ± 0.2320
ACB	email	0.6011	0.2603	0.5850 ± 0.0085	0.3913 ± 0.0514	0.1310 ± 0.0514
ACB	name	0.4500	0.0000	0.4142 ± 0.0075	0.1302 ± 0.0407	0.1302 ± 0.0407
ACB	animal presence	0.9219	0.5000	0.8281 ± 0.0563	0.5000 ± 0.0000	0.0000 ± 0.0000
ACB	food presence	0.9138	0.5000	0.7874 ± 0.0263	0.5057 ± 0.0100	0.0057 ± 0.0100
ACB	vehicle presence	0.7125	0.5000	0.6375 ± 0.1269	0.5000 ± 0.0000	0.0000 ± 0.0000
ACB	order ID	0.9375	0.0000	0.6458 ± 0.0753	0.0042 ± 0.0072	0.0042 ± 0.0072
ACB	person presence	0.9508	0.5246	0.6175 ± 0.1609	0.5519 ± 0.0473	0.0273 ± 0.0473
ACB	phone	0.9625	0.0000	0.7417 ± 0.0505	0.0875 ± 0.0760	0.0875 ± 0.0760
ACB	template ID	0.7875	0.0000	0.5042 ± 0.3178	0.0000 ± 0.0000	0.0000 ± 0.0000

Training cost. The additional $K = 2$ off-target scoring increases Molmo training time per step from approximately 1.03–1.05 s for ACB w/o off-target preservation to 1.39–1.41 s for ACB, corresponding to roughly 34–35% training-step overhead. This is training-time cost; the two Molmo variants share the same deployment-time bottleneck architecture.

E Robustness and adaptive probing

E.1 Fixed-representation paraphrases

For each denied example, ACB is run once on the canonical denied query. The resulting protected representation is cached and reused for paraphrased queries; ACB is not re-run after the wording changes. The evaluated capabilities are phone, name, email, address, order ID, date, amount, template ID, and document presence, with at most four paraphrases per target.

The table reports the aggregate stress-test statistics produced by the paraphrase evaluator. An individual failure on the `name` capability for "Please read the person's name from the document." reaches UQG $= 1.0$. The experiment therefore supports robustness only over the evaluated fixed query set, not unrestricted prompt search.

E.2 Adaptive person-presence probe

To test whether information remains decodable beyond the frozen deployed MLLM, a separate MLP probe is trained for person presence. The probe uses hidden dimension 256, 1,000 optimization steps, batch size 16, learning rate 3×10^{-4}, and seeds 20260821/20260822/20260823 with strict train/validation/test separation.

For ACB-denied versus No-Image, the mean score difference is 0.3333 with bootstrap 95% CI $[0.2951, 0.3770]$. The probe therefore recovers substantial residual signal from the protected repre-

Table 13: Molmo2-O-7B fixed-representation matrix under the Qwen-matched `max_per_target` $= 16$ protocol

Method	Subset	n	Retention	Excess capability	Frac. ret.> 0.25	Frac. ret.> 0.50
ACB w/o off-target preservation	diagonal	477	0.1455 ± 0.3589	0.0696	0.1992	0.1321
ACB w/o off-target preservation	off-diagonal	4,932	0.2822 ± 0.4339	0.1762	0.3745	0.2924
ACB	diagonal	477	0.2594 ± 0.6310	0.0921	0.2809	0.1761
ACB	off-diagonal	4,932	0.4980 ± 0.5979	0.2885	0.5440	0.4767

Table 14: Paired Molmo2-O-7B comparison between ACB w/o off-target preservation and ACB

Subset	Metric	n	W/o preservation	ACB	Δ	95% CI
Off-diagonal	Retention	4,932	0.2822	0.4980	+0.2158	[0.2003, 0.2312]
Off-diagonal	Excess capability	4,932	0.1762	0.2885	+0.1122	[0.1034, 0.1213]
Diagonal	Retention	477	0.1455	0.2594	+0.1138	[0.0644, 0.1671]
Diagonal	Target UQG	477	0.0696	0.0921	+0.0225	[0.0019, 0.0427]

sentation, showing that behavioral capability suppression should not be interpreted as representation-level information erasure.

F Efficiency and reproducibility details

All efficiency measurements use the same GPU, input slice, generation limit, and attention implementation across compared conditions. Profiling uses an NVIDIA RTX PRO 6000 Blackwell Server Edition, CUDA execution with SDPA attention, 16 inputs, two warm-up calls, three repeats, 50 bottleneck-only iterations, and `max_new_tokens`=32.

ACB's trainable parameter count is approximately 0.01993% of the frozen primary backbone. The reported 29,859,840 FLOPs/sample are the theoretical low-rank correction cost. The measured bottleneck path is sub-millisecond; the larger end-to-end difference includes the full generation path and framework overhead.

Reproducibility summary. The held-out primary evaluation uses seeds 20260821, 20260822, and 20260823. Rank and λ_{off} are selected only from primary-backbone validation data before the held-out test is evaluated. The full Molmo-13 cross-backbone evaluation uses seeds 20260924, 20260925, and 20260926 and directly transfers $r = 2$, $\lambda_{\text{off}} = 0.10$, and $K = 2$ without Molmo-specific retuning. Its fixed-representation matrix uses the same `max_per_target` $= 16$ and Raw–No-Image gap > 0.05 informative protocol as the primary evaluation. Reference answers are generated by the corresponding frozen model and reused during training.

Table 15: Fixed-representation paraphrase robustness

Metric	Mean ± std
Canonical target UQG	0.0211 ± 0.0107
Mean paraphrase UQG	0.0265 ± 0.0122
Aggregate worst-paraphrase UQG	0.1778 ± 0.0184
Robust denial rate	0.7685 ± 0.0262

Table 16: Adaptive person-presence probe

Representation condition	Probe test score
Raw	0.8798 ± 0.0095
ACB w/o off-target preservation, denied	0.5301 ± 0.0809
ACB, denied	0.8087 ± 0.0413
No-Image	0.4754 ± 0.0000

Table 17: Matched efficiency profile for the primary backbone

Method	E2E latency (s)	Bottleneck latency (s)	Peak GPU (GB)	Trainable params	Low-rank FLOPs/sample	Overhead
Raw	0.3590 ± 0.0029	0	15.617	0	0	0.00%
ACB w/o off-target preservation	0.4067 ± 0.0016	0.000523 ± 0.000042	15.626	1,652,227	29,859,840	13.29%
ACB	0.4010 ± 0.0007	0.000488 ± 0.000012	15.623	1,652,227	29,859,840	11.69%