M2AC: POLICY-AWARE ACCESS CONTROL FOR MULTIMODAL RETRIEVAL-AUGMENTED GENERATION SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

To address the fine-grained access control problem for cross-modal data in multimodal Retrieval-Augmented Generation (RAG) systems, this paper proposes a policy-aware access control framework, M2AC (Multimodal Access Control). Centered on multimodal derived data such as text, images, OCR text, captions, summaries, and embedding vectors, this framework extends the traditional binary allow/deny decision into a policy control mechanism that covers the entire process of retrieval, reranking, fusion, summarization, citation, and generation. M2AC comprises two core components: the Derived Permission Propagation Engine (DPPE) and the Modality-Aware Policy Enforcer (MAPE). Experiments on a LangChain-based multimodal RAG with three vector databases, namely Milvus, Weaviate, and pgvector, show that M2AC's policy generation takes 48.3 seconds, on the same order of magnitude as recent RAG-specific access control methods while using the fewest rules; the F1 score of end-to-end policy enforcement reaches 96.8%, an improvement of 3.4 to 7.2 percentage points over these methods. Its average retrieval latency increases by only 4.7% compared to the lightest baseline, and its average defense success rate against four types of cross-modal attacks, such as OCR bypass, reaches 96.2%.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has become a core paradigm for Large Language Model (LLM) applications and, with the advent of vision-language models (VLMs; Liu et al., 2023; Alayrac et al., 2022), has been extended to multimodal knowledge bases containing product images, engineering drawings, and scanned documents. Derived data generated from original objects (e.g., medical images containing sensitive information) through Optical Character Recognition (OCR), captioning, summarization, and embedding are distributed across the stages of the pipeline, forming complex permission propagation chains.

Existing permission management for RAG relies on Role-Based Access Control (RBAC) (Ferraiolo et al., 2001) and Attribute-Based Access Control (ABAC) (Hu et al., 2014). Recent efforts, including metadata-based retrieval filtering (Chen & Taipalus, 2025), permission-aware retrieval delegated to native Identity and Access Management (IAM) (Jeong & Lee, 2025), retrieval-time selective disclosure (Al Masoud et al., 2026), and privacy-isolated retrieval (Zhou et al., 2025), also largely remain at the document or text granularity and suffer from three limitations. First, binary allow/deny decisions cannot express rich access semantics, such as permitting access to an image description while prohibiting access to the raw image pixels. Second, permission inheritance for derived data is ill-defined. For example, it is unclear whether the OCR text of a restricted image inherits the restrictions of the original image, or whether the summary of a sensitive document can be shared more broadly. Third, end-to-end policy control is lacking, allowing attackers to bypass the permission constraints on original objects through derived channels such as OCR text, captions, or summaries. Studies on prompt injection (Greshake et al., 2023) and knowledge base poisoning (Zou et al., 2025) have revealed the vulnerabilities of RAG, and policy engines (Apache Software Foundation, 2026) and information flow control (Myers & Liskov, 1997) offer partial insights, but none of them systematically addresses the above problems.

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.

This paper proposes M2AC (Multimodal Access Control), a policy-aware access control framework for multimodal RAG. Its Derived Permission Propagation Engine (DPPE) models permission inheritance between original objects and derived data using a Directed Acyclic Graph (DAG), supporting permission attenuation, degraded access, and conditional propagation. The Modality-Aware Policy Enforcer (MAPE) integrates with Casbin and deploys policy enforcement points at six stages, namely retrieval, reranking, fusion, summarization, citation, and generation, for real-time adjudication. Experiments based on LangChain (LangChain, 2023) with Milvus (Wang et al., 2021), Weaviate (Weaviate, 2023), and pgvector (Kane, 2021) demonstrate the advantages of M2AC in policy quality, enforcement accuracy, and security defense. The main contributions are as follows: (1) We propose a DAG-based multimodal derived permission propagation model that supports permission attenuation and degraded access. (2) We design a modality-aware policy enforcement mechanism covering the entire process from retrieval to generation, enabling multi-stage and multi-granularity policy control. (3) We propose a systematic defense against four types of attacks, namely OCR bypass, caption leakage, cross-modal existence inference, and summary over-privilege.

2 Related Work

Access control has evolved from discretionary and mandatory access control (Bell & LaPadula, 1973) to RBAC (Ferraiolo et al., 2001) and ABAC (Hu et al., 2014). The static roles of RBAC can hardly express the dynamic permission requirements of multimodal pipelines. ABAC and its standardized policy language (OASIS, 2013) are complex to author and unaware of derivation semantics. Relationship-Based Access Control (ReBAC) (Pang et al., 2019) likewise cannot characterize data derivation chains. Casbin (Apache Software Foundation, 2026) supports the unified expression of multiple models but has not been extended for multimodal RAG. Metadata filtering (Chen & Taipalus, 2025), IAM-delegated retrieval (Jeong & Lee, 2025), and selective disclosure (Al Masoud et al., 2026) move permission checks forward into the retrieval pipeline, yet they remain centered on text documents.

CLIP (Radford et al., 2021), BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023), and Flamingo (Alayrac et al., 2022) have advanced unified image-text representation and vision-language generation, while MuRAG (Chen et al., 2022), RA-CM3 (Yasunaga et al., 2023), and UniIR (Wei et al., 2024) have extended retrieval augmentation to multiple modalities. These systems involve derivation operations such as OCR (Du et al., 2022), image captioning (Li et al., 2022), and document summarization (Zhang et al., 2020), yet they generally overlook permission propagation for derived data.

RAG security is receiving increasing attention. On the attack side, prompt injection (Greshake et al., 2023; Perez & Ribeiro, 2022) manipulates LLM output through malicious prompts, knowledge base poisoning (Zou et al., 2025; Zhong et al., 2023) misleads retrieval through fake documents, and data poisoning (Carlini et al., 2024) contaminates the embedding model during the training phase. In multimodal scenarios, the attack surface further expands: adversarial images (Goodfellow et al., 2015) can mislead visual encoders, and cross-modal injection (Bailey et al., 2024) exploits text embedded in images to manipulate VLMs. On the defense side, information flow control (Myers & Liskov, 1997) provides a theoretical foundation for tracking data propagation, and differential privacy (Dwork, 2006) provides formal leakage guarantees, but they face efficiency and usability challenges in practical multimodal RAG deployment.

Regarding vector database security, Milvus (Wang et al., 2021) supports collection-level RBAC, Weaviate (Weaviate, 2023) provides API-key-based authentication and simple RBAC, and pgvector (Kane, 2021), as a PostgreSQL extension, inherits its row-level security policies. These native mechanisms enforce access control only at the storage layer, ignoring the permission requirements of subsequent stages such as retrieval, reranking, fusion, and summarization, and they lack permission awareness of multimodal derived data; no work has yet embedded control policies into pipeline stages such as retrieval query execution, reranking, or fusion. In the data governance domain, Apache Atlas (Apache Software Foundation, 2020) and Amundsen (Lyft Engineering, 2020) provide data lineage tracking, but lack deep integration with access control.

In summary, existing technologies face three shortcomings in multimodal RAG: traditional RBAC/ABAC and even recent RAG-specific methods still lack the ability to model the permission propagation of multimodal derived data, existing RAG security research focuses on attack analysis

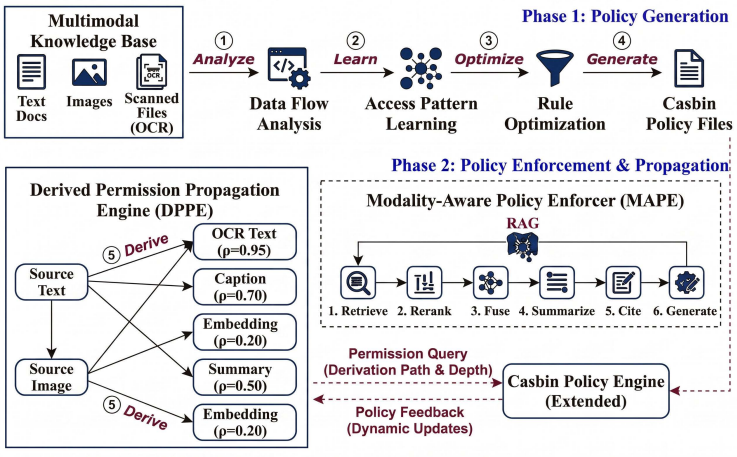


Figure 1: The M2AC framework for policy-aware access control in multimodal RAG systems.

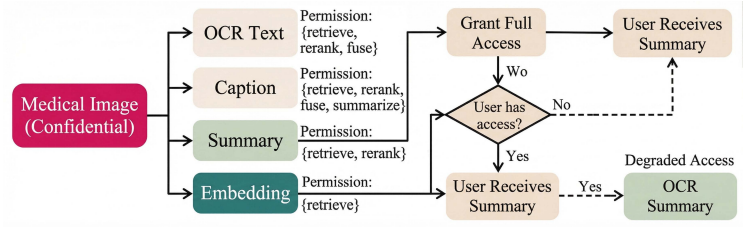


Figure 2: An example of derived permission propagation in DPPE.

but lacks a systematic end-to-end defense framework, and native vector database mechanisms cannot cover multiple stages of the RAG pipeline. Based on a derived permission propagation model, M2AC achieves policy-aware access control for the whole process of multimodal RAG.

3 METHODOLOGY

As shown in Figure 1, DPPE models permission inheritance and attenuation between original objects and derived data using a DAG, and MAPE deploys policy enforcement points at six pipeline stages. The two interact bidirectionally, with MAPE querying DPPE for effective permissions and degradation paths at each stage, and DPPE updating its permission cache according to the adjudication results. The automatic policy generation module analyzes data flow graphs and access patterns to construct a rule set covering all modalities and derivation paths.

3.1 DERIVED PERMISSION PROPAGATION

DPPE models derived permission propagation with a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node $v \in \mathcal{V}$ carries a modality label $\mu(v) \in \{text, image, ocr, caption, summary, embedding\}$ and a base permission set $\Pi_0(v)$. A directed edge $(u, v) \in \mathcal{E}$ indicates that v is derived from u, and it carries an operation type $\delta(u, v) \in \{ocr, caption, summarize, embed, crop, translate\}$ and a permission attenuation function $\lambda_\delta : 2^\Pi \to 2^\Pi$. Figure 2 takes a medical image as an example, where the original image ("Confidential") produces derived data through four operations. Each edge is annotated with its information retention ratio ρ and attenuation rule, and the color intensity of each node reflects the size of its effective permission set.

The effective permissions of a derived node v are computed by the following recursive formula:

$$\Pi_{eff}(v) = \Pi_0(v) \cap \bigcap_{(u,v)\in\mathcal{E}} \lambda_{\delta(u,v)}(\Pi_{eff}(u)) \tag{1}$$

The recursion proceeds top-down in topological order. When a node has multiple predecessors, the intersection of the permissions propagated from all predecessors is taken, following the principle of least privilege. The permission space $\Pi = \{retrieve, rerank, fuse, summarize, cite, generate\}$ corresponds to the six stages of RAG. The function λ_δ prunes permissions according to the information retention ratio $\rho(\delta) \in [0, 1]$. OCR preserves the text of the original image almost completely and thus has the highest retention ratio, embedding can hardly recover the original content and thus has the lowest, and caption and summarize lie in between. When $\rho(\delta)$ falls below the degradation threshold τ, cite and generate are no longer propagated, and when it decreases further, summarize and fuse are also removed, retaining only retrieve and rerank.

Degraded access is a key mechanism of M2AC. For example, a medical image containing sensitive patient information should not be returned to ordinary users, but its de-identified text summary can be. When a user lacks the requested permission π on v, the degradation function $\Gamma : \mathcal{V} \times \Pi \to \mathcal{V} \cup \{\perp\}$ follows the principle of information minimization to search the derivation subgraph of v for substitute data that satisfies the conditions and has the lowest information retention ratio. Candidates are restricted to human-understandable modalities such as summaries and captions, with embeddings excluded by default. When multiple derivation paths exist, Γ ranks candidates uniformly by the product of the ρ values accumulated along the derivation edges.

3.2 Policy Enforcement and Cross-Modal Defense

MAPE registers custom matching functions through the ABAC model of Casbin, treating modality, pipeline stage, and derivation depth as extended attributes of the request context, thereby supporting multimodal semantics without modifying the Casbin core engine. The policy model is defined as a seven-tuple $\mathcal{P} = (\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{M}, \mathcal{ST}, \mathcal{D}, \mathcal{C})$, where $\mathcal{S}$ denotes subjects, $\mathcal{O}$ denotes resource objects, $\mathcal{A} = \{allow, deny, degrade\}$ denotes actions, $\mathcal{M}$ and $\mathcal{ST}$ denote the aforementioned six modalities and six pipeline stages respectively, $\mathcal{D} \in \mathbb{N}$ denotes the maximum allowed derivation depth, and $\mathcal{C}$ denotes condition predicates. Each rule is expressed as:

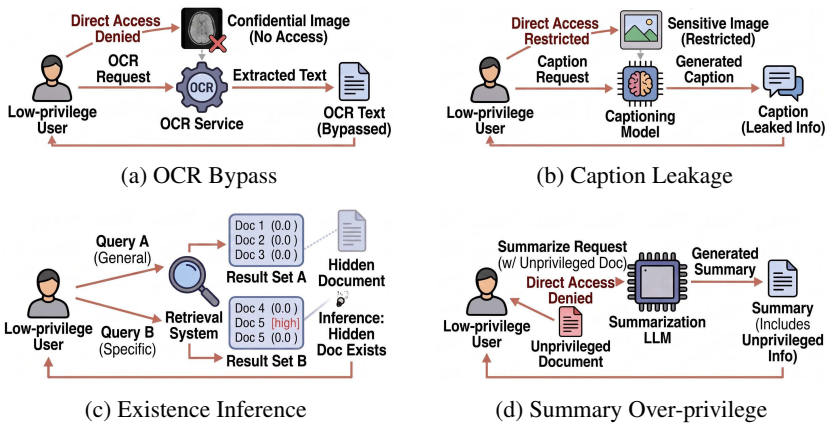
$$p = (s, o, a, m, st, d, c) : \text{subject } s \text{ may } a \text{ object } o \\ \text{of modality } m \text{ at stage } st \text{ within depth } d \text{ if } c \tag{2}$$

For a request $req = (s_{req}, o_{req}, m_{req}, st_{req})$, MAPE first queries DPPE for the derivation depth d_{req} of o_{req}, then retrieves the matching rules, and uses condition predicates to evaluate runtime context such as time windows and request frequency. The adjudication function Θ is defined as:

$$\Theta(req) = \begin{cases} allow, & \text{if } \exists p \in \mathcal{P}_{match},\ p.a = allow \quad \wedge \nexists p' \in \mathcal{P}_{match},\ p'.a = deny \\ degrade, & \text{if } \exists p \in \mathcal{P}_{match},\ p.a = degrade \quad \wedge \Gamma(o_{req}, st_{req}) \neq \perp \\ deny, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{P}_{match}$ is the subset of rules that match on the four dimensions of subject, resource, modality, and stage and satisfy $d_{req} \leq p.d$ and $p.c$. M2AC adopts deny-override semantics, under which any matching deny overrides all allows, and degrade takes effect only when there is neither an allow nor a deny, ensuring that adjudication under conflicts is deterministic and biased toward safety. Enforcement differs across stages. The retrieval stage evaluates permissions routed by modality, the reranking stage sets the scores of restricted documents to a neutral value to eliminate existence leakage, the fusion stage aligns permissions across modalities, the summarization stage sanitizes unauthorized inputs and constrains the LLM through prompt engineering, the citation stage filters or replaces restricted sources, and the generation stage performs post-processing audits and removes restricted fragments.

The threat model assumes that the attacker is an insider holding legitimate low-privilege credentials who can issue normal RAG requests but cannot tamper with the DPPE permission graph or the Casbin policies. Figure 3 illustrates four cross-modal attacks and their defenses. The OCR bypass attack performs OCR on restricted images to obtain their text. Since OCR approaches a complete copy in terms of information retention, DPPE makes the derived text inherit the constraints of the original image almost entirely. The caption leakage attack has a lower information retention ratio



(a) OCR Bypass (b) Caption Leakage

(c) Existence Inference (d) Summary Over-privilege

Figure 3: Overview of four cross-modal bypass attacks and M2AC's corresponding defense mechanisms.

but may still leak sensitive information, so M2AC makes captions inherit the corresponding restrictions when the sensitivity level of the original image exceeds a threshold. Cross-modal existence inference exploits differences in the number or confidence of retrieval results to reveal the existence of restricted documents, which M2AC defends against with constant-size result sets padded with placeholders and calibrated score distributions. The summary over-privilege attack requests a summary of a result set containing restricted documents. M2AC removes unauthorized fragments before passing them to the LLM and prohibits the LLM from speculating about the removed content in the system prompt.

3.3 Automatic Policy Generation

M2AC automatically generates policy rules in three phases. In the derivation relationship discovery phase, data flow analysis statically traces all derived data of each original object in the pipeline $\mathcal{L} = (l_1, l_2, \ldots, l_n)$ to construct $\mathcal{G}$, and runtime monitoring supplements dynamic derivation paths such as real-time OCR triggered by user requests. In the access pattern learning phase, access logs $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ containing subjects, resource objects, modalities, pipeline stages, and timestamps are collected during integration testing and canary deployment. Using subject role, resource modality, pipeline stage, and derivation depth as features, hierarchical clustering with Euclidean distance and the Ward linkage criterion is performed (with the number of clusters determined by the silhouette coefficient), and candidate rules are generated for each group of similar requests. In the policy rule optimization phase, rule set minimization is formulated as a constrained set cover problem:

$$\begin{aligned} \min_{\mathcal{P}'} \quad & |\mathcal{P}'| \\ \text{s.t.} \quad & \forall r \in \mathcal{R}_{\text{legit}},\ \exists p \in \mathcal{P}' :\ \Theta(r,p) = \mathit{allow}, \\ & \forall r \in \mathcal{R}_{\text{forbid}},\ \nexists p \in \mathcal{P}' :\ \Theta(r,p) = \mathit{allow}, \\ & \forall (u,v) \in \mathcal{E},\ \forall p \in \mathcal{P}',\ p.o = v \wedge p.a = \mathit{allow} \Rightarrow p.st \in \lambda_{\delta(u,v)}\big(\Pi_{\text{eff}}(u)\big) \cup \Pi_0(v). \end{aligned} \tag{4}$$

The three constraints respectively ensure that legitimate requests are allowed, that illegitimate requests are denied, and that the authorized stage $p.st$ of any allow rule acting on a derived node v does not exceed the propagation upper bound, thereby explicitly coupling policy optimization with the derivation consistency of DPPE. This NP-hard problem is solved with a greedy algorithm that, at each step, selects the rule covering the most uncovered legitimate requests without introducing over-privilege, achieving an approximation ratio of $(1 + \ln|\mathcal{R}_{legit}|)$ and a complexity of $O(|\mathcal{P}_{cand}| \cdot |\mathcal{R}_{legit}|)$. It can be completed within minutes at the scale of thousands of original objects and dozens of derivation relationships. The generated rules are encoded as Casbin policy files

Table 1: Statistics of the multimodal experimental datasets

Metric	Mean	Min	Max
#Access requests per DB	71.5K	9.2K	182K
#Unique access patterns per DB	3,260	412	9,840
#Collections per DB	18.4	5	52
#Modality types per collection	3.6	2	6
#Derived objects per original object	4.2	1	8
Derivation depth	2.1	1	4

for MAPE to load. To cope with changes in access patterns and new derivation types, MAPE incrementally feeds new observations back to DPPE through the policy feedback channel, and DPPE recomputes permission propagation only for the affected subgraphs and locally updates the rules, supporting online policy adaptation.

4 EVALUATION

The experimental environment consists of an Intel Xeon Gold 6338 and Ubuntu 22.04. The software includes Milvus 2.3.4, Weaviate 1.24, PostgreSQL 16, pgvector 0.6.0, LangChain 0.1.0, CLIP ViT-L/14, and Casbin v1.36.1, and the pipeline integrates Tesseract (OCR), BLIP-2 (captioning), and PEGASUS (summarization). M2AC is integrated in middleware mode via a gRPC interceptor (Milvus), HTTP middleware (Weaviate), and a row-level security policy extension (pgvector), respectively.

4.1 DATASETS

Table 1 presents the dataset statistics. We construct three multimodal knowledge bases for enterprise documents (product manuals, engineering drawings, product images, etc.), healthcare (medical records, medical images, examination reports, etc.), and legal affairs (contracts, seal images, legal documents, etc.). The data are drawn from real business operations and de-identified, and each knowledge base is deployed on all three databases. Access logs were collected from approximately six weeks of integration testing and canary deployment covering 3 institutions from different industries and 214 real users, and were split chronologically into training and test sets at a ratio of 7:3. Policies are generated on the training set and metrics are computed on the test set. Eighteen engineers with an average of 3.5 years of experience performed the annotation based on the least privilege criterion while considering the reasonableness of permission propagation across modalities. Each data item was independently annotated by at least two annotators, yielding a Cohen's kappa of 0.85. Evaluation metrics include policy generation time, number of policy rules, six-stage policy adjudication F1 score, request pass rate (the proportion of legitimate requests that are allowed), retrieval latency, and throughput.

4.2 PERFORMANCE

The baselines include Metadata-Aware RAG (MA-RAG) (Chen & Taipalus, 2025), which performs metadata filtering; Permission-Aware RAG (PA-RAG) (Jeong & Lee, 2025), which delegates permission-aware retrieval to native IAM; Selective Disclosure RAG (SD-RAG) (Al Masoud et al., 2026), which performs retrieval-time selective disclosure; OpenFGA-based ReBAC (Pang et al., 2019); and ModalAC (Modal Access Control), a naive baseline that configures permissions independently for each modality without considering derivation relationships. For a fair comparison, all methods are connected to the same automated policy generation pipeline and uniformly use the finest-grained native permissions of each database (e.g., collection level in Milvus and table level in PostgreSQL), differing only at the access control model level. As shown in Table 2, M2AC takes 48.3 seconds to generate policies, on the same order of magnitude as PA-RAG (47.1 seconds) and MA-RAG (51.2 seconds), indicating that its advantages do not stem from differences in the degree of automation. M2AC produces the fewest rules (68), 10.5% to 39.3% fewer than the other methods (Cohen's $d \geq 0.72$), and achieves an F1 score of 96.8%, which is 6.6, 5.0, 4.4, 3.4, and 7.2 percentage points higher than MA-RAG, PA-RAG, SD-RAG, ReBAC, and ModalAC, respectively

Table 2: Time overhead and policy quality comparison of different methods

Method	Time (s)↓	#Rules↓	F1 (%)↑	Method	Time (s)↓	#Rules↓	F1 (%)↑
MA-RAG	51.2	84	90.2	ReBAC	65.7	91	93.4
PA-RAG	**47.1**	76	91.8	ModalAC	72.4	112	89.6
SD-RAG	55.6	89	92.4	M2AC	48.3	**68**	**96.8**

Table 3: Pass rate, latency, and throughput comparison of different methods

Methods	Pass (%) ↑	Lat. (ms) ↓	Thpt. (req/s) ↑	Methods	Pass (%) ↑	Lat. (ms) ↓	Thpt. (req/s) ↑
MA-RAG	91.4	**10.32**	**8,624**	ReBAC	94.6	11.58	8,184
PA-RAG	93.2	11.06	8,352	ModalAC	91.6	11.82	7,986
SD-RAG	93.8	11.72	8,098	M2AC	**96.4**	10.81	8,510

($p < 0.05$, Wilcoxon signed-rank test), accurately capturing the actual usage patterns of multimodal RAG.

4.3 Usability and Efficiency

Table 3 compares the request pass rate, retrieval latency, and throughput. M2AC achieves a pass rate of 96.4%, the highest among all methods. MA-RAG and ModalAC suffer from more false rejections because they cannot perform degraded access (with pass rates of 91.4% and 91.6%, respectively), whereas M2AC can still return substitute data with less information when permissions on the original object are insufficient. MA-RAG performs only lightweight metadata filtering and thus has the lowest latency (10.32 ms) and the highest throughput (8,624 req/s). M2AC additionally introduces DAG permission propagation, degradation resolution, and six-stage enforcement, resulting in an average latency of 10.81 ms, 4.7% higher than MA-RAG, and a throughput of 8,510 req/s. Because DPPE precomputes and caches the propagation results of the derived permission graph during initialization, adjudication at runtime requires only constant-time hash lookups. M2AC thus avoids the overhead of relationship graph traversal in ReBAC and retrieval-time sanitization in SD-RAG, and still outperforms PA-RAG, SD-RAG, ReBAC, and ModalAC in latency and throughput.

Table 4 evaluates scalability under different request scales. The time and memory overhead of M2AC grows approximately linearly with the request scale, requiring only about 13.4 minutes for 1 million requests. The memory reported in the table represents the additional overhead of the DPPE permission graph, the MAPE policy engine, and the policy cache, excluding the vector database and the RAG pipeline, of which maintaining the derived permission graph accounts for approximately 0.3 to 0.8 GB.

4.4 Security Analysis

Against the multimodal attacks in the threat model, we evaluated M2AC's defense capability. Table 5 summarizes the defense success rates against different attacks. Each attack was designed with 100 simulations, all launched by a simulated user holding minimal permissions (only public text retrieve permission). The construction of the four types of attacks refers to existing multimodal and RAG attack literature (Bailey et al., 2024; Zou et al., 2025) and adopts red-teaming methods (Perez et al., 2022), implemented by personnel independent of the defense design. Specifically, the OCR bypass attack uses Tesseract to perform OCR on restricted images and attempts to obtain content through the text retrieval channel, the caption leakage attack requests the BLIP-2 description of restricted images via API to indirectly obtain semantics, the cross-modal existence inference attack infers the existence of restricted documents by observing the count and confidence distribution of retrieval results, and the summary over-privilege attack indirectly obtains content by requesting summary generation over a result set containing restricted documents.

Regarding the OCR bypass attack, M2AC propagates the original image's permission constraints to the OCR-derived text through DPPE's high-information-retention propagation ($\rho(ocr) = 0.95$), achieving a defense success rate of 99.1%, whereas MA-RAG and PA-RAG achieve 45.6% and

Table 4: Scalability test of M2AC on large-scale requests

Request Scale	Time (min)	Memory (GB)	Request Scale	Time (min)	Memory (GB)
100K	2.1	0.9	1M	13.4	3.2
500K	7.2	1.9	2M	25.8	5.4

Table 5: Defense success rate (%) against multimodal RAG-specific attacks

Attack Type	MA-RAG	PA-RAG	ModalAC	M2AC
OCR Bypass (OCR-B)	45.6	58.2	42.8	**99.1**
Caption Leakage (Cap-L)	41.2	52.4	48.2	**95.8**
Cross-Modal Existence Inference (CME)	30.4	38.6	34.6	**93.6**
Summary Over-Privilege (Sum-O)	38.2	46.8	40.4	**96.3**
Average	38.9	49.0	41.5	**96.2**

58.2% respectively, due to their lack of complete derived data permission propagation capability. Regarding the caption leakage attack, M2AC imposes restrictions on sensitive image captions through conditional permission propagation, achieving a success rate of 95.8%. Regarding the cross-modal existence inference attack, M2AC eliminates the information leakage channel through constant-size result sets and calibrated score distributions, achieving a success rate of 93.6%. Regarding the summary over-privilege attack, M2AC enforces input sanitization and output auditing at the summarization stage, achieving a success rate of 96.3%.

Table 6 further presents the defense success rates by scenario and target modality. M2AC performs best on the medical knowledge base (97.3% on average), because medical data carry explicit privacy level annotations, enabling DPPE to propagate permissions more accurately. By target modality, OCR achieves the highest rate (96.2% on average), while summary is relatively lower (94.4% on average), as the latter is affected by the uncertainty of LLM outputs.

4.5 Ablation Study

As shown in Table 7, removing any component degrades access control quality. Without DPPE, the system degenerates into independent per-modality control, with F1 dropping to 88.4% and the defense rates against OCR bypass and caption leakage plummeting to 24.2% and 30.6%, respectively, indicating that DPPE is key to cross-modal defense. Without degraded access, only binary decisions are possible, and the pass rate drops to 91.2%. When policies are enforced only at the retrieval stage (w/o Pipeline), F1 drops to 89.6%, and the defense rates against summary over-privilege and cross-modal existence inference decrease to 31.2% and 42.8%, respectively. Replacing Casbin with a simple rule matching engine reduces F1 to 92.4% due to the lack of efficient policy indexing and flexible model adaptation. Using only access pattern learning without data flow graph analysis (Log Only) reduces F1 to 84.6%, showing that static analysis is essential for discovering derivation paths not covered by the logs.

Parameter Sensitivity Analysis. As shown in Table 8, we apply random perturbations to the estimated ρ of each operation and recompute permission propagation and adjudication. Under a perturbation of ± 0.1, F1 fluctuates within ± 0.9 percentage points, and the defense success rates against OCR bypass and caption leakage change by less than 1.5 percentage points. This is because λ_δ prunes permissions in a piecewise manner depending on whether ρ crosses two calibrated thresholds rather than through a continuous linear mapping, making it insensitive to small perturbations.

5 Discussion

M2AC still has several limitations. First, the information retention ratio ρ relies on a calibration dataset, and its optimal values may vary across scenarios. Future work will explore end-to-end automatic calibration based on information theory. Second, centralized policy adjudication may suffer from synchronization delays and single-point bottlenecks in geographically distributed or multi-tenant scenarios, which could be addressed by an enforcement architecture based on distributed

Table 6: Detailed defense success rate (%) of M2AC across scenarios and target modalities

Attack Type	By Scenario			By Target Modality					
	Enterprise	Medical	Legal	Text	Image	OCR	Caption	Summary	Embedding
OCR-B	98.7	99.8	98.8	–	–	99.1	–	–	–
Cap-L	95.3	97.0	95.1	–	–	–	95.8	–	–
CME	93.1	94.8	92.9	93.8	94.0	93.2	93.4	92.4	94.8
Sum-O	95.9	97.4	95.6	–	–	–	–	96.3	–
Average	95.8	97.3	95.6	93.8	94.0	96.2	94.6	94.4	94.8

"–" indicates that the attack type is not applicable to that modality.

Table 7: Ablation study of M2AC. Def. denotes the average defense success rate over the four attacks

Configuration	F1 (%)↑	Pass (%)↑	Lat. (ms)↓	Def. (%)↑
M2AC (full)	**96.8**	**96.4**	10.81	**96.2**
w/o DPPE	88.4	90.8	10.14	38.4
w/o Degrade	96.2	91.2	10.52	95.4
w/o Pipeline	89.6	93.2	**9.88**	52.4
w/o Casbin	92.4	94.6	11.32	92.8
Log Only	84.6	88.4	10.68	82.6

Table 8: Sensitivity of M2AC to perturbations of the information retention ratio ρ. $\Delta\rho$ denotes the magnitude of the random perturbation applied to the ρ of each operation; OCR-B and Cap-L report defense success rates

$\Delta\rho$	F1 (%)	OCR-B (%)	Cap-L (%)	$\Delta\rho$	F1 (%)	OCR-B (%)	Cap-L (%)
0	96.8	99.1	95.8	$\pm$0.10	96.1	98.2	94.8
$\pm$0.05	96.6	98.9	95.5	$\pm$0.20	95.3	96.9	93.6

consensus protocols. Third, only the text and image modalities have been validated so far, and extending to streaming modalities such as audio and video requires redesigning the temporal semantics of the derived permission graph and the real-time enforcement mechanism. Fourth, degraded access degenerates into complete denial when all derived data exceed the security threshold, and differential privacy could be introduced to provide minimal information services. In addition, M2AC does not yet protect the policy metadata itself, and attackers may reverse-engineer permission boundaries through probing queries at different granularities, which needs to be mitigated by combining query auditing with randomized degradation decisions. Finally, the number of rules may grow significantly with pipeline complexity, calling for research on hierarchical compression and incremental update mechanisms.

6 Conclusion

This paper proposes M2AC, a policy-aware access control framework for multimodal RAG. DPPE models permission inheritance and attenuation between original objects and derived data and supports degraded access, while MAPE integrates with Casbin to extend binary allow/deny decisions into policy control covering the entire process from retrieval to generation. Experiments on Milvus, Weaviate, and pgvector show that, compared with recent methods such as MA-RAG, PA-RAG, and SD-RAG, M2AC produces the fewest rules at comparable generation time, improves F1 by 3.4 to 7.2 percentage points, increases average retrieval latency by only 4.7% over the most lightweight baseline, and achieves an average defense success rate of 96.2% against four types of cross-modal attacks. Future work will explore automatic calibration of the information retention parameter and extend M2AC to streaming modalities such as audio and video.

AI USE STATEMENT

Apart from the models evaluated as components of the M2AC experimental pipeline, no generative AI tools were used at any stage of this work, including research ideation, method design, experimentation, and paper writing.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our research aims to strengthen data security and privacy protection in multimodal RAG systems. The use of real user data was approved by an Institutional Review Board (IRB), and all enterprise, medical, and legal data, as well as user access logs, were de-identified prior to use and release. All cross-modal attacks in this work were conducted solely by authorized personnel in an isolated experimental environment and did not target any real production system. Moreover, the attack types considered are drawn from existing public research and are described only to support defense design, and we believe the security benefits of this work substantially outweigh its potential for misuse.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we release the source code of M2AC and the de-identified experimental data in anonymized form at `https://anonymous.4open.science/r/iclr27-13401`, enabling reviewers and readers to reproduce our experimental results.

REFERENCES

Aiman Al Masoud, Marco Arazzi, and Antonino Nocera. SD-RAG: A prompt-injection-resilient framework for selective disclosure in retrieval-augmented generation. *arXiv preprint arXiv:2601.11199*, 2026.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Apache Software Foundation. Apache Atlas: Data governance and metadata framework for Hadoop. `https://atlas.apache.org/`, 2020. Accessed: 2026-07-05.

Apache Software Foundation. Apache Casbin: An authorization library that supports access control models. `https://casbin.apache.org/`, 2026. Accessed: 2026-07-05.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: adversarial images can control generative models at runtime. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 2443–2455, 2024.

David Elliott Bell and Leonard J LaPadula. Secure computer systems: Mathematical foundations. Technical Report MTR-2547, Vol. I, The MITRE Corporation, 1973.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024.

Bingxiang Chen and Toni Taipalus. Metadata-aware RAG for enforcing access control and metadata-based filtering: Proof of concept and evaluation. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems (RACS)*, pp. 1–7, 2025. doi: 10.1145/3769002.3769952.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. SVTR: Scene text recognition with a single visual model. In *International Joint Conference on Artificial Intelligence*, pp. 884–890, 2022.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer, 2006.

David F Ferraiolo, Ravi Sandhu, Serban Gavrila, D Richard Kuhn, and Ramaswamy Chandramouli. Proposed NIST standard for role-based access control. *ACM Transactions on Information and System Security (TISSEC)*, 4(3):224–274, 2001.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023.

Vincent C Hu, David Ferraiolo, Rick Kuhn, Adam Schnitzer, Kenneth Sandlin, Robert Miller, and Karen Scarfone. Guide to attribute based access control (ABAC) definition and considerations. Technical Report NIST Special Publication 800-162, National Institute of Standards and Technology, 2014.

Jooyoung Jeong and Sang-Goo Lee. Permission-Aware RAG: Identity and access management (IAM)-based access filtering in multi-resource environments. *IEEE Access*, 13:192819–192835, 2025. doi: 10.1109/ACCESS.2025.3628960.

Andrew Kane. pgvector: Open-source vector similarity search for PostgreSQL. `https://github.com/pgvector/pgvector`, 2021. Accessed: 2026-07-05.

LangChain. LangChain: Building applications with LLMs through composability. `https://github.com/langchain-ai/langchain`, 2023. Accessed: 2026-07-05.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Lyft Engineering. Amundsen: Open source data discovery and metadata engine. `https://www.amundsen.io/`, 2020. Accessed: 2026-07-05.

Andrew C Myers and Barbara Liskov. A decentralized model for information flow control. *ACM SIGOPS Operating Systems Review*, 31(5):129–142, 1997.

OASIS. eXtensible Access Control Markup Language (XACML) version 3.0. *OASIS Standard*, 2013.

Ruoming Pang, Ramon Caceres, Mike Burrows, Zhifeng Chen, Pratik Dave, Nathan Germer, Alexander Golynski, Kevin Graney, Nina Kang, Lea Kissner, et al. Zanzibar: Google's consistent, global authorization system. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, pp. 33–46, 2019.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, 2022.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, pp. 2614–2627, 2021.

Weaviate. Weaviate: Open-source vector search engine. `https://weaviate.io/`, 2023. Accessed: 2026-07-05.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*, pp. 39755–39769. PMLR, 2023.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pp. 11328–11339. PMLR, 2020.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775, 2023.

Pengcheng Zhou, Yinglun Feng, and Zhongliang Yang. Privacy-aware RAG: Secure and isolated knowledge retrieval. *arXiv preprint arXiv:2503.15548*, 2025.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025.