CVE-SAI: Counterfactual Visual Evidence-Guided Selective Attribute Indexing for Risk-Controlled E-commerce Search

Xiaolong Sun[1] Qichao Wang[2] Hangyu Li[3] Liang Chen[1,*]
[1]Sun Yat-Sen University, Guangzhou, China
[2]Nanyang Technological University, Singapore
[3]Tencent, Shenzhen, China
sunxlong@mail2.sysu.edu.cn qichao001@e.ntu.edu.sg
masonhyli@tencent.com chenliang6@mail.sysu.edu.cn
*Corresponding author

Abstract

Multimodal product models can complete missing e-commerce attributes, yet current methods still optimize attribute-answer accuracy without verifying visual support, conflate transient prediction with persistent index admission, and lack explicit risk control over factually incorrect or visually unsupported values. We address these gaps with Counterfactual Visual Evidence-Guided Selective Attribute Indexing (CVE-SAI), which first infers and freezes an ontology-constrained candidate from the primary image and attribute question without catalog text, and then decides whether that candidate should enter the index. Focus-Zone Distortion (FZD) constructs an attribute-specific visual-dependence proxy through a controlled counterfactual intervention, and Evidence-Guided Attention Redistribution (EGAR) uses the proxy to refine ontology-constrained scoring. The canonical candidate is frozen before evidence necessity, evidence retention, nuisance-transformation stability, and candidate-specific catalog-text conflict audits; catalog text can only tighten admission and cannot revise the candidate. Independent family-level calibration selects one policy with a simultaneous one-sided finite-sample bound under a 5% unsafe-admission budget. Experiments on five visual attributes derived from Amazon Berkeley Objects show that CVE-SAI improves attribute inference and evidence localization, achieves the highest certified admission coverage under the shared risk protocol, and yields the strongest controlled retrieval performance with the lowest unsafe auto-induced exposure among automatic-admission systems. Separating inference from admission therefore enables visually supported attribute completion to improve retrieval while limiting persistent index contamination.

1 Introduction

Product attributes support product retrieval, ranking, recommendation, and catalog understanding, yet large catalogs often leave attribute values missing or incomplete. MAVE [1] documents this problem across diverse product categories, and MOON [2] shows that visual and textual product content can support attribute prediction and cross-modal retrieval. Product imagery therefore offers a practical source for completing catalog records. The same prediction becomes consequential when it is stored and repeatedly reused by search infrastructure rather than consumed once.

Multimodal large language models can infer product attributes from images and jointly represent heterogeneous product content. Their fluency does not ensure that every output is visually supported. Visual Contrastive Decoding (VCD) [3] mitigates unsupported generation by contrasting outputs from clean and distorted images. For product attributes, language-model confidence, ontology validity, and support from the primary image are distinct properties. A candidate may satisfy the first two and still be unsuitable for a persistent catalog update.

Existing work addresses either product attribute inference and representation learning or visually unsupported multimodal generation. Neither direction jointly verifies support from the current product image and controls whether a prediction enters a persistent index. A visually plausible answer may therefore remain unsuitable as a durable search signal.

A one-time prediction error becomes persistent retrieval contamination once the inferred attribute is indexed and reused for matching and ranking. A suitable admission rule should respond to changes in attribute-relevant evidence while remaining stable under semantics-preserving nuisance variation, mirroring the counterfactual behavior studied in robust retrieval [6]. This distinction calls for an index-admission decision that is separate from attribute inference.

We formulate risk-controlled selective product attribute indexing as a decision separate from attribute inference. The model may answer or abstain, but only a frozen canonical value can be considered for admission. Image

truth records what the primary image supports, whereas item truth records the trusted product value. SelectiveNet [7] motivates the reject option and risk–coverage trade-off, while Learn then Test [8] supports policy selection under a prespecified risk budget. We set the unsafe-admission budget to 5% on an independent family-level calibration population.

As illustrated in Figure 1, we propose Counterfactual Visual Evidence-Guided Selective Attribute Indexing (CVE-SAI), which separates product-text-free attribute inference from risk-controlled index admission. Focus-Zone Distortion (FZD) constructs an attribute-specific visual-dependence proxy, and Evidence-Guided Attention Redistribution (EGAR) uses it to refine scores over the ontology. CVE-SAI freezes the resulting canonical candidate before necessity, retention, nuisance-transformation stability, and candidate-specific catalog-text conflict audits. Independent calibration then determines whether the candidate enters the isolated auto-attribute field.

We construct a five-attribute benchmark and controlled retrieval evaluation from Amazon Berkeley Objects (ABO) [10]. Across the five visually auditable attributes, CVE-SAI achieves the highest Macro Visual Answer Accuracy (Macro VAA) and Answerability Macro-F1 (Ans.-F1) among the evaluated visual-answering systems, while FZD obtains the highest Macro Patch AUPRC among evidence-localization methods. Under the prespecified 5% unsafe-admission budget, CVE-SAI attains the highest Certified Write Coverage (CWC@5%) among the evaluated risk-controlled systems. Its admitted attributes also yield the highest NDCG@10 and the lowest Unsafe Auto-Induced Exposure@10 (UAIE@10) among automatic-admission systems in controlled Lucene retrieval. Our contributions are as follows:

- We formulate risk-controlled selective product attribute indexing, which separates attribute inference from persistent index admission. The formulation distinguishes image truth from item truth, supports explicit abstention, and defines unsafe-admission risk for admitted canonical values.
- We introduce the visual front end of CVE-SAI. FZD derives an attribute-specific visual-dependence proxy from a controlled counterfactual edit, and EGAR uses this proxy to improve ontology-constrained candidate scoring. The frozen candidate is then examined through evidence necessity, evidence retention, nuisance-transformation stability, and candidate-specific catalog-text conflict audits.
- We develop a risk-controlled admission strategy that uses independent finite-sample calibration to select the feasible policy with the highest coverage under a prespecified budget for unsafe admission. The selected policy is frozen before test, and only admitted canonical values enter the isolated auto-attribute field.
- We construct a five-attribute benchmark from ABO and evaluate CVE-SAI across attribute inference, evidence localization, index admission, and retrieval with a fixed Lucene pipeline. CVE-SAI achieves the highest certified admission coverage under the 5% risk budget, improves retrieval effectiveness, and reduces unsafe auto-induced exposure among the evaluated automatic-admission systems.

2 Related Work

2.1 Product Attribute Extraction

Product attributes support product ranking, retrieval, and recommendation, motivating benchmarks such as MAVE, which combines titles, descriptions, features, and specifications across diverse categories and exposes incomplete values and a challenging zero-shot split. MBSD [11] transfers knowledge across product modalities through self-distillation, whereas MOON learns generative multimodal representations for attribute prediction and cross-modal retrieval. These representation-learning approaches predict or encode product content but do not decide whether inferred attributes should enter a persistent index. MXT [12] formulates large-scale multimodal extraction as question answering over images and text, while DEFLATE [13] generates explicit and implicit values before assessing candidate credibility with a discriminator. EIVEN [14] efficiently adapts multimodal language models for implicit values, and ImplicitAVE [15] provides a curated benchmark for this setting. Under image-only inference, ViOC-AG [17] transfers textual attribute knowledge to a visual generator and corrects out-of-domain values, whereas MICE [18] combines specialized captioning experts to provide complementary visual descriptions. HyperPAVE [16] models higher-order relations induced by user behavior and product inventory in a heterogeneous hypergraph and uses inductive link prediction for unseen values, while Hypergraph PAVE [19] integrates visual and textual product information into multimodal hypergraphs for zero-shot attribute prediction. MSIT [20] extends attribute mining to the open world through multimodal self-correction, and TACLR [21] uses taxonomy-aware retrieval to handle implicit and out-of-distribution values while producing normalized outputs. Beyond product attributes, HADSF [22] uses structured aspect–opinion signals to improve downstream recommendation, illustrating the value of evaluating extracted information through its system-level effect. These methods improve attribute generation, extraction, correction, or normalization, but generally treat prediction as the endpoint. CVE-SAI instead separates product-text-free candidate generation from risk-controlled persistent index admission.

2.2 Reliable Multimodal Prediction

POPE [23] evaluates object hallucination through binary questions about whether mentioned objects are present in an image rather than scoring unconstrained free-form descriptions. Visual Contrastive Decoding contrasts outputs from clean and distorted images, while M3ID [24] strengthens dependence on the visual prompt through mutual-information decoding. These decoding methods reduce unsupported content without retraining the base model. Their interventions affect current responses rather than persistent catalog state. AGLA [4] assembles global and local attention, whereas CMAC [5] calibrates cross-modal attention to balance visual and textual influence during training-free decoding. Same Attention, Different Truths [25] shows that similar visual-attention magnitudes can correspond to different object-hallucination outcomes, motivating logit-aware diagnosis. VES-RFT [26] rewards visual-evidence sensitivity during reinforcement fine-tuning, whereas CausalLens [27] intervenes on sensitivity-selected attention heads. EnAR [28] guides generation with counterfactual visual impressions. Together, these methods motivate measuring visual dependence separately for each attribute under controlled interventions. Along a textual evidence channel, Ext2Gen [29] extracts evidence before generation. KnowFC [30] studies conflicts between external evidence and parametric knowledge. These methods use retrieved text for generation or fact verification, whereas CVE-SAI reads catalog text only after freezing the visual candidate and uses it solely to tighten admission. SelectiveNet integrates a reject option and jointly optimizes prediction and selection. Learn then Test selects predictive procedures through finite-sample tests, while Conformal Risk Control [31] controls monotone losses through calibration. Two-stage Risk Control [32] separately controls candidate retrieval and ranking quality at query time. These risk-control methods operate on query-time predictions or rankings rather than persistent catalog updates. Counterfactual dense retrieval promotes sensitivity to relevance-bearing changes and stability under irrelevant variation through unsupervised contrastive learning. CVE-SAI introduces FZD as an attribute-specific visual-dependence proxy and EGAR to use that proxy in ontology-constrained scoring. The visual candidate is frozen before necessity, retention, nuisance-transformation stability, and catalog-text conflict audits; catalog text can tighten admission but cannot revise the candidate. Independent finite-sample calibration then selects a policy under the prespecified unsafe-admission budget, and only admitted canonical values enter the isolated auto-attribute field.

3 Problem Formulation

Task and Decision. For product–attribute pair i, let I_i be the primary product image, X_i the catalog text, and a_i the target attribute. The frozen ontology is $\mathcal{V}_{a_i} = \{v_1, \ldots, v_{m_{a_i}}\}$. The visual-answer space is $\mathcal{C}_{a_i} = \mathcal{V}_{a_i} \cup \{o, \perp\}$, where o denotes an image-supported value outside the ontology and $\perp$ denotes that the primary image is insufficient to answer. A pre-split applicability table fixes which product–attribute pairs enter the visual-answer population; once included, abstention, an out-of-ontology response, model fallback, technical failure, and withholding remain in the evaluation denominator.

The visual stage predicts $\hat{y}_i \in \mathcal{C}_{a_i}$ using only I_i, a_i, and the frozen ontology. An admission policy returns $d_{\theta,i} \in \{0, 1\}$, where 1 denotes admission and 0 denotes withholding. Catalog text is unavailable until the visual candidate is frozen and may affect only $d_{\theta,i}$. Consequently, only a canonical $\hat{y}_i \in \mathcal{V}_{a_i}$ can enter the auto-attribute field.

Dual Truths and Visual Support. Image truth and item truth answer different questions. When image-truth adjudication is complete, $y_i^{\text{img}} \in \mathcal{C}_{a_i}$ records what the primary image supports. When item-truth adjudication is complete, y_i^{item} records the product fact using trusted structured fields and independent auxiliary views; the title, description, and primary image are excluded from this adjudication. Item truth is canonical, out of ontology, or indeterminate (the internal `ITEM_UNKNOWN` state). For canonical value $v \in \mathcal{V}_{a_i}$ with adjudicated evidence mask $\mathbf{G}_i(a_i, v)$, visual support is

$$g_i(a_i, v) = \mathbf{1}[y_i^{\text{img}} = v]\, \mathbf{1}[\|\mathbf{G}_i(a_i, v)\|_1 > 0]. \tag{1}$$

A canonical image truth requires a complete nonempty mask for its matching value. An out-of-ontology or visually unanswerable image truth gives zero support to every canonical value.

Observable Certification Population. Risk certification starts from a prespecified, label-blind family sample. Before any truth annotation or model output is available, one applicable pair is fixed for each product family in subset s, forming $\mathcal{D}_{\text{pre}}^{(s)}$. This selected-pair list remains fixed after annotation attrition and defines the product family as the certification unit. Let $L_i = 1$ when image truth, item truth, and the required evidence annotation are complete and item truth is canonical or out of ontology. The risk-observable population is

$$\mathcal{D}_{\text{cert}}^{(s)} = \{i \in \mathcal{D}_{\text{pre}}^{(s)} : L_i = 1\}. \tag{2}$$

Incomplete truth, indeterminate item truth, or incomplete evidence annotation causes unreplaced attrition. Membership in $\mathcal{D}_{\text{cert}}^{(s)}$ depends only on label observability; model fallback, refusal, failure, and withholding remain in its denominator. The certification analysis assumes independently and identically sampled family-level units.

Unsafe Admission and Objective. For an admitted canonical candidate, factual risk captures disagreement with the product fact and grounding risk captures the

absence of visual support:

$$\begin{aligned} R_i^{\text{fact}} &= \mathbf{1}[\hat{y}_i \neq y_i^{\text{item}}], \\ R_i^{\text{ground}} &= \mathbf{1}[g_i(a_i, \hat{y}_i) = 0], \\ R_i^{\text{unsafe}} &= R_i^{\text{fact}} \vee R_i^{\text{ground}}. \end{aligned} \tag{3}$$

The union counts an admission once even when both risks occur. Let $\mathcal{A}_\theta^{(s)} = \{i \in \mathcal{D}_{\text{cert}}^{(s)} : d_{\theta,i} = 1\}$, $N_s = |\mathcal{D}_{\text{cert}}^{(s)}|$, $n_\theta^{(s)} = |\mathcal{A}_\theta^{(s)}|$, and $k_\theta^{(s)} = \sum_{i \in \mathcal{A}_\theta^{(s)}} R_i^{\text{unsafe}}$. Coverage and the unsafe fraction among admitted values are

$$\begin{aligned} \text{Coverage}^{(s)}(\theta) &= \frac{n_\theta^{(s)}}{N_s}, \\ \text{UnsafeWWR}^{(s)}(\theta) &= \frac{k_\theta^{(s)}}{n_\theta^{(s)}}. \end{aligned} \tag{4}$$

When no value is admitted, coverage is zero and UnsafeWWR is undefined. Let $U_{\text{CP}}(\theta)$ be the simultaneous one-sided upper confidence bound defined in Section 4.4. CVE-SAI maximizes calibration coverage under the risk and minimum-admission constraints

$$\begin{aligned} \max_{\theta \in \Theta} \quad & \text{Coverage}^{(\text{cal-risk})}(\theta) \\ \text{s.t.} \quad & U_{\text{CP}}(\theta) \leq \alpha, \qquad n_\theta \geq n_{\min}. \end{aligned} \tag{5}$$

The main protocol fixes $\alpha = 0.05$ and $n_{\min} = 172$; the policy family and unique selection rule are specified below.

4 Methodology

4.1 Framework Overview

Figure 1 summarizes three stages: product-text-free visual inference, frozen-candidate auditing, and independently calibrated index admission. Given a primary image, an attribute question, and a frozen ontology, FZD constructs an attribute-specific visual-dependence proxy from a counterfactually weakened image. EGAR uses that proxy to refine ontology-constrained scores, after which the canonical gate immediately freezes the candidate. The auditing block then measures evidence necessity, evidence retention, nuisance-transformation stability, and candidate-specific catalog-text conflict. The final block applies the independently selected policy and writes only admitted values to the isolated auto-attribute field.

Candidate generation and visual auditing use the primary image, attribute question, ontology, and frozen verbalizers. Catalog text is read only after the candidate and visual scores are frozen; it can raise admission requirements but cannot regenerate, replace, or edit the frozen value.

4.2 Counterfactual Visual Candidate Generation

Focus-Zone Distortion. Raw question-to-visual attention may cover the complete product rather than the region that resolves the requested attribute. FZD uses the question-conditioned attention map $A(I) \in \mathbb{R}_+^{N_v}$ to produce a localized counterfactual view I^{cf}, as shown in the FZD block of Figure 1. It retains the positive attention decrease caused by weakening that region and normalizes over the visual patches:

$$\begin{aligned} \Delta_j &= \big[A_j(I) - A_j(I^{\text{cf}})\big]_+, \\ E_j &= \frac{\Delta_j}{\sum_{r=1}^{N_v} \Delta_r}. \end{aligned} \tag{6}$$

The proxy E is attribute-specific because both the probe attention and the counterfactual intervention are conditioned on the attribute question. It is computed before any candidate is frozen. A nonpositive normalization term or inconsistent patch geometry invalidates the route and leads to withholding.

Evidence-Guided Attention Redistribution. EGAR strengthens patches supported by E without discarding the model's original visual distribution. In the EGAR block of Figure 1, let q_{probe} be the normalized probe map, q_t the visual-key distribution at a teacher-forced token row, and z_t its attention logits. On the same patch grid, EGAR computes

$$\begin{aligned} \gamma &= \gamma_{\max} \frac{\text{JSD}(q_{\text{probe}}, E)}{\log 2}, \\ q_{t,j}^\star &= (1-\gamma) q_{t,j} + \gamma E_j, \\ z_{t,j}^\star &= z_{t,j} + \log \frac{q_{t,j}^\star}{q_{t,j}}. \end{aligned} \tag{7}$$

The Jensen–Shannon divergence makes the correction adaptive while keeping $\gamma \leq \gamma_{\max}$. The same pair-level γ is used for every teacher-forced row and ontology verbalizer in the fixed attention subset; all other heads and nonvisual logits remain unchanged. Because q_t and $q_t^\star$ are normalized, the correction preserves the total exponential mass over visual keys.

Ontology-Constrained Freezing. Each ontology value, the out-of-ontology outcome, and abstention has one frozen verbalizer. CVE-SAI scores them with length-normalized teacher-forced log-likelihoods and a temperature-scaled softmax. A fixed ontology order breaks exact ties, and a canonical value is returned only when its probability reaches the refusal threshold. The canonical gate then freezes $(\hat{y}, p_0, E, \gamma)$, where p_0 is the original-image probability of $\hat{y}$. Every later view rescores this same value. An out-of-ontology response, abstention, base fallback, or invalid route is withheld before auditing.

4.3 Frozen-Candidate Auditing

Evidence Necessity and Retention. As shown in the auditing block of Figure 1, the evidence mask M_E is the smallest set of patches whose descending E mass reaches 0.70, followed by one-patch dilation on the visual grid. The necessity view I^{-E} weakens the masked region,

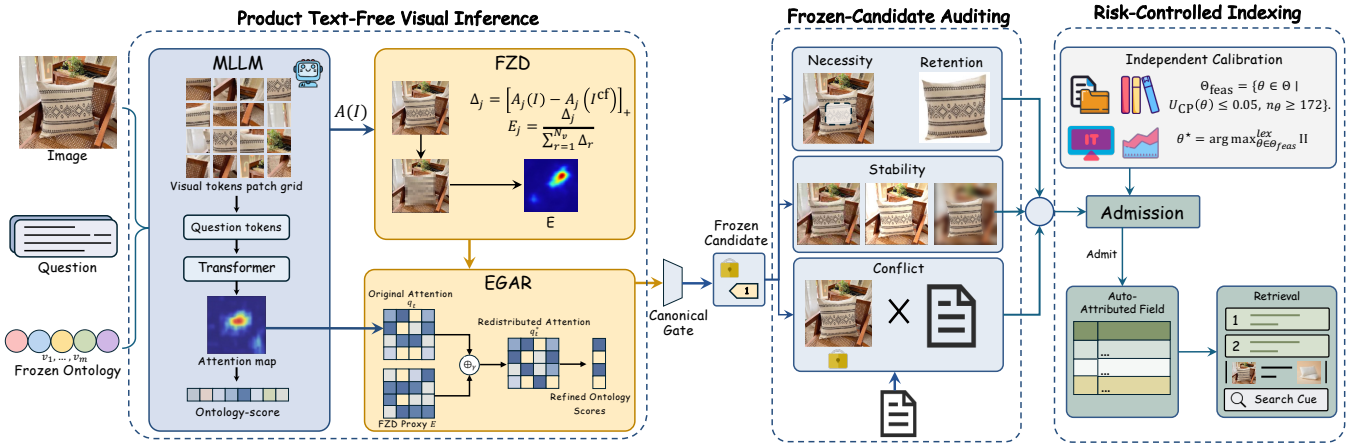

Figure 1: Overview of CVE-SAI. Given a primary image, an attribute question, and a frozen ontology, FZD constructs the attribute-specific visual-dependence proxy E, and EGAR redistributes visual attention to refine ontology-constrained scores. The resulting canonical candidate is frozen before evidence necessity, evidence retention, nuisance-transformation stability, and candidate-specific catalog-text conflict audits. Independent calibration selects the admission policy under the 5% risk budget and the minimum-admission requirement; only admitted values enter the isolated auto-attribute field for retrieval.

whereas the retention view I^{+E} preserves it and weakens the complement. Both views reuse the frozen candidate, proxy, redistribution strength, mask, and scorer. Let $s_J(\hat{y})$ be the candidate probability on view J and $c_J^\dagger$ its top candidate. Writing $s^- = s_{I-E}(\hat{y})$, $s^+ = s_{I+E}(\hat{y})$, and $c_+^\dagger = c_{I+E}^\dagger$, the evidence scores are

$$\begin{aligned} S_{\text{nec}} &= [p_0 - s^-]_+, \\ S_{\text{ret}} &= \mathbf{1}[c_+^\dagger = \hat{y}]\big(1 - [p_0 - s^+]_+\big), \\ S_{\text{evi}} &= \min(S_{\text{nec}}, S_{\text{ret}}). \end{aligned} \tag{8}$$

Necessity rewards a support decrease after removing the proposed evidence; retention requires the same candidate to remain top-ranked when that evidence is preserved.

Nuisance-Transformation Stability. Visual support should survive changes unrelated to the target attribute. Let $I^{(1)}, I^{(2)}, I^{(3)}$ be the frozen JPEG-compression, linear-RGB brightness, and outside-mask blur views. With $s_k = s_{I^{(k)}}(\hat{y})$ and top candidate $c_k^\dagger$, the nuisance-transformation stability score is

$$S_{\text{sta}} = \frac{1}{3}\sum_{k=1}^{3} \mathbf{1}[c_k^\dagger = \hat{y}]\big(1 - |s_k - p_0|\big). \tag{9}$$

A top-candidate change contributes zero. The mask and all three transformed views are mandatory; an invalid element fails the common technical gate.

Candidate-Specific Catalog-Text Conflict. After the visual scores are fixed, an ontology-bound parser compares the frozen candidate with the title and description and returns severity $S_T \in [0,1]$. Missing text yields $S_T = 0$ and is recorded as unobservable. Because the threshold adjustment below is nonnegative, catalog conflict can only make admission more demanding; it leaves $\hat{y}$, E, M_E, and every visual score unchanged.

4.4 Risk-Controlled Index Admission

Admission Policy. Let $\mathcal{B} = \{\text{evi}, \text{sta}\}$. Each policy $\theta \in \Theta$ has base thresholds τ_b^0 and nonnegative conflict penalties λ_b, and adjusts the two audit thresholds as

$$\tau_b^\theta(x) = \text{clip}\big(\tau_b^0 + \lambda_b S_T(x), 0, 1\big), \qquad b \in \mathcal{B}. \tag{10}$$

A technically valid canonical candidate is admitted only when both S_{evi} and S_{sta} reach their adjusted thresholds; all other candidates are withheld. Section 5.2 fixes the complete finite grid before risk labels are opened.

Finite-Sample Calibration. Let $\mathcal{D}_R = \mathcal{D}_{\text{cert}}^{(\text{cal-risk})}$, $M = |\Theta|$, and $0 < \alpha, \delta < 1$. For policy θ, n_θ is the number admitted in $\mathcal{D}_R$ and k_θ is the number unsafe. With $\eta = \delta/M$, the Bonferroni-corrected [34] one-sided Clopper–Pearson bound [9] is

$$U_{\text{CP}}(\theta) = \begin{cases} 1, & n_\theta = 0 \text{ or } k_\theta = n_\theta, \\ \mathrm{B}_{1-\eta}^{-1}(k_\theta + 1, n_\theta - k_\theta), & \text{otherwise}, \end{cases} \tag{11}$$

where B_u^{-1} is a Beta quantile. The minimum count is fixed as

$$n_{\min} = \left\lceil \frac{\log(\delta/M)}{\log(1-\alpha)} \right\rceil = 172 \quad (\alpha = \delta = 0.05,\ M = 324). \tag{12}$$

Thus Figure 1, Eq. (5), and the formal certification rule use the same feasible set:

$$\begin{aligned} \Theta_{\text{feas}} &= \{\theta \in \Theta : U_{\text{CP}}(\theta) \le \alpha,\ n_\theta \ge 172\}, \\ \theta^\star &= \operatorname*{arg\,max}^{\text{lex}}_{\theta \in \Theta_{\text{feas}}} \Pi(\theta). \end{aligned} \tag{13}$$

The lexicographic tuple $\Pi(\theta)$ orders larger calibration coverage, smaller upper risk bound, larger minimum

Table 1: Statistics of the ABO-derived benchmark. Product families define the data splits and resampling units.

Split	Families	$\mathcal{D}_{\text{elig}}$	$\mathcal{D}_{\text{pre}}$	$\mathcal{D}_{\text{cert}}$
Development	8,000	25,312	8,000	7,424
Validation	3,000	9,492	3,000	2,784
Calibration	5,000	15,820	5,000	4,640
Test	4,000	12,656	4,000	3,712
Total	20,000	63,280	20,000	18,560

admitted score margin, and earlier preregistered policy ID. This order makes $\theta^\star$ unique. The selected policy is applied to test without refitting or recertification. If no policy is feasible, the frozen fallback admits no automatic attributes.

Isolated Index Update. Only an admitted canonical value emits a token to the searchable auto-attribute field. Audit scores and version identifiers remain in a nonsearchable provenance record. The auto-attribute field is isolated from merchant-provided fields and can be rolled back without modifying them.

5 Experimental Setup

5.1 Datasets and Tasks

We construct five visual-attribute tasks from Amazon Berkeley Objects (ABO): color, pattern, item shape, finish type, and style. Table 1 reports label-blind product-family splits. The test split contains 12,656 eligible product–attribute pairs from 4,000 families. The preselected certification list fixes one pair per family before annotation or inference; $\mathcal{D}_{\text{cert}}$ retains pairs with complete image truth, canonical or out-of-ontology item truth, and complete evidence annotation. Indeterminate or incomplete truth and incomplete evidence cause unreplaced attrition, while visual unanswerability remains eligible. Two annotators label independently and a third adjudicates disagreements. Localization uses 750 validation pairs and a disjoint 1,000-pair test set, balanced across the five attributes and with one pair per family. Controlled retrieval uses 3,712 test certification listings and 800 query families: 120 attribute queries and 40 controls for validation, plus 480 attribute queries and 160 controls for test. Predictions, refusals, and technical failures remain in all applicable denominators.

5.2 Frozen Models, Parameters, and Metrics

The primary CVE-SAI backbone is Qwen2.5-VL-3B-Instruct [37]; InternVL3-2B [38] is used only in the cross-backbone analysis. On the primary backbone, $A(I)$ averages question-token attention to visual tokens over decoder layers 28–35 and all 16 attention heads. FZD weakens the top 20% of patches under $A(I)$ with Gaussian blur $\sigma = 12$ pixels, and EGAR uses $\gamma_{\max} = 0.35$. The evidence mask uses the 0.70 cumulative-mass and one-patch-dilation rule in Section 4.3; necessity and retention use the same $\sigma = 12$ blur. The three nuisance-transformation stability views use JPEG quality 50, a 1.15× linear-RGB brightness multiplier, and outside-mask Gaussian blur $\sigma = 8$. Candidate probabilities use $T_{\text{cal}} = 0.85$ and refusal threshold $\tau_{\text{ans}} = 0.55$. Catalog fields are Unicode-normalized and matched against frozen ontology verbalizers before assigning S_T. Calibration families are partitioned 30%/10%/60% for probability calibration, policy construction, and independent risk certification. The policy grid is $\tau^0_{\text{evi}} \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, $\tau^0_{\text{sta}} \in \{0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$, $\lambda_{\text{evi}} \in \{0, 0.10, 0.20\}$, and $\lambda_{\text{sta}} \in \{0, 0.05, 0.10\}$, giving $M = 324$ and $n_{\min} = 172$. With $\alpha = \delta = 0.05$, calibration selects $\theta^\star = (0.15, 0.85, 0.10, 0.05)$ in the order $(\tau^0_{\text{evi}}, \tau^0_{\text{sta}}, \lambda_{\text{evi}}, \lambda_{\text{sta}})$ before test. We report Macro VAA and Ans.-F1 for attribute inference, Macro Patch AUPRC for evidence localization, CWC@5% and UnsafeWWR for admission, and NDCG@10 [45] and UAIE@10 for retrieval.

5.3 Baselines and Retrieval Protocol

We compare contrastive encoders CLIP [35], SigLIP2 [36], and FashionCLIP [46], together with GME [47], MM-Embed [48], InternVL3, Qwen2.5-VL, and MOON. Every method receives the primary image, attribute question, and frozen ontology, freezes one candidate, and uses the same family-level risk-selection protocol; external systems use scalar confidence, while CVE-SAI uses FZD, EGAR, and the four audits. For localization, we compare FZD with raw question-to-visual attention, Attention Rollout [39], and Grounding DINO [40]. Retrieval uses Lucene 9.12.1 and BM25 [43] with $k_1 = 1.2$ and $b = 0.75$. Titles and descriptions use the English analyzer; merchant and auto-attribute fields use keyword tokenization with lowercasing and accent folding. Query boosts are 2.0 for title, 1.0 for description, and 2.5 for each attribute field. Three-level qrels are fixed before automatic attributes are generated: grade 2 requires product-type relevance and an exact trusted target-attribute match, grade 1 denotes product-type relevance without a verified exact match, and grade 0 denotes nonrelevance. Each automatic run is paired with the same system's No-Auto run, which differs only by leaving the auto-attribute field empty. UAIE@10 counts top-10 positions newly occupied by grade-0 products through unsafe automatic values. The 160 test controls contain no target-attribute cue, so the query builder omits both merchant-attribute and auto-attribute clauses. Ties use ascending listing ID.

6 Results and Analysis

We first examine attribute inference and evidence localization, then test whether these gains yield broader

Table 2: Results on the ABO-derived test sets: (a) attribute inference and (b) evidence localization. Values are percentages; best and second-best results are shown in bold and underlined.

(a) Attribute inference

Method	Macro VAA↑	Ans.-F1↑
SigLIP2	57.26	72.44
InternVL3	65.42	78.88
Qwen2.5-VL	67.84	80.31
MOON	<u>69.41</u>	<u>81.72</u>
CVE-SAI Front End	**72.36**	**84.47**

(b) Evidence localization

Method	Patch AUPRC↑
Grounding DINO	30.84
Attention Rollout	<u>38.27</u>
FZD	**49.16**

certified admission and stronger retrieval. Further analyses cover individual components, attributes, backbones, risk budgets, and withholding outcomes.

6.1 Attribute Inference and Evidence Localization

Table 2(a) compares image-only attribute inference under the shared image-only input protocol. On 12,656 eligible test pairs, CVE-SAI obtains 72.36% Macro VAA and 84.47% Ans.-F1, outperforming the strongest baseline, MOON, by 2.95 and 2.75 percentage points. These differences correspond to relative improvements of 4.25% and 3.37%, respectively. The simultaneous gains indicate that the front end improves exact ontology-value prediction without trading away recognition of image-unanswerable cases. Because every eligible pair remains in the denominator, fallback and failed-inference outcomes directly lower both scores rather than being removed from evaluation.

Table 2(b) evaluates evidence localization with Macro Patch AUPRC. FZD reaches 49.16%, improving over Attention Rollout by 10.89 points and Grounding DINO by 18.32 points, or 28.46% and 59.40% in relative terms. Generic object localization tends to cover the complete product, although a queried attribute may depend on a small pattern, material, or surface region. FZD instead assigns patch importance according to the positive decrease in question-conditioned visual attention caused by the focus-zone intervention. As shown in the front-end path of Figure 1, the resulting attribute-specific visual-dependence proxy is computed before candidate freezing and guides EGAR in refining ontology-constrained scores. The gain in Table 2(a) is therefore accompanied by a substantially more selective estimate of where the supporting visual information lies.

Table 3: Risk-controlled index admission on 3,712 test certification pairs. CWC@5% uses policies selected under the shared 5% risk protocol; UnsafeWWR is observed on test. Values are percentages; best and second-best results are shown in bold and underlined.

Method	CWC@5%↑	UnsafeWWR↓
CLIP	20.47	3.03
SigLIP2	24.65	2.95
FashionCLIP	25.67	2.94
GME	27.02	2.91
MM-Embed	26.35	2.97
InternVL3	28.53	2.88
Qwen2.5-VL	29.09	3.06
MOON	<u>33.19</u>	<u>2.84</u>
CVE-SAI	**44.50**	**2.36**

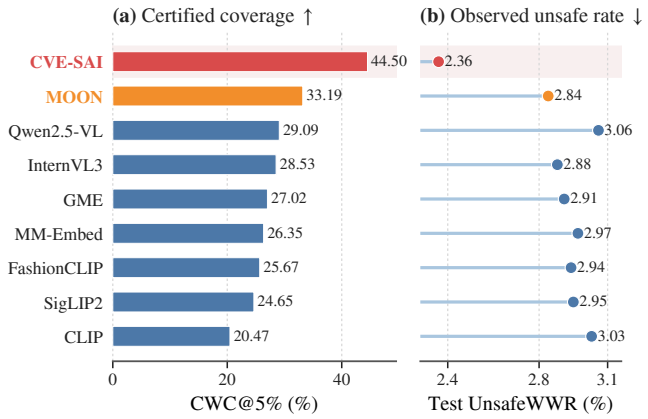

Figure 2: Risk-controlled index admission under the shared 5% certification protocol: (a) CWC@5% and (b) observed test unsafe-admission rate. Higher CWC@5% and lower UnsafeWWR are better.

Answerability and localization capture different failure modes. Ans.-F1 penalizes a system that answers when the primary image is insufficient or refuses when the attribute is visible, whereas Patch AUPRC tests whether the estimated support coincides with the annotated attribute region. CVE-SAI improves both quantities, so its higher answer accuracy is not produced by answering more pairs without regard to visual support. This distinction is important for indexing: a correct-looking value still requires localized evidence before it can become a persistent retrieval signal.

6.2 Risk-Controlled Index Admission

Table 3 and Figure 2 report risk-controlled index admission under the shared 5% unsafe-admission budget. The comparison spans contrastive encoders, product-specific representations, universal multimodal embedders, and generative multimodal models under the common protocol described in Section 5.3.

For CVE-SAI, the frozen $\theta^\star = (0.15, 0.85, 0.10, 0.05)$ admits 1,184 risk-certification candidates, of which 32 are unsafe. Its empirical UnsafeWWR is 2.70%, while the selector-corrected one-sided bound is 4.830%. The 2.13-point gap reflects finite-sample uncertainty and simultaneous search over the prespecified $M = 324$ policies. The bound remains 0.170 points below the 5% limit; this same policy is then applied to test without refitting or recertification.

After transfer to the test population, the observed 2.36% unsafe rate is 0.34 points below the calibration empirical rate and 2.47 points below its corrected upper bound. The same frozen thresholds produce these test outcomes; no test-specific operating point is selected. The calibration result and the subsequent test count therefore describe one policy from selection through test evaluation.

CVE-SAI attains 44.50% CWC@5%, compared with 33.19% for MOON. This is an 11.31-point absolute gain and a 34.08% relative increase over the strongest external system. Its observed UnsafeWWR simultaneously decreases from 2.84% to 2.36%, a 0.48-point absolute and 16.90% relative reduction. Thus, the additional coverage is not obtained by admitting a larger unsafe fraction.

The underlying counts make this difference concrete. On 3,712 test certification pairs, CVE-SAI admits 1,652 candidates, including 1,613 that are both factually correct and visually supported and 39 that are unsafe. MOON admits 1,232 candidates, with 1,197 safe and 35 unsafe admissions. CVE-SAI therefore adds 420 admitted attributes: 416 safe and four unsafe. Safe values account for 99.05% of this incremental set, and the total number of safe admitted attributes increases by 34.75%. These gains directly enlarge the searchable attribute inventory while preserving a lower unsafe proportion.

When normalized by the complete test certification population, CVE-SAI supplies 43.45 safe admitted values per 100 eligible certification pairs, compared with 32.25 for MOON. The resulting 11.21-point gain closely tracks the 11.31-point CWC improvement because nearly all additional admissions are safe. At the same time, the unsafe count rises by only four while the admitted set grows by 420. The selector therefore uses the available risk budget primarily to recover useful attributes rather than to relax support requirements uniformly.

Scalar confidence alone cannot distinguish a plausible canonical value from one that lacks attribute-relevant visual support. CVE-SAI evaluates evidence necessity, evidence retention, nuisance-transformation stability, and candidate-specific catalog-text conflict after candidate freezing. The first two audits test whether removing or retaining the localized region changes support in the expected direction; the nuisance-transformation stability audit rejects values that react excessively to nuisance variation; and the conflict audit raises the admission requirement when catalog text disagrees with the frozen value. Their combination separates visual support from answer confidence and explains why coverage can rise while observed unsafe admission falls.

Table 4: Controlled Lucene retrieval on 480 attribute-bearing test queries using the admitted values from Table 3. A dash denotes no automatic-attribute exposure; best and second-best results are shown in bold and underlined.

Method	NDCG@10↑	UAIE@10 (%)↓
No-Auto	0.6128	–
CLIP	0.6226	2.06
SigLIP2	0.6318	1.78
FashionCLIP	0.6345	1.64
GME	0.6402	1.49
MM-Embed	0.6376	1.57
InternVL3	0.6465	1.27
Qwen2.5-VL	0.6499	1.25
MOON	<u>0.6587</u>	<u>0.91</u>
CVE-SAI	**0.6749**	**0.50**

The external front ends all use the same risk-selection procedure, so calibration alone cannot account for the gap in Table 3. Their CWC@5% values range from 20.47% to 33.19%, even though their observed UnsafeWWR values remain within a narrow 2.84%–3.06% interval. CVE-SAI moves beyond this cluster in both directions, reaching 44.50% coverage at 2.36% observed unsafe admission. The evidence audits provide additional ordering information among candidates whose scalar scores alone offer similar risk–coverage choices.

The minimum-admission condition prevents a policy from appearing certifiable by writing only a handful of easy cases. The selected policy admits 1,184 risk-certification candidates, well above the required 172, so its operating point is not determined by the count floor. Its 4.830% upper bound instead reflects the 32 observed unsafe admissions, the admitted sample size, and the Bonferroni correction over the frozen policy family. CWC@5% therefore evaluates the coverage of one statistically eligible policy rather than a threshold chosen retrospectively from test outcomes.

6.3 Retrieval Effectiveness

Table 4 and Figure 3 report the retrieval effect of the attributes admitted in Table 3. Each policy supplies canonical values to the same isolated auto-attribute field under the fixed Lucene configuration.

CVE-SAI reaches 0.6749 NDCG@10. This exceeds MOON by 0.0162, or 2.46%, and the paired No-Auto reference by 0.0621, or 10.13%. Within each paired run, the corpus, analyzers, BM25 parameters, query, qrels, boosts, and tie breaking are fixed; the only changed input is that system's admitted auto-attribute tokens. Under this controlled protocol, the within-system ranking difference is therefore attributable to those tokens rather

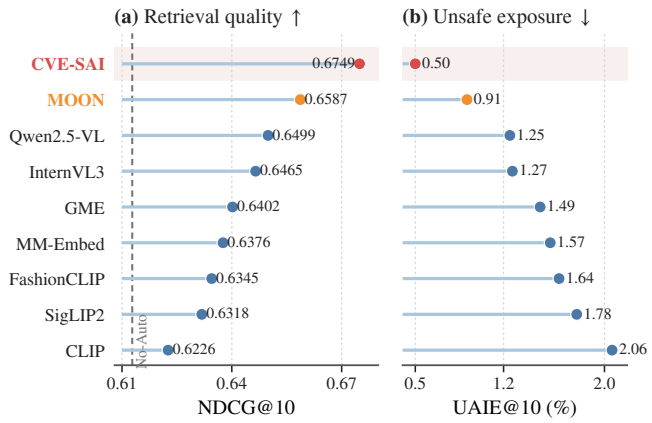

Figure 3: Controlled retrieval comparison: (a) NDCG@10 with the No-Auto reference and (b) unsafe auto-induced exposure among automatic-admission systems. Higher NDCG@10 and lower UAIE@10 are better.

than to a changed retrieval pipeline.

The progression among the strongest systems further connects admission quality to retrieval utility. Qwen2.5-VL reaches 29.09% CWC@5% and 0.6499 NDCG@10; MOON increases these values to 33.19% and 0.6587; CVE-SAI reaches 44.50% and 0.6749. Unsafe exposure decreases over the same sequence from 1.25% to 0.91% and then 0.50%. Thus, the ranking gain is associated with a larger pool of admitted attributes whose unsafe share is lower, rather than with indiscriminate expansion of the auto-attribute field.

CVE-SAI also obtains the lowest UAIE@10 among automatic-admission systems. Its value of 0.50% is 0.41 percentage points below MOON, a 45.05% relative reduction. This result complements NDCG@10: the system retrieves more relevant products while exposing fewer non-relevant products through unsafe automatic attributes. The admission gain in Table 3 therefore survives downstream use, improving ranking quality instead of merely increasing the number of indexed values. The 160 control queries activate neither attribute field; every paired ranking is unchanged, with ΔNDCG@10$=0.0000$ and UAIE@10$=0$. The retrieval effect is therefore confined to queries that contain the evaluated attribute cue. The attribute and control queries serve complementary roles: attribute queries activate the exact-match clauses and test the intended retrieval route, whereas controls retain the same corpus and lexical pipeline without an attribute cue. Their zero paired differences rule out score changes caused by analyzer drift, index construction, or listing-ID tie handling. The NDCG@10 and UAIE@10 changes are observed only when the auto-attribute field is eligible to contribute.

Table 5: Component analysis under the shared 5% risk protocol. Each variant is recalibrated independently; best and second-best results are shown in bold and underlined.

Variant	CWC@5%↑	NDCG@10↑
Raw attention for FZD	35.64	0.6578
w/o attention redistribution	38.36	0.6632
Necessity only	31.84	0.6582
Retention only	33.62	0.6604
w/o nuisance-transformation stability	38.79	0.6658
w/o conflict audit	<u>41.11</u>	<u>0.6687</u>
Full CVE-SAI	**44.50**	**0.6749**

6.4 Component Analysis

The variants in Table 5 correspond to the visual-inference and auditing components shown in Figure 1. Replacing FZD with raw attention reduces CWC@5% from 44.50% to 35.64% and NDCG@10 from 0.6749 to 0.6578, absolute drops of 8.86 points and 0.0171. Removing EGAR lowers the same metrics to 38.36% and 0.6632, drops of 6.14 points and 0.0117. FZD thus supplies a more useful attribute-specific visual-dependence proxy than raw attention, while EGAR converts that proxy into better ontology-constrained scores before freezing the candidate.

Necessity alone trails the full system by 12.66 coverage points and 0.0167 NDCG@10, while retention alone trails it by 10.88 points and 0.0145. Each audit observes only one side of the intervention: necessity measures the effect of removing evidence, whereas retention checks whether the retained region preserves the candidate. Their joint use provides a stronger admission signal than either direction separately.

Removing nuisance-transformation stability decreases CWC@5% by 5.71 points and NDCG@10 by 0.0091. Removing candidate-specific catalog-text conflict produces smaller but consistent drops of 3.39 points and 0.0062. The latter audit acts only on admission, leaving the frozen visual value unchanged. Finally, an empirical threshold without finite-sample certification reaches 50.03% coverage but also 5.82% UnsafeWWR. The 5.53-point coverage increase over full CVE-SAI comes with an unsafe rate 0.82 points above the budget. Finite-sample calibration is therefore essential to the observed balance between admitted coverage and unsafe admission.

The ordering of the ablations separates upstream evidence quality from downstream admission checks. Raw attention causes the largest drop among the two front-end variants, followed by removing EGAR, which agrees with the localization advantage in Table 2(b). Among the audits, using only one counterfactual direction loses more coverage than removing either nuisance-transformation stability or candidate-specific catalog-text conflict from the complete selector. Necessity and retention therefore form the core visual-support test, while these two audits recover complementary failure modes that remain after localized evidence has been assessed.

Every structural variant is recalibrated independently

on the same family-level splits, rather than inheriting thresholds optimized for the full system. The comparison therefore asks whether each altered score set can support a high-coverage policy under the same 5% risk protocol. The aligned reductions in CWC@5% and NDCG@10 show that upstream evidence quality and downstream audits affect the usefulness of the resulting ordering even after each variant receives its own best certifiable operating point.

6.5 Cross-Attribute and Backbone Robustness

Across the five predefined attributes, observed coverage ranges from 36.30% for style to 50.26% for color, a span of 13.96 points. Observed UnsafeWWR ranges from 1.28% to 3.89%, and even the largest slice-level value remains 1.11 points below 5%. All slices inherit the same globally selected policy, with no attribute-specific threshold. Color therefore offers more admissible visual evidence than style under the shared rule, while the unsafe proportions remain consistently low across attributes.

The 2.61-point spread in observed UnsafeWWR is substantially smaller than the 13.96-point coverage spread. The shared policy consequently adapts mainly through the number of candidates satisfying its evidence conditions, not through large attribute-wise changes in the unsafe fraction. For a catalog containing heterogeneous attributes, slices that satisfy the common evidence criteria more often contribute more indexable values, while the same rule withholds a larger share of the remaining slices.

Qwen2.5-VL-3B-Instruct is the primary backbone. On the secondary InternVL3-2B backbone, applying the CVE-SAI front end raises Macro VAA from 65.42% to 69.81%, an absolute gain of 4.39 points and a relative gain of 6.71%. The corresponding gain with Qwen2.5-VL is 4.52 points, differing by only 0.13 points. InternVL3 with CVE-SAI reaches 0.6691 NDCG@10, 0.0226 above the unmodified InternVL3 retrieval result and only 0.0058 below the primary CVE-SAI system. The candidate-generation and admission design therefore transfers across the two evaluated multimodal backbones.

6.6 Risk-Budget Sensitivity

The risk budget controls how much certified coverage the selector can retain. Tightening the budget to 1% yields 31.25% test coverage, the primary 5% budget yields 44.50%, and relaxing it to 10% yields 64.01%. Coverage increases by 13.25 points from 1% to 5% and by a further 19.51 points from 5% to 10%, for a total span of 32.76 points. The 5% operating point retains substantially more coverage than the 1% setting while its selected policy has a 4.830% calibration bound. This operating point is used for every main comparison and all downstream retrieval runs.

Measured per additional percentage point of risk budget, the first interval recovers 3.31 coverage points and the second recovers 3.90. The selector therefore has a broad set of candidates near the stricter evidence thresholds, but admitting them requires a correspondingly looser certified risk level. Fixing the operating point before the test evaluation prevents this trade-off from being chosen retrospectively from test coverage.

The sensitivity curve is a genuine risk–coverage trade-off rather than a cost-free gain. The 10% setting admits many candidates that fail the stricter evidence requirements of the primary policy, but it also permits a larger certified unsafe-admission budget. Conversely, the 1% setting withholds additional visually plausible values to obtain a tighter guarantee. The 5% operating point is therefore interpreted only as the prespecified deployment budget used for the main study, not as evidence that the same threshold is optimal for every catalog or application.

6.7 Withholding Analysis

The frozen policy withholds 2,060 of the 3,712 test candidates, or 55.50%. Insufficient evidence or nuisance-transformation stability accounts for 1,301 cases (63.16%), making it the dominant reason. Another 376 candidates (18.25%) fall outside the frozen ontology, and 180 (8.74%) are withheld after candidate-specific catalog-text conflict raises the admission requirement. These three substantive outcomes comprise 90.15% of all withheld cases and prevent unsupported, noncanonical, or conflict-sensitive values from becoming searchable attributes.

The remaining 167 cases (8.11%) have unusable FZD or mask outputs, and 36 (1.75%) encounter another technical failure. Together they constitute 9.85% of withheld candidates. Catalog text remains asymmetric in this analysis: it can make admission more demanding but cannot replace the frozen visual value. The distribution of outcomes shows that withholding is driven mainly by evidence quality and ontology fit rather than by technical attrition.

Evidence or nuisance-transformation stability insufficiency alone is 6.41 times as frequent as the two technical categories combined. Out-of-ontology values and catalog conflicts account for another 556 cases, or 26.99% of all withholding. These counts locate the principal constraint on coverage in ambiguous visual support and catalog compatibility, while implementation failures form a comparatively small tail. The withheld set therefore reflects the intended selectivity of the admission stage rather than a large loss of otherwise usable candidates to system errors.

The three final dispositions exhaust the 3,712-pair certification population: 1,613 safe admissions, 39 unsafe admissions, and 2,060 withheld candidates. Per 100 certification pairs, the policy contributes 43.45 safe indexed values, admits 1.05 unsafe values, and withholds 55.50 candidates. This partition exposes the operational trade-

off directly: coverage is limited chiefly by withholding, while unsafe additions remain a small part of the resulting auto-attribute field.

7 Conclusions and Future Work

We formulate risk-controlled selective product attribute indexing and propose CVE-SAI, which separates attribute inference from persistent index admission. FZD constructs an attribute-specific visual-dependence proxy, EGAR uses it to refine ontology-constrained scoring, and the canonical candidate is frozen before evidence necessity, evidence retention, nuisance-transformation stability, and candidate-specific catalog-text conflict audits. A finite policy family and independent one-sided calibration select a unique admission rule under a 5% unsafe-admission budget, after which admitted values enter only a reversible auto-attribute field. Across five ABO attributes, CVE-SAI improves attribute inference and evidence localization, achieves the highest CWC@5% with the lowest observed UnsafeWWR, and provides the highest NDCG@10 with the lowest UAIE@10 among automatic-admission systems. Future work will extend the framework to multi-image products, evolving ontologies, and cross-catalog calibration while preserving the separation among inference, auditing, and admission.

References

[1] L. Yang, Q. Wang, Z. Yu, A. Kulkarni, S. Sanghai, B. Shu, J. Elsas, and B. Kanagal, "MAVE: A product dataset for multi-source attribute value extraction," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. ACM, 2022, pp. 1256–1265. [Online]. Available: https://doi.org/10.1145/3488560.3498377

[2] D. Zhang, C. Fu, Z. Nie, J. Liu, W. Guan, Y. Gao, J. Song, P. Wang, J. Xu, and B. Zheng, "MOON: Generative MLLM-based multimodal representation learning for e-commerce product understanding," in *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. ACM, 2026, pp. 924–933. [Online]. Available: https://doi.org/10.1145/3773966.3777958

[3] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing, "Mitigating object hallucinations in large vision-language models through visual contrastive decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2024, pp. 13 872–13 882. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html

[4] W. An, F. Tian, S. Leng, J. Nie, H. Lin, Q. Wang, P. Chen, X. Zhang, and S. Lu, "Mitigating object hallucinations in large vision-language models with assembly of global and local attention," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. IEEE, June 2025, pp. 29 915–29 926. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/html/An_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_with_Assembly_of_Global_CVPR_2025_paper.html

[5] J. Li, J. Zhang, Z. Jie, L. Ma, M. Li, X. Luo, and G. Li, "Cross-modal attention calibration for LVLM hallucination mitigation," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. IEEE, June 2026, pp. 40 186–40 196. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2026/html/Li_Cross-Modal_Attention_Calibration_for_LVLM_Hallucination_Mitigation_CVPR_2026_paper.html

[6] H. Chen, Q. Ai, Y. Zhou, X. Wang, Y. Liu, F. Lin, and Q. Liu, "Unsupervised dense retrieval with conterfactual contrastive learning," in *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. ACM, 2026, pp. 47–57. [Online]. Available: https://doi.org/10.1145/3773966.3778015

[7] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 2151–2159. [Online]. Available: https://proceedings.mlr.press/v97/geifman19a.html

[8] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then test: Calibrating predictive algorithms to achieve risk control," *The Annals of Applied Statistics*, vol. 19, no. 2, pp. 1641–1662, June 2025. [Online]. Available: https://doi.org/10.1214/24-AOAS1998

[9] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, December 1934. [Online]. Available: https://doi.org/10.1093/biomet/26.4.404

[10] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik, "ABO: Dataset and benchmarks for real-world 3d object understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, Louisiana, USA: IEEE, June 2022, pp. 21 094–21 104. [Online]. Available: https://openaccess.

thecvf.com/content/CVPR2022/html/Collins_ABO_Dataset_and_Benchmarks_for_Real-World_3D_Object_Understanding_CVPR_2022_paper.html

[11] S. Liu, L. Li, J. Song, Y. Yang, and X. Zeng, "Multimodal pre-training with self-distillation for product understanding in e-commerce," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. ACM, 2023, pp. 1039–1047. [Online]. Available: https://doi.org/10.1145/3539597.3570423

[12] A. Khandelwal, H. Mittal, S. Kulkarni, and D. Gupta, "Large scale generative multimodal attribute extraction for e-commerce attributes," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 305–312. [Online]. Available: https://aclanthology.org/2023.acl-industry.29/

[13] Y. Zhang, S. Wang, P. Li, G. Dong, S. Wang, Y. Xian, Z. Li, and H. Zhang, "Pay attention to implicit attribute values: A multi-modal generative framework for AVE task," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 13 139–13 151. [Online]. Available: https://aclanthology.org/2023.findings-acl.831/

[14] H. P. Zou, G. H. Yu, Z. Fan, D. Bu, H. Liu, P. Dai, D. Jia, and C. Caragea, "EIVEN: Efficient implicit attribute value extraction using multimodal LLM," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 453–463. [Online]. Available: https://aclanthology.org/2024.naacl-industry.40/

[15] H. P. Zou, V. Samuel, Y. Zhou, W. Zhang, L. Fang, Z. Song, P. S. Yu, and C. Caragea, "ImplicitAVE: An open-source dataset and multimodal LLMs benchmark for implicit attribute value extraction," in *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, August 2024, pp. 338–354. [Online]. Available: https://aclanthology.org/2024.findings-acl.20/

[16] J. Gong and H. Eldardiry, "Multi-label zero-shot product attribute-value extraction," in *Proceedings of the ACM Web Conference 2024*. ACM, 2024, pp. 2259–2270. [Online]. Available: https://doi.org/10.1145/3589334.3645649

[17] J. Gong, M. Cheng, H. Shen, P.-Y. Vandenbussche, J. Jenq, and H. Eldardiry, "Visual zero-shot e-commerce product attribute value extraction," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*. Albuquerque, New Mexico: Association for Computational Linguistics, April 2025, pp. 460–469. [Online]. Available: https://aclanthology.org/2025.naacl-industry.38/

[18] J. Gong, H. Shen, and J. Jenq, "MICE: Mixture of image captioning experts augmented e-commerce product attribute value extraction," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 1151–1160. [Online]. Available: https://aclanthology.org/2025.acl-industry.80/

[19] J. Hu, J. Gong, H. Shen, and H. Eldardiry, "Hypergraph-based zero-shot multi-modal product attribute value extraction," in *Proceedings of the ACM on Web Conference 2025*. ACM, 2025, pp. 4853–4862. [Online]. Available: https://doi.org/10.1145/3696410.3714714

[20] J. Li, Y. Li, X. Shen, C. Zhang, G. Qi, and S. Bi, "Open-world attribute mining for e-commerce products with multimodal self-correction instruction tuning," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 1702–1714. [Online]. Available: https://aclanthology.org/2025.acl-long.85/

[21] Y. Su, H. Zou, L. Sun, T. Zhang, H. Yang, C. L. Yu, D. Lo, Q. Zhang, S. Han, and J. Chen, "TACLR: A scalable and efficient retrieval-based method for industrial product attribute value identification," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 31 526–31 538. [Online]. Available: https://aclanthology.org/2025.acl-long.1521/

[22] Z. Nie and P. Sun, "HADSF: Aspect aware semantic control for explainable recommendation," in *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. ACM, 2026, pp. 509–519. [Online]. Available: https://doi.org/10.1145/3773966.3778013

[23] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, December 2023, pp. 292–305. [Online]. Available: https://aclanthology.org/2023.emnlp-main.20/

[24] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, "Multi-modal hallucination control by visual information grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2024, pp. 14 303–14 312. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html

[25] Z. Wang, S. Yang, B. Peng, Z. Tang, Y. Li, B. Dong, and J. Dong, "Same attention, different truths: Put logit-lens over visual attention to detect and mitigate LVLM object hallucination," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026, pp. 25 315–25 325. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Same_Attention_Different_Truths_Put_Logit-Lens_over_Visual_Attention_to_CVPR_2026_paper.html

[26] X. Hou, W. Li, Y. Li, H. Shu, Y. Wang, X. Chen, and S. Wang, "VES-RFT: Rewarding visual evidence sensitivity to mitigate hallucinations in large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026, pp. 4168–4177. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2026/html/Hou_VES-RFT_Rewarding_Visual_Evidence_Sensitivity_to_Mitigate_Hallucinations_in_Large_CVPR_2026_paper.html

[27] J. Ji, Q. Liu, W. Yang, and Z. He, "CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026, pp. 4199–4209. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2026/html/Ji_CausalLens_Sensitivity-Guided_Multi-Head_Causal_Intervention_for_Hallucination_Mitigation_in_Large_CVPR_2026_paper.html

[28] Y. Liang, F. Shi, R. Zhu, X. Li, X. Chen, Z. Liu, B. Li, and X. Xue, "Envision, attend, then respond: Counterfactual hallucination mitigation in large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026, pp. 18 261–18 272. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2026/html/Liang_Envision_Attend_Then_Respond_Counterfactual_Hallucination_Mitigation_in_Large_Vision-Language_CVPR_2026_paper.html

[29] H. Song, J. Choi, and M. Kim, "Aligning extraction and generation for robust retrieval-augmented generation," in *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. ACM, February 2026, pp. 596–606. [Online]. Available: https://doi.org/10.1145/3773966.3777939

[30] Y. Zhang, S. Zhou, X. Li, Z. Tian, Y. Gao, S. Zhang, W. Hou, Y. Liu, and B. Zhou, "KnowFC: Navigating knowledge conflicts in large language model-based fact-checking," in *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. ACM, February 2026, pp. 996–1006. [Online]. Available: https://doi.org/10.1145/3773966.3777935

[31] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=33XGfHLtZg

[32] Y. Xu, M. Ying, W. Guo, and Z. Wei, "Two-stage risk control with application to ranked retrieval," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, J. Kwok, Ed. International Joint Conferences on Artificial Intelligence Organization, August 2025, pp. 9104–9111, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2025/1012

[33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 586–595. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html

[34] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936. [Online]. Available: https://books.google.com/books?id=3CY-HQAACAAJ

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser.

Proceedings of Machine Learning Research, vol. 139. Virtual Event: PMLR, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[36] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," February 2025. [Online]. Available: https://arxiv.org/abs/2502.14786

[37] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-VL technical report," February 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[38] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, Z. Gao, E. Cui, X. Wang, Y. Cao, Y. Liu, X. Wei, H. Zhang, H. Wang, W. Xu, H. Li, J. Wang, N. Deng, S. Li, Y. He, T. Jiang, J. Luo, Y. Wang, C. He, B. Shi, X. Zhang, W. Shao, J. He, Y. Xiong, W. Qu, P. Sun, P. Jiao, H. Lv, L. Wu, K. Zhang, H. Deng, J. Ge, K. Chen, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang, "InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models," April 2025. [Online]. Available: https://arxiv.org/abs/2504.10479

[39] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4190–4197. [Online]. Available: https://aclanthology.org/2020.acl-main.385/

[40] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Computer Vision – ECCV 2024*, ser. Lecture Notes in Computer Science, vol. 15105. Milan, Italy: Springer, 2024, pp. 38–55. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-72970-6_3

[41] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. Sydney, Australia: PMLR, 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[42] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023. [Online]. Available: https://doi.org/10.1561/2200000101

[43] S. E. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009. [Online]. Available: https://doi.org/10.1561/1500000019

[44] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: https://www.jstor.org/stable/4615733

[45] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002. [Online]. Available: https://doi.org/10.1145/582415.582418

[46] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, and J. Tagliabue, "Contrastive language and vision learning of general fashion concepts," *Scientific Reports*, vol. 12, no. 1, p. 18958, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-23052-9

[47] X. Zhang, Y. Zhang, W. Xie, M. Li, Z. Dai, D. Long, P. Xie, M. Zhang, W. Li, and M. Zhang, "Bridging modalities: Improving universal multimodal retrieval by multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 9274–9285. [Online]. Available: https://doi.org/10.1109/CVPR52734.2025.00866

[48] S.-C. Lin, C. Lee, M. Shoeybi, J. Lin, B. Catanzaro, and W. Ping, "MM-Embed: Universal multimodal retrieval with multimodal LLMs," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=i45NQb2iKO