Rethinking Visual Privacy: A Compositional Privacy Risk Framework for Severity Assessment with VLMs

Efthymios Tsaprazlis[1], Tiantian Feng[1], Anil Ramakrishna[3], Sai Praneeth Karimireddy[1], Rahul Gupta[2], and Shrikanth Narayanan[1]

[1] University of Southern California, Los Angeles CA 90089, USA
[2] Amazon AGI
[3] Meta Superintelligence Labs
tsaprazl@usc.edu

Abstract. Existing visual privacy benchmarks largely treat privacy as a binary property, labeling images as private or non-private based on visible sensitive content. We argue that privacy is fundamentally compositional. Attributes that are benign in isolation may combine to produce severe privacy violations. We introduce the Compositional Privacy Risk Taxonomy (CPRT), a regulation-aware framework that organizes visual attributes according to standalone identifiability and compositional harm potential. CPRT defines four graded severity levels and is paired with an interpretable scoring function that assigns continuous privacy severity scores. We further construct a taxonomy-aligned dataset of 6.7K images and derive compositional risk scores. By evaluating frontier and open-weight VLMs we find that frontier models align well with compositional severity when provided structured guidance, but systematically underestimate composition-driven risks. Smaller models struggle to internalize graded privacy reasoning. To bridge this gap, we introduce a deployable 8B SFT model that closely matches frontier-level performance on compositional privacy assessment. Our dataset and models are publicly available at: https://huggingface.co/collections/timtsapras23/cprt.

Keywords: Privacy · Vision-Language Models · Taxonomy

1 Introduction

Vision-language models (VLMs) are increasingly deployed in applications that process sensitive visual data, from healthcare documentation to social media moderation. As the reasoning capabilities of such models improve over time [12, 35], so do the privacy risks associated with their training or inference data. It is now widely recognized that VLMs can infer sensitive information through reasoning [28, 30]. This creates a fundamental challenge: how can we leverage the expressive capabilities of VLMs while preventing them from revealing or amplifying privacy-related information?

Prior work on visual privacy has largely focused on detecting explicit identifiers (e.g., faces, license plates, or government IDs) while treating privacy as a binary property inherent to individual attributes [21, 24, 34]. However, this framing overlooks a critical threat. *Privacy violations often emerge through the composition of individually benign attributes* [9, 22, 29]. Consider an image of a patient's wristband, a clinic room number, and a distinctive treatment device in the background. None of these elements alone can uniquely identify the patient. In combination, however, these cues could reveal a specific medical context and institutional setting, thus collectively exposing information that no single attribute discloses on its own. In a hospital setting, this aggregation can reduce the anonymity set to a small number of relevant individuals. When linked to scheduling data or other records, the patient may become uniquely identifiable.

This compositional effect is well established in structured data. k-anonymity [29] demonstrated that 87% of the U.S. population can be uniquely identified from the seemingly innocuous combination of {*postal code, birth date, gender*}. Despite decades of research on compositional privacy in databases, existing privacy frameworks largely assign risk based on the most severe detected attribute, without modeling how weaker signals may aggregate to expose identity [21]. This limitation becomes increasingly problematic as VLMs improve in reasoning capabilities [28, 32]. Attributes once considered benign, such as demographic cues, can later become identifiable when linked to auxiliary information [14, 30]. While regulatory frameworks indeed reflect this compositional view, and GDPR [7] distinguishes between data that are inherently identifying or sensitive (Art. 9) and personal data that can link to a person's identity (Art. 4). However, none of the existing visual privacy frameworks provides a composition-aware severity assessment aligned with these legal distinctions.

We address this gap by formalizing compositional visual privacy as a graded risk assessment problem. We introduce the **Compositional Privacy Risk Taxonomy (CPRT)**, a four-level hierarchy that organizes attributes by (i) standalone identifiability and (ii) compositional harm potential under plausible auxiliary information. CPRT is paired with a continuous scoring function with provable *lexicographic dominance*, i.e., the presence of a higher-severity attribute strictly outweighs any combination of lower-severity attributes. This yields interpretable, quantitative severity estimates aligned with GDPR [7], HIPAA [33], CCPA [5] and EU AI Act [8] provisions.

We construct a dataset of 6,736 images annotated for 22 privacy attributes with ground-truth compositional scores to enable systematic evaluation of VLM privacy reasoning. Across a benchmark on eight frontier and open-weight models, we observe that, while leading systems capture coarse severity distinctions, especially under structured taxonomic guidance, they struggle in modeling intermediate, composition-driven risks. Moreover, an 8B-parameter model (Qwen3-VL [4]) approaches frontier-level performance under taxonomy prompting, suggesting that compositional privacy reasoning can be distilled for edge deployment. Overall, our proposed composition-aware risk evaluation provides empirical evidence of how modern VLMs approximate compositional privacy reasoning.

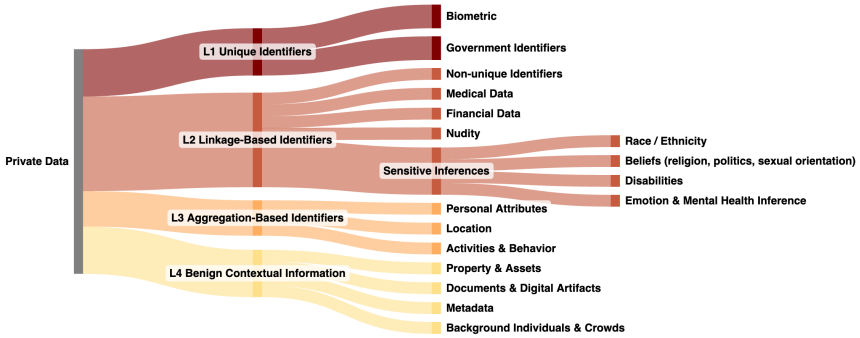

Fig. 1: Compositional Privacy Risk Taxonomy. Privacy risks are organized by inherent sensitivity and compositional identification potential. L_1 contains intrinsically identifying attributes that do not require additional information to constitute severe violations. As we move downward, both sensitivity and identifying ability decrease, while the need for composition increases. L_2 includes attributes that require minor composition to become harmful, L_3 contains attributes that become risky through aggregation and L_4 consists of attributes that contribute minimally to risk and only in specific contexts. Representative subcategories are shown on the right.

2 Related Works

2.1 Sensitive Attribute Inference with VLMs

The ability of modern foundation models to infer latent private attributes has emerged as a central privacy concern. Prior work formalizes the distinction between memorization and inference, demonstrating that large language models can deduce latent private attributes from indirect contextual cues rather than recalling them verbatim [27]. This concern extends to the visual domain, where vision–language models have been shown to infer demographics, geographic location, socioeconomic status, lifestyle, and other sensitive traits from seemingly innocuous images [30]. Beyond single-image analysis, recent studies highlight how aggregating signals across multiple images enables increasingly detailed personal profiling, underscoring the compositional nature of privacy leakage [17,28]. In response, a growing body of work focuses on auditing and mitigation. Prior efforts analyze biometric leakage and evaluate refusal behavior under privacy-sensitive prompts [15], characterize disclosure and retention risks in multimodal models [6], and explore adversarial training or machine unlearning techniques to suppress sensitive attribute inference while preserving downstream utility [1,18].

2.2 Visual Privacy Benchmarks

Early visual privacy benchmarks framed privacy as a binary image-level classification task, labeling images as private or not based on the presence of sensitive content. Datasets such as PicAlert [37], PrivacyAlert [38], VizWiz-Priv [11],

DIPA2 [34], and BIV-Priv [25] established this paradigm using coarse taxonomies and subjective annotations. These datasets have been influential, yet they collapse heterogeneous risks into binary labels and do not model how attributes interact to enable identification. VISPR [21] introduced a more structured taxonomy of 68 privacy attributes and defined an image's risk as the maximum user-rated severity among detected attributes. Still, VISPR relies on subjective preference scores and conflates inherently identifying attributes with quasi-identifiers that become harmful only through aggregation.

More recent work evaluates privacy through VLM-based protocols. PrivBench and PrivBench-H [24] provide compact, high-quality benchmarks with clearer definitions and reduced noise, showing that minimal supervision can improve VLM privacy awareness. MultiPriv [28] evaluates identity-level reasoning across multiple images, exposing risks that emerge from cross-image attribute aggregation. Visual Privacy Taxonomy [31] propose a taxonomy-grounded evaluation framework emphasizing structured attribute reasoning and regulatory alignment.

Our framework is complementary yet distinct. Rather than focusing on binary classification, subjective scoring, or cross-image attribute aggregation alone, we ground privacy assessment in compositional harm potential at the single-image level. Unlike prior benchmarks that treat privacy as binary or rely on maximum-attribute heuristics, we introduce a formally grounded continuous severity framework that explicitly models compositional harm for standalone images.

3 Compositional Privacy Taxonomy

3.1 Problem Statement

Visual privacy assessment faces a fundamental challenge, as privacy violations emerge through attribute compositions across images and data sources, more than from individual sensitive elements. Our main motivation is grounded in the same principle underlying k-anonymity [29].

We formalize visual privacy as a multi-attribute composition problem. Let $\mathcal{A} = \{a_1, \ldots, a_{|A|}\}$ denote a set of privacy-relevant attributes and an image I containing a subset $\mathcal{A}_I \subseteq \mathcal{A}$. We argue that the privacy risk of image I depends on (1) the severity of attributes present and (2) their compositional potential with attributes that an adversary may possess from auxiliary sources.

We further stratify attributes by two properties: (1) *atomic sensitivity*: whether an attribute is inherently sensitive on its own, and (2) *compositional identification potential*: whether the attribute requires auxiliary information to enable identity leakage and how such combinations amplify identifiability.

3.2 Taxonomy Structure

We define four severity levels $\mathcal{L} = \{L_1, L_2, L_3, L_4\}$ and attributes are partitioned into subcategories within each level. The full taxonomy is visualized in Figure 1.

Level 1 (L_1): Unique Identifiers Attributes that uniquely and directly identify a specific individual on their own. Examples of L_1 attributes include biometric data such as a recognizable face and fingerprints, or government-issued identifiers such as a passport number or SSN.
Level 2 (L_2) Linkage-Based Identifiers Attributes that can reference a person or reveal sensitive personal information. They may not uniquely identify an individual on their own, but can link to a person's identity with auxiliary information. For example, a full legal name is not necessarily unique, as multiple individuals may share the same name, however, it can reference a specific person within a particular system or database. Similarly, sensitive inferences such as medical conditions or mental health diagnoses do not identify an individual in isolation, but they constitute highly sensitive information that becomes harmful or discriminatory once associated with a specific identifiable person.
Level 3 (L_3) Aggregation-Based Identifiers Attributes that are non-sensitive and non-identifying in isolation, but can contribute to identity linkage or profiling when combined with other non-uniquely identifying information. For example, the triplet postal code, date of birth, gender can uniquely identify the majority of the US population [29]. Examples of L_3 attributes are age, gender, occupation, location cues, activities or lifestyle indicators.
Level 4 (L_4): Benign Contextual Information Attributes that are generally benign and non-identifying, but may be regarded as private information depending on the context (e.g., personal belongings, home interiors, metadata, or background individuals).

Each attribute a_j is assigned to a severity level $\ell(a_j) \in \{1, 2, 3, 4\}$ using a simple decision tree containing four ordered questions: (Q1) Is a_j permanently tied to one person and sufficient to identify them by itself? (Q2) Is a_j assigned to a person or does it reveal sensitive personal information? (Q3) Is a_j non-identifying, but capable of helping identify or profile someone when combined with other details? (Q4) Is a_j non-identifying and generally benign, but potentially private depending on the situation? Based on the answers, each attribute is deterministically assigned to a unique severity level, as illustrated in Figure 2.

Therefore, we model an adversary as a triplet (A, C, G): the auxiliary information available (e.g., public web data, organizational rosters), inference capabilities (e.g., face recognition, OCR, OSINT chaining), and goal (unique identification or sensitive-attribute inference). Severity levels correspond to the minimum A required to achieve G given C. L_1 attributes require no auxiliary data, L_2 require minimal linking data, L_3 require cross-source aggregation, and L_4 cannot achieve G under realistic (A, C).

3.3 Continuous Privacy Risk Scoring

We require a scoring function that, given the set of attributes present in an image I, produces a continuous privacy severity score $S \in [0, 1]$. For interpretability, we assume uniform atomic contribution among attributes within the same severity level. Accordingly, we define a scoring function $S : \mathbb{N}^4 \to [0, 1]$ that maps attribute counts (c_1, c_2, c_3, c_4) to a severity value. S is designed to satisfy three fundamental properties. First, *complete coverage*: the entire interval $[0, 1]$

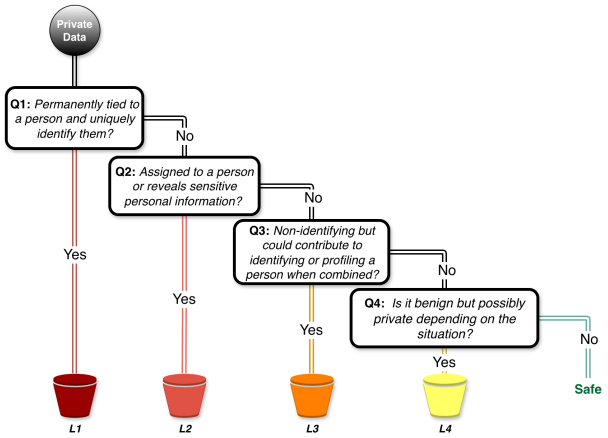

Fig. 2: Question Decision Tree. We provide a deterministic method to assign each attribute to each privacy level. This structure accommodates novel attributes without taxonomy modification, e.g., Pregnancy inference: Q1: No, Q2: Yes $\rightarrow L_2$ (sensitive when linked to identity).

must be attainable, ensuring a gap-free, continuous allocation of severity scores. Second, *lexicographic dominance*: the presence of any attribute from a higher-severity level L_i must strictly outweigh any combination of attributes from lower levels L_j, $j > i$. Consequently, an image containing at least one attribute from level L_i is always assigned a higher severity score than any image containing only attributes from lower levels. Third, *monotonicity*: increasing any attribute count c_i must strictly increase the score, guaranteeing that, within each level, additional attributes correspond to higher privacy severity. Together, these properties ensure a scoring function that spans the full $[0, 1]$ interval, preserves strict cross-level ordering, and maintains intuitive within-level scaling consistent with the hierarchical structure of the taxonomy.

Let $L = \min\{i : c_i > 0\}$ denote the determined level. We compute:

$$S_{\text{lex}} = \sum_{k=L}^{4} c_k \cdot w_k, \;\; w_i > \sum_{j>i} |A_j| w_j \tag{1}$$

where weights $w = (330, 30, 5, 1)$ (Sup. A.3) satisfy the lexicographic constraint, ensuring any L_i attribute outweighs all $L_{j>i}$ combinations. Here $|A_i|$ denotes attribute cardinality at level i. We normalize $r = S_{\text{lex}}/S^{(L)}_{\max}$ where $S^{(L)}_{\max} = \sum_{k=L}^{4} |A_k| w_k$. To eliminate gaps, we apply ratio stretching:

$$r_{\text{norm}} = \frac{r - r^{(L)}_{\min}}{1 - r^{(L)}_{\min}}, \quad r^{(L)}_{\min} = \frac{w_L}{S^{(L)}_{\max}} \tag{2}$$

We derive the final score via square-root interpolation to empirically-derived boundaries:

$$S(c_1, c_2, c_3, c_4) = b_{\min}^{(L)} + (b_{\max}^{(L)} - b_{\min}^{(L)}) \cdot \sqrt{r_{\text{norm}}} \tag{3}$$

Boundary extraction Boundaries $\{b_{\min}^{(i)}\}$ are derived from learned attribute embeddings. We train a 16-dimensional embedding space via ordinal triplet loss to cluster attributes by maximum severity level, followed by Inverse Distance Weighting (IDW) interpolation to obtain a smooth severity manifold. We then extract the 5th percentile threshold per level to define non-overlapping boundary intervals (Sup. A.2). This procedure yields the following partition of the score space: $B_1 = [0.711, 1]$ for Level 1, $B_2 = [0.514, 0.711)$ for Level 2, $B_3 = [0.292, 0.514)$ for Level 3, and $B_4 = [0, 0.292)$ for Level 4 privacy severity.

3.4 Coverage and Extensibility

Our taxonomy is grounded in existing legal frameworks, with severity levels aligned with the relative risk articulated in data protection law. Higher levels correspond to categories explicitly recognized as high-risk. For example, GDPR [7] Art. 9 (special categories of personal data) maps directly to the upper levels of our taxonomy: biometric data aligns with L_1, while health data, sexual orientation, religious or political beliefs, and ethnic origin align with L_2. More broadly, many L_2 attributes fall under GDPR Art. 4, which defines personal data whose sensitivity emerges once linked to an identifiable individual. Lower taxonomy levels align with provisions addressing contextual or indirect privacy risks. Attributes in L_3 correspond to profiling-related risks under GDPR Art. 22, where harm arises through aggregation rather than stand-alone exposure. L_4 primarily reflects context-dependent personal data. Overall, higher taxonomy levels map to more severe or high-risk processing, while lower levels correspond to contextual or ancillary categories. This consistency indicates that the ordering of our taxonomy aligns with the legal notion of privacy severity. A complete mapping to GDPR [7], EU AI Act [8], HIPAA [33] is provided in Sup. B.2.

The taxonomy is structurally robust to the addition or removal of attributes. Consider constructing the taxonomy before the introduction of the EU AI Act [8], when categories such as emotional inference were not explicitly regulated. Following the adoption of the EU AI Act, emotional inference is identified as high-risk when derived from biometric data. Within our framework, emotional inference naturally maps to L_2, and its addition does not alter the structure or ordering of the taxonomy. Moreover, high risk arises when combined with L_1 identifiers. This example shows that new regulatory concepts integrate seamlessly into the taxonomy while preserving both its structure and severity ordering.

4 Experimental Setup

4.1 Visual Privacy Risk Dataset

Dataset Generation We construct a visual privacy dataset based on VISPR [21] by filtering images to align with our compositional privacy taxonomy. We first identify images clearly containing privacy-sensitive content (e.g., faces, medical information, credit cards, passports) to populate the higher-severity categories (Levels 1 and 2). To populate the lower-severity levels (Levels 3 and 4), we sample images labeled as safe in VISPR. To better represent intermediate privacy risks, we further enrich the dataset with Level 2 attributes that do not involve direct identifiability by filtering for sensitive categories while explicitly excluding faces. This increases the coverage of compositional but non-identifying privacy risks. We use VISPR's test split for evaluation, resulting in a dataset of 6,736 images, and draw a subset from the validation split for model fine-tuning.

Annotation Protocol Each image is annotated for 22 privacy attributes corresponding to CPRT subcategories (see Sup. Table 4). Annotations are generated using GPT-5.1 and Gemini 3 Flash as primary annotators. Following Staab et al. (2023) [27], we query whether each attribute is *inferable* rather than directly inferred, allowing assessment of privacy-relevant signals while avoiding safety-induced refusals. Each annotation includes a binary decision and a short rationale to support localization and consistency checks. We adopt a three-valued annotation scheme $y \in \{0, 0.5, 1\}$, where 1 denotes clear presence, 0.5 denotes ambiguity (e.g., partial visibility or uncertain inference), and 0 denotes absence. For training and evaluation, ambiguous labels ($y = 0.5$) are conservatively mapped to 0, and $y = 1$ is retained only when both annotators agree.

Privacy Level	**Count**	**(%)**
Level 1	2,860	42.5
Level 2	1,025	15.2
Level 3	1,363	20.2
Level 4	1,488	22.1
Total	**6,736**	**100**

Table 1: Dataset distribution by privacy severity level.

Annotator	Agreement	Cohen's κ
Human-Human	**87.4%**	**0.652**
`GPT-5.1`-Human	81.5%	0.559
`Llama 4`-Human	82.4%	0.565
`Gemini`-Human	74.6%	0.391
`GPT-5.1-Llama 4`	81.0%	0.519
`GPT-5.1-Gemini`	73.7%	0.418
`Llama 4-Gemini`	75.9%	0.483

Table 2: Annotation quality metrics. Human inter-annotator agreement and model-human alignment on 65 validation images across 22 attributes.

Quality Validation To assess annotation reliability, we conduct a human validation study on 65 randomly sampled images, with 2–3 annotators per image. Due to the sensitive nature of the content, crowdsourcing is not used; instead, annotation is performed by trusted collaborators. In total, 17 participants answer the same 22 binary questions used in model-based annotation, enabling

direct comparison. Human inter-annotator agreement reaches 87.4% (Cohen's $\kappa = 0.652$), indicating moderate to high agreement. Model–human agreement is reported in Table 2, with frontier models approaching human consensus within 5–6 percentage points. Per-attribute analysis shows over 90% agreement on objective attributes (e.g., medical information, financial data, nudity), while more subjective attributes (e.g., race/ethnicity, emotional state, lifestyle) exhibit lower agreement (55–68%), reflecting inherent ambiguity.

Across severity levels, human agreement is highest for L_2 (94.5%), which includes sensitive attributes with limited interpretive variation. Agreement decreases to 84.4% for L_1, where annotators occasionally disagreed on the visibility or identifiability of faces, and to 80% for L_3, reflecting its inherently subjective and aggregation-based nature. Model–human and model–model agreement are consistently lower but follow similar trends. Agreement is highest for L_2 (78.6% model–human; 76.9% model–model on average), remains around 78–79% for L_1, and drops to 64–67% for L_4 and 57–59% for L_3, further confirming the subjectivity of aggregation-based cues. We also observe systematic calibration differences. Models tend to over-classify privacy-relevant attributes compared to humans, particularly overpredicting L_1, L_3, and L_4 attributes by more than 10%. The only consistently underpredicted category is race/ethnicity inference, which is under-classified by approximately 25.5%.

Score Derivation Privacy scores are computed from the resulting binary attribute vectors using the compositional scoring function defined in Section 3.3 (Equation 3). For each image, attribute counts across severity levels are mapped to the continuous privacy severity score in $[0, 1]$.

Dataset Statistics The final dataset contains 6,736 images distributed across CPRT severity levels as shown in Table 1.

4.2 Evaluation Metrics for Privacy Risk Assessment

To assess whether VLMs approximate compositional privacy severity, we evaluate both ranking fidelity and magnitude calibration.

Ranking Consistency We first compute Spearman correlation (ρ) between predicted scores $\hat{S}$ and taxonomy-derived scores S to measure whether models preserve the relative ordering of privacy risk across images. We further evaluate the pairwise ranking accuracy over curated inter-level and intra-level image pairs, as shown below:

$$\mathrm{acc}_{\mathrm{pairwise}} = \frac{1}{N} \sum_{(i,j)} \mathbb{1}[(S_i - S_j)(\hat{S}_i - \hat{S}_j) > 0],$$

where (i, j) denotes an image pair. This directly tests whether models discriminate between different privacy levels and capture fine-grained compositional distinctions within the same level.

Magnitude Calibration Beyond ordering, we assess whether predicted scores reflect the magnitude of privacy severity. We report Pearson correlation (r) to measure linear agreement and Mean Absolute Error (MAE) to quantify deviation from ground-truth severity. To evaluate systematic bias, we compute the mean signed error $\frac{1}{N}\sum(\hat{S} - S)$, indicating whether models consistently underestimate or overestimate privacy risk across the scale.

4.3 Implementation Details

To assess whether models internalize and reason about compositional privacy severity, we evaluate three prompting strategies that progressively incorporate our taxonomy:

- **Zero-Shot:** Models are directly asked to assign a continuous privacy severity score in $[0, 1]$ to an image without additional guidance.
- **Intuition Zero-Shot:** Prompts include high-level intuition about atomic sensitivity and compositional privacy risk as we defined in Section 3.1 , but do not reveal the explicit taxonomy structure.
- **Taxonomy-Guided:** The full compositional privacy taxonomy is provided in the prompt, enabling structured reasoning over severity levels.

In all settings, we do not provide the empirically derived score boundaries or the scoring function used to compute ground-truth severity. This prevents reverse-engineering of the scoring mechanism and allows us to observe whether models independently internalize the taxonomy and reason about privacy severity.

We evaluate several VLMs spanning open-source and proprietary systems. Open-source models (Llama 3.2-Vision-11B [3], Qwen3-VL-8B [4], MiniCPM-V 2.6 [36], Pixtral-12B [2]) are run locally using `vLLM` [16] on a single NVIDIA A100 GPU with temperature fixed to 0 for deterministic generation. Proprietary and larger models (GPT-5.2 [20], Gemini 3 Flash [10], Qwen3-VL-32B [4], Llama 4 Maverick [19]) are evaluated via API using default inference settings.

5 Results

5.1 Overall Alignment with Compositional Severity

Table 3 reports the evaluation metrics comparing model predictions against the taxonomy-derived compositional severity scores. Across models, Spearman correlation (ρ) measures preservation of relative risk ordering, while Pearson correlation (r), MAE, and bias capture magnitude alignment. Frontier models achieve strong overall alignment under taxonomy-guided prompting. Gemini 3 Flash ($\rho = 0.872$, $r = 0.884$) and GPT-5.2 ($\rho = 0.844$, $r = 0.850$) show close agreement with the theoretical severity ordering and low absolute error. These results indicate that leading models can approximate the hierarchical structure encoded in CPRT, particularly when provided structured guidance. In contrast, smaller open-weight models exhibit weaker correlations, higher error, and stronger negative bias, reflecting systematic *underestimation* of privacy severity. While some

Prompting	Model	Pearson ↑	Spearman ↑	MAE ↓	Bias	Level Acc ↑	Inter-Acc ↑	Intra-Acc ↑
Zero-Shot	Gemini 3 Flash	0.781	0.802	0.203	-0.166	0.403	0.848	0.662
	GPT-5.2	0.770	0.809	0.225	-0.197	0.316	0.884	0.645
	Llama 4 Maverick	0.673	0.695	0.255	-0.231	0.384	0.806	0.537
	Llama 3.2-VL (11B)	0.267	0.339	0.345	-0.206	0.298	0.646	0.388
	Qwen3-VL (32B)	0.603	0.724	0.292	-0.250	0.315	0.827	0.633
	Qwen3-VL (8B)	0.377	0.383	0.389	-0.370	0.290	0.665	0.575
	MiniCPM-V (8B)	0.509	0.566	0.305	-0.247	0.274	0.721	0.444
	Pixtral (12B)	0.595	0.691	0.279	-0.234	0.381	0.789	0.538
Intuition	Gemini 3 Flash	0.752	0.792	0.244	-0.220	0.302	0.857	0.639
	GPT-5.2	0.733	0.798	0.257	-0.230	0.281	0.878	0.667
	Llama 4 Maverick	0.719	0.762	0.278	-0.264	0.284	0.848	0.675
	Llama 3.2-VL (11B)	0.460	0.571	0.344	-0.304	0.299	0.729	0.478
	Qwen3-VL (32B)	0.616	0.724	0.299	-0.269	0.298	0.815	0.679
	Qwen3-VL (8B)	0.558	0.678	0.331	-0.311	0.296	0.807	0.684
	MiniCPM-V (8B)	0.616	0.610	0.237	-0.160	0.311	0.749	0.540
	Pixtral (12B)	0.622	0.716	0.286	-0.253	0.308	0.812	0.629
Taxonomy	Gemini 3 Flash	**0.884**	**0.872**	**0.140**	0.009	**0.703**	**0.938**	**0.862**
	GPT-5.2	0.850	0.844	0.158	-0.046	0.632	0.919	0.805
	Llama 4 Maverick	0.728	0.763	0.233	-0.199	0.387	0.857	0.588
	Llama 3.2-VL (11B)	0.307	0.354	0.349	-0.253	0.295	0.629	0.364
	Qwen3-VL (32B)	0.726	0.753	0.224	-0.181	0.416	0.852	0.572
	Qwen3-VL (8B)	0.636	0.751	0.291	-0.263	0.314	0.808	0.649
	MiniCPM-V (8B)	0.476	0.526	0.326	-0.252	0.371	0.714	0.418
	Pixtral (12B)	0.616	0.720	0.311	-0.290	0.293	0.802	0.658
	Qwen3-VL (8B) + SFT	0.799	0.762	**0.140**	0.061	0.633	0.849	0.745

Table 3: Compositional privacy risk evaluation grouped by prompting strategy. Frontier models achieve strong alignment with compositional risk scores under taxonomy-guided prompting, exhibiting high correlation, low error, and minimal bias. In contrast, smaller open models struggle in zero-shot settings and remain biased even under structured prompting. Taxonomy guidance yields consistent improvements across models.

models preserve coarse ordering between clearly benign and clearly severe cases, their absolute severity estimates deviate substantially from ground truth. Inter-level pairwise ranking accuracy exceeds 80% for most capable models, suggesting reliable discrimination between high- and low-severity categories. However, this aggregate alignment masks more subtle structural limitations, particularly at intermediate levels of compositional risk.

5.2 Structured Taxonomy Guidance on Compositional Severity

To assess whether compositional privacy reasoning emerges spontaneously or requires structural scaffolding, we compare zero-shot, intuition-based, and taxonomy-guided prompting. In zero-shot settings, frontier models achieve moderate correlation ($\rho = 0.70$-0.81), but exhibit strong negative bias, indicating consistent underestimation of privacy severity. Providing only high-level intuition about compositionality yields limited or inconsistent gains, suggesting that abstract descriptions alone are insufficient to induce structured reasoning. In contrast, explicitly supplying the compositional taxonomy substantially improves alignment for frontier systems, increasing correlation and reducing error. This improvement is also reflected in higher discrete level accuracy. These findings suggest that the taxonomy functions as a structural scaffold, enabling capable models to operationalize graded severity more effectively. Smaller models generally benefit less from large structured prompting, indicating constraints in their ability

to incorporate hierarchical context [23, 26, 31]. A notable exception is Qwen3-VL (8B) [4], whose correlation nearly doubles under taxonomy guidance and approaches frontier-level alignment. This demonstrates that compositional privacy assessment can, to some extent, be distilled into deployable models when provided explicit structural supervision.

5.3 Systematic Underestimation of Intermediate Privacy Risk

Although models reliably discriminate between distinct privacy levels, this performance does not extend to fine-grained distinctions within the same severity band. Intra-level pairwise ranking accuracy drops substantially, with smaller models performing near random chance. Only GPT-5.2 and Gemini 3 Flash achieve statistically significant intra-level discrimination. Our results reveal two consistent patterns. First, models implicitly treat privacy as a near-binary concept. In zero-shot settings, even clearly severe cases are compressed toward lower values. Under taxonomy prompting, extreme categories are better separated, but intermediate images frequently collapse toward one of the scale endpoints, most often the low-risk end. Figure 3 visualizes this behavior for Gemini 3 Flash. Second, models systematically underestimate compositional risks. Images whose severity arises from aggregation-based identifiers (e.g., location cues combined with demographic signals) are recognized as riskier than benign scenes, yet are positioned lower than the intermediate severity range. Conversely, models sometimes overestimate risk in cases commonly associated with high sensitivity. For example, images containing credit cards are often assigned elevated severity scores, even when the image provides limited opportunity for direct identity linkage (e.g., no visible name). In some cases, models also overpredict risk when objects merely resemble credit cards, indicating confusion driven by visual similarity rather than actual identification potential. Taken together, these findings indicate that while modern VLMs can approximate coarse compositional ordering, they lack stable internal representations for graded intermediate severity. This systematic collapse of composition-driven risk motivates the need for explicit compositional modeling as provided by CPRT.

5.4 Closing the gap between deployable VLMs and frontier models

Although closed-source frontier models demonstrate consistently stronger performance in assessing privacy risks and approximating compositional privacy severity, both under zero-shot prompting and taxonomy-guided prompting, we investigate whether fine-tuning open-source models, such as Qwen3-VL [4], can close the gap between locally deployable VLMs and frontier systems. Our goal is to determine whether a compact model can approximate the behavior of large proprietary agentic models in the VLM-as-a-Judge setting. A locally deployable model of modest scale (e.g., 8B parameters) provides an attractive alternative to perform privacy-sensitive reasoning directly on-device while maintaining reasonable computational requirements.

To explore this possibility, we create a supervised fine-tuning (SFT) dataset using taxonomy-guided instructions. Each instruction–response pair follows the

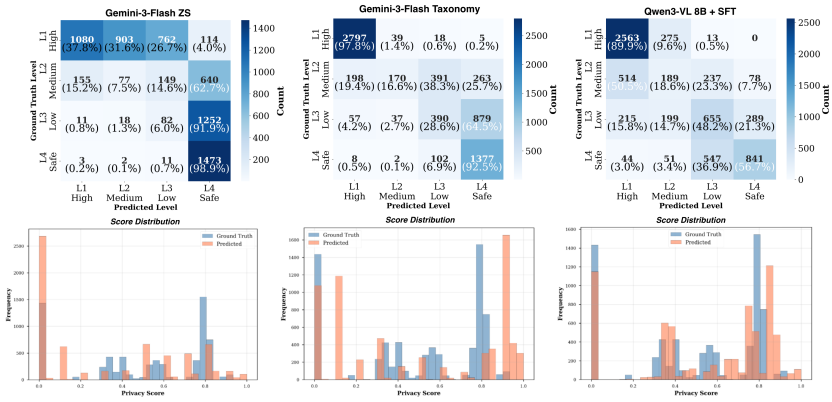

Fig. 3: Models exhibit near-binary privacy behavior. In the zero-shot setting, Gemini 3 Flash compresses most non-identifying risks toward near-zero scores while dispersing severe cases across higher levels. Taxonomy guidance improves separation, though intermediate categories remain partially collapsed. The SFT-trained Qwen3-VL 8B model achieves substantially better calibration, aligning both confusion patterns and score distributions more closely with ground truth.

prompt structure used in our earlier experiments. We fine-tune the Qwen3-VL-8B-Instruct model using low-rank adaptation (LoRA) [13], with a rank of 64 and a batch size of 128. We set the epoch for the SFT to 5 and a learning rate in $[10^{-5}, 2\times10^{-5}]$. The SFT results, presented in Table 3, indicate that fine-tuning effectively improves performance in the privacy risk evaluations. Compared with the Gemini, Qwen3-VL 8B + SFT shows strong predictions for L_3 cases, where a larger proportion of predictions fall closer to the correct levels rather than collapsing to the "Safe" category. However, Gemini 3 Flash shows a strong bias toward predicting L_4 for lower-risk categories. These results suggest that the SFT-tuned Qwen3 produces a more balanced severity distribution.

Ablation: Counterfactual Examples To evaluate whether models properly reason about intermediate privacy risks, we construct a set of counterfactual examples. Specifically, we select 100 images for which all models were highly confident in their severity assessment, predominantly L_1 cases containing faces and government-issued identifiers. Our goal is to examine whether models can recognize and appropriately weigh other privacy-relevant signals once dominant identifiers are removed. To this end, we instruct the model to assume that all visible faces are perfectly anonymized and should not be considered for identity inference. We conduct this experiment using Gemini 3 Flash.

Our results show that for images containing government-issued identifiers, Gemini 3 Flash remains robust after facial removal, maintaining high severity scores based on the remaining strong identifiers. However, in cases where privacy risk arises primarily from sensitive contextual cues (e.g., medical settings), the zero-shot model often fails to recognize the severity of the situation. While the model may acknowledge the presence of sensitive information in its reasoning, it

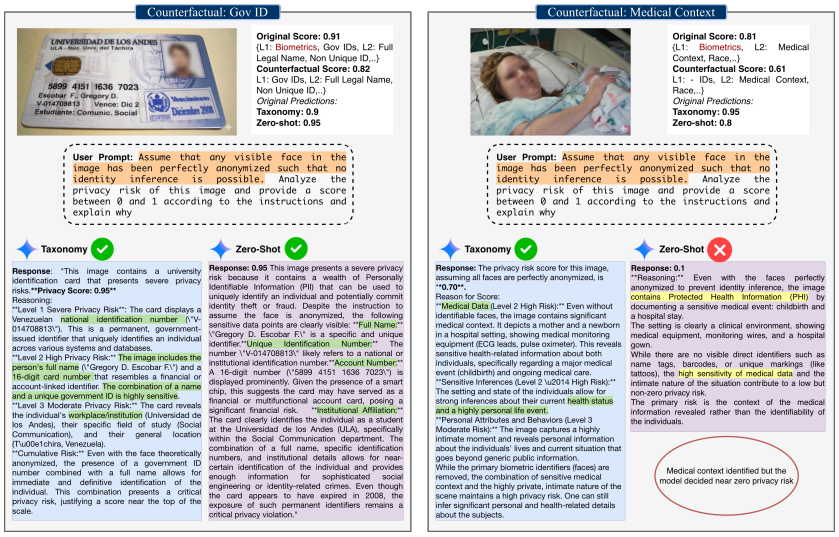

Fig. 4: Taxonomy guidance improves robustness on intermediate privacy categories. When facial cues are removed, Gemini 3 Flash preserves high scores in the presence of strong institutional identifiers (e.g., government IDs). In contrast, salient biometric cues can otherwise dominate predictions, leading to underestimation of contextual risks once removed. Structured taxonomy prompting reduces this sensitivity and improves compositional reasoning, as illustrated in the counterfactual examples.

frequently assigns a near-zero severity score if no explicit identity linkage is visible. Taxonomy-guided prompting substantially improves this behavior, enabling the model to maintain appropriately elevated severity scores in such compositional contexts and to adjust scores more consistently with the taxonomy. A qualitative example is shown in Figure 4.

6 Discussion

This work advances privacy assessment from binary classification toward a compositional and graded understanding of risk. CPRT formalizes privacy severity along two principled axes, *atomic sensitivity* and *compositional identification potential* providing a structured framework to reason about how attributes interact to produce harm. Importantly, the decision structure underlying the taxonomy is modality-invariant, and it operates over attribute properties rather than modality-specific signals, making it extensible to multimodal settings where analogous attributes can be defined.

We further argue that privacy should be evaluated in context. Whether content is considered private depends on situational, cultural, and relational factors but static attributes. Nevertheless, context-aware evaluation requires a stable foundation. CPRT provides such a foundation by offering a context-agnostic estimation of global privacy risk potential grounded in composition and sensitivity.

Finally, our results highlight the importance of deployable privacy assessment. While frontier systems demonstrate stronger compositional reasoning, we show that such capabilities can be distilled into an 8B supervised fine-tuned model that approaches frontier performance. This suggests that structured taxonomic supervision can enable practical, on-device privacy evaluation without reliance on proprietary systems.

Limitations. Our scoring mechanism assumes uniform contribution among attributes within the same severity level. In practice, attributes within a level may differ substantially in identifying power, and certain combinations may exhibit stronger synergistic effects than others. Additionally, benchmark annotation relies on GPT-5.1 and Gemini for scalability. However, the annotation and evaluation tasks are structurally different. Specifically, annotation focuses on attribute presence while evaluation requires holistic severity assessment.

7 Conclusion

We introduced the Compositional Privacy Risk Taxonomy (CPRT), a regulation-aligned framework that categorizes privacy risk. We constructed a taxonomy-aligned dataset of 6.7K images annotated across 22 privacy attributes and derived continuous privacy risk scores to enable systematic evaluation. Through comprehensive benchmarking of frontier and open-weight models, we reveal systematic limitations in modeling intermediate compositional risk. We further demonstrate that compositional privacy reasoning can be distilled into a deployable 8B vision–language model, enabling practical privacy assessment at scale. Together, these contributions establish a principled foundation for graded, composition-aware privacy evaluation and open the path toward robust, deployable multimodal systems aligned with regulatory risk frameworks.

Acknowledgments

This work was supported by USC Amazon Center for Secure and Trusted Machine Learning.

References

1. Abdulaziz, S., D'amicantonio, G., Bondarev, E.: Evaluation of human visual privacy protection: Three-dimensional framework and benchmark dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5893–5902 (2025)
2. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., Monicault, B.D., Garg, S., Gervet, T., Ghosh, S., Héliou, A., Jacob, P., Jiang, A.Q., Khandelwal, K., Lacroix, T., Lample, G., Casas, D.L., Lavril, T., Scao, T.L., Lo, A., Marshall, W., Martin, L., Mensch, A., Muddireddy, P., Nemychnikova, V., Pellat, M., Platen, P.V., Raghuraman, N., Rozière, B., Sablayrolles, A., Saulnier, L., Sauvestre, R., Shang, W., Soletskyi, R., Stewart, L., Stock, P.,

Studnia, J., Subramanian, S., Vaze, S., Wang, T., Yang, S.: Pixtral 12b (2024), https://arxiv.org/abs/2410.07073
3. AI, M.: Llama 3.2: Advancing ai for vision, edge, and mobile devices. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ (2024), accessed: 2025-03-03
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
5. California CA: California Consumer Privacy Act. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5 (2018)
6. Chen, T., Li, P., Zhou, K., Chen, T., Wei, H.: Unveiling privacy risks in multi-modal large language models: Task-specific vulnerabilities and mitigation challenges. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 4573–4586 (2025)
7. European Union: General Data Protection Regulation. https://gdpr-info.eu/ (2016)
8. European Union: Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act) (2024), http://data.europa.eu/eli/reg/2024/1689/oj, eLI: http://data.europa.eu/eli/reg/2024/1689/oj
9. Ganta, S.R., Kasiviswanathan, S.P., Smith, A.: Composition attacks and auxiliary information in data privacy. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 265–273 (2008)
10. Google Deepmind: Gemini-3-Flash: : Frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/ (2025), accessed: 2025-03-05
11. Gurari, D., Li, Q., Lin, C., Zhao, Y., Guo, A., Stangl, A., Bigham, J.P.: Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 939–948 (2019). https://doi.org/10.1109/CVPR.2019.00103
12. Hao, Y., Gu, J., Wang, H.W., Li, L., Yang, Z., Wang, L., Cheng, Y.: Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. arXiv preprint arXiv:2501.05444 (2025)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
14. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1548–1558 (2021)
15. Kim, Y., Swetha, S., Kagdi, F., Shah, M.: Safe-llava: A privacy-preserving vision-language dataset and benchmark for biometric safety. arXiv preprint arXiv:2509.00192 (2025)

16. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
17. Liu, F., Zhang, Y., Huang, X., Peng, Y., Li, X., Wang, L., Shen, Y., Duan, R., Qin, S., Jia, X., et al.: The eye of sherlock holmes: Uncovering user private attribute profiling via vision-language model agentic framework. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 4875–4883 (2025)
18. Liu, Z., Dou, G., Jia, M., Tan, Z., Zeng, Q., Yuan, Y., Jiang, M.: Protecting privacy in multimodal large language models with mllmu-bench. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 4105–4135 (2025)
19. Meta: The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/` (2025), accessed: 2025-03-05
20. OpenAI: Introducing GPT-5.2. `https://openai.com/index/introducing-gpt-5-2/` (2025), accessed: 2025-03-05
21. Orekondy, T., Schiele, B., Fritz, M.: Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In: Proceedings of the IEEE international conference on computer vision. pp. 3686–3695 (2017)
22. Patil, V., Stengel-Eskin, E., Bansal, M.: The sum leaks more than its parts: Compositional privacy risks and mitigations in multi-agent collaboration. arXiv preprint arXiv:2509.14284 (2025)
23. Pei, R., Li, H., Guo, F., Zhu, Q.: Vera: Identifying and leveraging visual evidence retrieval heads in long-context understanding. arXiv preprint arXiv:2602.10146 (2026)
24. Samson, L., Barazani, N., Ghebreab, S., Asano, Y.M.: Little data, big impact: Privacy-aware visual language models via minimal tuning. arXiv preprint arXiv:2405.17423 (2024)
25. Sharma, T., Stangl, A., Zhang, L., Tseng, Y.Y., Xu, I., Findlater, L., Gurari, D., Wang, Y.: Disability-first design and creation of a dataset showing private visual information collected with people who are blind. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2023)
26. Song, D., Chen, S., Chen, G.H., Yu, F., Wan, X., Wang, B.: Milebench: Benchmarking mllms in long context. arXiv preprint arXiv:2404.18532 (2024)
27. Staab, R., Vero, M., Balunović, M., Vechev, M.: Beyond memorization: Violating privacy via inference with large language models. arXiv preprint arXiv:2310.07298 (2023)
28. Sun, X., Li, H., Zhang, J., Yang, Y., Liu, K., Feng, R., Tan, W.J., Lim, W.Y.B.: Multipriv: Benchmarking individual-level privacy reasoning in vision-language models. arXiv preprint arXiv:2511.16940 (2025)
29. Sweeney, L.: k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems **10**(05), 557–570 (2002)
30. Tömekçe, B., Vero, M., Staab, R., Vechev, M.: Private attribute inference from images with vision-language models. Advances in Neural Information Processing Systems **37**, 103619–103651 (2025)
31. Tsaprazlis, E., Feng, T., Ramakrishna, A., Gupta, R., Narayanan, S.: Assessing visual privacy risks in multimodal ai: A novel taxonomy-grounded evaluation of

vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4292–4301 (June 2026)
32. Tu, Y., Liu, X., Qin, L., Jin, H.: Privacyreasoner: Can llm emulate a human-like privacy mind? arXiv preprint arXiv:2601.09152 (2026)
33. United States: Health Insurance Portability and Accountability Act (HIPAA). `https://www.hhs.gov/hipaa/for-professionals/index.html` (1996), accessed: 2025-03-03
34. Xu, A., Zhou, Z., Miyazaki, K., Yoshikawa, R., Hosio, S., Yatani, K.: Dipa2: An image dataset with cross-cultural privacy perception annotations. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **7**(4), 1–30 (2024)
35. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2087–2098 (October 2025)
36. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
37. Zerr, S., Siersdorfer, S., Hare, J.: Picalert! a system for privacy-aware image classification and retrieval. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 2710–2712 (2012)
38. Zhao, C., Mangat, J., Koujalgi, S., Squicciarini, A., Caragea, C.: Privacyalert: A dataset for image privacy prediction. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 16, pp. 1352–1361 (2022)