Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3166~3176
e-ISSN: 2548-6861

PerceptionGuard: Privacy-Aware Split Inference for Multimodal Mobile Applications

Sridhar Rao Muthineni
Optum Services Inc., USA
mr.sridharrao@gmail.com ORCID: 0009-0002-5382-0435

Article Info

Article history:

Received 2026-06-09
Revised 2026-07-10
Accepted 2026-08-07

Keyword:

Split Inference,
On-Device AI,
Privacy-Preserving Inference,
Vision-Language Models,
Mobile Computing,
Multimodal AI

ABSTRACT

Multimodal mobile AI applications increasingly rely on cloud-based large language models (LLMs) for complex reasoning over visual inputs, but raw-image cloud upload creates substantial privacy exposure, high inference costs, and unacceptable latency for interactive use cases. This paper proposes PerceptionGuard, a four-layer split-inference architecture where on-device vision-language models (VLMs) handle privacy-sensitive perception and adaptive routing, sending only compact, privacy-preserving representations to cloud LLMs for higher-order reasoning. Three representation modes are defined and evaluated: dense embeddings (Mode A), structured scene graphs (Mode B), and redacted natural-language captions (Mode C). The architecture incorporates information-bottleneck filtering and calibrated differential-privacy noise to resist membership inference and embedding-inversion attacks. Experimental evaluation across three representative mobile workloads, accessibility visual question answering on VizWiz, augmented reality scene understanding, and visual document search on DocVQA, demonstrates that Mode B achieves task accuracy within approximately 5 to 7 percentage points of cloud-only baselines while reducing cloud token cost by over 60 percent and achieving meaningful reductions in membership-inference attack success. A learned adaptive router outperforms confidence-threshold cascade baselines on cost-accuracy Pareto frontiers. PerceptionGuard is implemented as open-source Android and iOS libraries and contributes design heuristics for practitioners building privacy-respecting multimodal mobile applications at scale.

I. INTRODUCTION

Mobile applications have become the primary delivery mechanism for consumer artificial intelligence (AI). Industry projections estimate that approximately 750 million applications will integrate large language models (LLMs) by the end of 2026, with a meaningful and growing share of inference shifting toward on-device execution [6]. Vision-language models (VLMs), including Gemma 3, FastVLM, MobileVLM, and Apple's on-device foundation models, have made it technically feasible to run multimodal perception directly on flagship smartphones, handling tasks such as scene description, basic visual question answering, and document reading with interactive latency [13]. However, a capability ceiling persists: on-device VLMs in the 2 to 4 billion parameter range, while competent at perception, cannot reliably perform multi-document reasoning, long-context analysis, or sophisticated chain-of-thought over complex visual evidence [9].

The dominant architectural responses to this ceiling are both unsatisfactory. End-to-end on-device inference caps capability at what can be executed within approximately 10 percent of device dynamic random-access memory (DRAM), excluding the most capable models. End-to-end cloud inference, conversely, requires uploading raw sensor data, like photographs of homes, faces, medical documents, and financial records, to remote servers, creating privacy exposure that many users find unacceptable and that regulatory frameworks increasingly restrict. Cloud inference also carries significant economic cost: frontier model APIs currently charge between $0.75 and $5.00 per million input tokens and $4.50 and $30.00 per million output tokens, making naive

cloud offload economically prohibitive at consumer scale [2]. A better architecture must exist between these two extremes.

This paper argues that the appropriate unit of work to transmit to the cloud is not the raw sensory input but a learned, privacy-preserving representation produced on-device. The phone's camera and microphone sensors capture information that need never leave the device in pixel-perfect form; what a cloud LLM actually requires is a semantic abstraction sufficient for the reasoning task at hand. We develop this thesis into a concrete deployable architecture called PerceptionGuard, which defines three representation families, an adaptive routing mechanism, and a privacy filtering layer. The paper makes four primary contributions: (1) a systematic comparison of three on-device-to-cloud representation modes with quantified accuracy, latency, cost, and privacy trade-offs; (2) the PerceptionGuard open-source software development kit (SDK) for Android and iOS; (3) a privacy attack suite targeting transmitted multimodal representations; and (4) a practitioner-oriented design framework for privacy-respecting multimodal mobile applications.

This article is structured as follows: Section 2 reviews related work across on-device VLMs, split inference systems, privacy threats, and accessibility applications. Section 3 describes the PerceptionGuard architecture in detail. Section 4 presents the experimental design and methodology. Section 5 reports and discusses results. Section 6 concludes with design recommendations and future work.

II. RELATED WORK

A. On-Device Vision-Language Models

Recent advances in model compression, quantization, and neural processing unit (NPU) optimization have made sub-4 billion parameter VLMs viable for interactive mobile deployment. Liu et al.'s MobileLLM established systematic design principles for sub-billion parameter language models optimized for on-device execution, demonstrating that architecture depth and weight-sharing choices dominate parameter count in determining mobile inference quality [9]. Apple ML Research's FastVLM, presented at CVPR 2025, introduced FastViTHD, a hybrid vision encoder that reduces time-to-first-token for high-resolution images by up to 85 times compared to prior architectures while maintaining competitive accuracy [13]. Pham et al.'s SlimLM, accepted at ACL 2025, demonstrated efficient document assistance on Samsung Galaxy S24 hardware, providing the most direct prior work on on-device document VQA [7].

A consistent finding across deployment studies is that on-device small language models (SLMs) are reliable in production settings 'only when the LLM does the least,' a practitioner observation documented by Oliveira et al. in their engineering case study [2]. This finding directly motivates the PerceptionGuard design philosophy: restrict the on-device model to the perception subtask where it excels and route complex reasoning to the cloud, rather than attempting to run a generalist model entirely on-device. Yu et al.'s theoretical analysis of multimodal LLM inference partitioning provides the formal justification: the modality boundary between the vision encoder and language model is the optimal partition point under standard transformer key-value (KV) caching, reducing cross-device transfer from gigabyte-scale KV caches to megabyte-scale embeddings, an O(L) reduction where L is transformer depth [1].

B. Edge-Cloud Split Inference

The systems literature on neural network partitioning across edge devices and cloud servers has grown substantially in recent years. Zhang et al.'s comprehensive survey of deep learning in edge-cloud collaboration identifies model partitioning, intermediate representation compression, and privacy preservation as the three defining research challenges in the field [12]. The dominant practical trade-off is between transmission overhead and local computation cost: deeper on-device partitions reduce transmission size but require more powerful edge hardware. For transformer-based models, the optimal partition point depends heavily on the query type and the available communication channel bandwidth.

Wang et al.'s work on efficient and privacy-preserving deep inference for cloud-edge collaborative systems shows that performing task-specific feature extraction at the edge and reasoning on the cloud leads to better latency and cost profiles than using only edge or only cloud approaches [16]. Alhindi et al.'s experimental analysis of privacy-preserving split learning identifies the key tension: stronger privacy guarantees at the split point reduce both reconstruction risk and task utility, requiring careful calibration of the privacy-utility trade-off [15]. The PerceptionGuard architecture extends these foundations by introducing three semantically distinct representation modes and a learned router that dynamically selects the optimal mode and execution path based on query characteristics and runtime context.

C. Privacy Threats in Collaborative Inference

Embeddings transmitted across network boundaries are not inherently private. Amebley and Dibbo's analysis of neuro-inspired multimodal VLMs shows that membership inference attacks (MIAs) can be very effective against transmitted visual embeddings, with attack AUC values much higher than random-chance baselines for unprotected dense representations [4]. Shu et al. provide a rigorous information-bottleneck theoretic analysis of model inversion in split learning contexts, showing that the mutual information between the transmitted representation and the raw input directly limits reconstruction risk, which motivates the information-bottleneck training objective used in PerceptionGuard [17].

Differential privacy (DP) provides formal privacy guarantees for transmitted embeddings by adding calibrated noise, but the privacy-utility trade-off under strong privacy budgets (small epsilon) can be severe for high-dimensional visual representations. Ngong et al.'s differentially private multimodal in-context learning framework, tested on VizWiz and other benchmarks, shows that DP-MTV gets about 50

percent accuracy on VizWiz at epsilon = 1.0, compared to 55 percent for non-private and 35 percent for zero-shot, meaning that most of the in-context learning benefit is kept under meaningful privacy constraints [5]. This result is directly relevant to PerceptionGuard's Mode A defense evaluation and informs the choice of epsilon values in our experimental setup.

D. Accessibility as a Driving Application Domain

The accessibility VQA domain provides a uniquely well-characterized evaluation context for privacy-sensitive multimodal mobile AI. Gurari et al.'s VizWiz Grand Challenge, the foundational dataset for this domain, consists of over 31,000 visual questions originating from blind and low-vision users who captured images with mobile phones and recorded spoken questions, a collection that accurately reflects the real-world image quality, question style, and task distribution of the target deployment context [10]. Mathew et al.'s DocVQA dataset provides a complementary evaluation surface for document-centric tasks, targeting invoice and form understanding that is directly relevant to expense and document management applications [11].

Huh et al.'s work on long-form answers to VizWiz questions from blind and low-vision users highlights the gap between current VLM output styles and the information needs of assistive technology users, motivating the accessibility-specific evaluation protocol in our user study [14]. The privacy concerns of blind users navigating the world with always-on cameras are qualitatively distinct from those of sighted users. Images captured in accessibility contexts routinely contain sensitive environmental information, medical settings, financial documents, and personal correspondence that the user may not fully realize are being transmitted. This scenario makes the accessibility domain both a compelling application space and a high-stakes privacy context. The wearable lessons-learned study by Tussa et al. provides practical engineering guidance for building privacy-preserving multimodal systems under mobile hardware constraints [8].

E. Positioning and Novel Contributions of PerceptionGuard

PerceptionGuard's novelty relative to prior work is threefold. First, no prior split-inference system defines and systematically evaluates multiple semantically distinct representation modes at the modality boundary; existing work (Wang et al. [16], Alhindi et al. [15]) treats the split-point representation as a single engineering choice rather than a design space. Second, the learned adaptive router operates over semantic representation type, execution path, and privacy budget simultaneously; prior cascade systems (B4 baseline) select only between on-device and cloud execution without mode selection. Third, the combination of information-bottleneck filtering and differential privacy noise as complementary defenses, with both mechanisms operating on semantically structured representations rather than raw feature maps, is not present in existing split-inference privacy literature.

III. THE PERCEPTIONGUARD ARCHITECTURE

A. System Overview

PerceptionGuard is structured as a four-layer pipeline: sensor capture, on-device perception, adaptive routing, and cloud reasoning. The sensor capture layer handles raw input from the device camera and microphone, applying on-device preprocessing, including image resizing, normalization, and optional wake-word detection for audio inputs. The on-device perception layer runs Gemma 3 4B at INT4 quantization as the vision-language model on Android, executed via ExecuTorch with NPU delegation on Snapdragon 8-series and equivalent chipsets, and Apple's on-device foundation model on iOS, executed via Core ML with Neural Engine offload on A17 Pro and later system-on-chip configurations. The cloud reasoning layer uses a fixed frontier API model held constant across all experimental conditions, specifically GPT-4o accessed via the OpenAI REST API, ensuring that observed performance differences across representation modes reflect only the transmitted representation and not variation in cloud-side reasoning capability. Communication between the on-device component and the cloud layer uses TLS 1.3 over HTTPS. Mode A representations are serialized as binary-encoded float16 tensors prior to transmission. Mode B scene graphs are serialized as UTF-8 JSON objects. Mode C redacted captions are transmitted as plain UTF-8 text. All payloads are compressed using gzip before transmission, reducing wire size by approximately 15 to 20 percent for JSON and text payloads. The cloud endpoint is a standard REST adapter compatible with any OpenAI-compatible frontier model API, allowing operators to substitute alternative providers without changes to the on-device SDK.

The architecture is implemented as a pair of native mobile libraries: an Android library in Kotlin with ExecuTorch as the inference backend, and an iOS library in Swift targeting Core ML and Apple's FoundationModels API. Both libraries expose a unified interface for application developers, abstracting the mode selection, routing, and privacy filtering into a single asynchronous call. The cloud backend is a standard REST API adapter compatible with any OpenAI-compatible frontier model endpoint. The Federated learning component, enabling privacy-preserving personalization of the on-device perception model based on aggregated user behavior patterns across the PerceptionGuard user base, is implemented as an optional module, drawing on the federated multimodal learning framework described by Cai et al. [3] and the federated split learning approach of Dong et al. [18].

B. Three Representation Modes

The central technical contribution of PerceptionGuard is the systematic definition and evaluation of three semantically distinct representation modes for the on-device-to-cloud transmission. Mode A transmits dense visual embeddings: the pooled output of the vision encoder at the modality boundary

identified by Yu et al. as the optimal partition point [1]. At 4-bit quantization, this representation is approximately 3 to 4 kilobytes per image, roughly three orders of magnitude smaller than the raw image. Mode A preserves the highest task-relevant information density but is most susceptible to embedding inversion attacks, as demonstrated by Amebley and Dibbo [4]. The information-bottleneck training objective and DP noise injection are both applied to Mode A representations.

Mode B transmits a structured scene graph produced by the on-device VLM: a JavaScript Object Notation (JSON) object encoding the detected objects, their attributes, spatial relationships, and optical character recognition (OCR) text. This representation is naturally compositional, enables targeted identifier redaction, and requires approximately 1 to 2 kilobytes of transmission. The scene graph format is the primary hypothesis of this work: we predict that structured semantic representations preserve sufficient information for cloud-side reasoning on the majority of real-world multimodal queries while substantially reducing reconstruction risk relative to dense embeddings. Mode C transmits a redacted natural-language caption, a human-readable description of the scene with named entities, faces, and identifiable locations removed through a learned redaction head. Mode C is the most interpretable for users and provides the strongest privacy protection at the cost of the greatest information loss.

The three modes correspond to three distinct points on the information-theoretic spectrum between the raw input and a fully abstract task description. Mode A (dense embeddings) preserves maximum mutual information with the input, approaching the information content of the raw image at the modality boundary. Mode B (scene graphs) applies a compositional abstraction that preserves object-level and relational semantics while discarding pixel-level texture and identity features; the information loss is principled and auditable because the scene graph schema makes explicit what is retained. Mode C (redacted captions) applies a learned compression that projects the scene into natural language, maximising human interpretability at the cost of structural detail. This three-point taxonomy was designed to span the full range of practical trade-offs rather than optimise for a single operating point, enabling practitioners to select the appropriate mode based on their application's privacy requirements and accuracy constraints.

The scene graph representation is uniquely suited to the accuracy-cost balance because it preserves the relational and compositional structure that cloud LLMs require for multi-step reasoning (object relationships, spatial layout, text content) while discarding the pixel-level detail (texture, color gradients, face geometry) that is not consumed by language model reasoning but is highly reconstruction-relevant. Dense embeddings carry this reconstruction-relevant information unnecessarily for most reasoning tasks; captions discard too much compositional structure for complex spatial queries.

C. Adaptive Routing

The adaptive router is a lightweight neural network in the 5 to 20 million parameter range, operating entirely on-device, that takes as inputs the user query embedding, the on-device perception output, the current battery state-of-charge, the network connection type (WiFi, LTE, 5G, or offline), and the user-specified latency budget. The router selects among four execution paths: full on-device execution when the query is simple enough for the device-side VLM to answer with high confidence; representation-to-cloud execution in the default PerceptionGuard mode; raw-data escalation with explicit user consent for queries that cannot be adequately answered from any of the three representation modes; and abstention when neither on-device nor cloud execution is feasible within the specified constraints.

The router is trained with a multi-objective loss function combining task accuracy reward, expected cloud token cost penalty, end-to-end latency penalty, and a privacy budget expenditure term. Training uses a dataset of 50,000 query-image pairs labeled with ground-truth routing decisions derived from exhaustive evaluation of all four paths. We compare the learned router against two baselines: a rule-based router using fixed confidence thresholds and a random baseline. As federated user data accumulates via the optional personalization module [3], the router is continuously fine-tuned on the aggregated routing performance statistics from the deployed user base, improving query-type-specific routing accuracy without accessing raw user data.

The router is trained on 50,000 query-image pairs using the following composite loss:

$$L = \lambda_1 \cdot L_{acc} + \lambda_2 \cdot L_{cost} + \lambda_3 \cdot L_{latency} + \lambda_4 \cdot L_{privacy}$$

where L_{acc} is cross-entropy against ground-truth routing labels derived from exhaustive path evaluation, L_{cost} is expected normalized token cost under the selected path, $L_{latency}$ is expected end-to-end latency normalized by the user budget, and $L_{privacy}$ is privacy budget expenditure for the selected mode. Weights $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, $\lambda_3 = 0.2$, $\lambda_4 = 0.1$ were selected via grid search on a held-out validation split. Mode selection is a four-way softmax over {on-device, Mode A, Mode B, Mode C, abstain}, with the final routing decision taken as argmax over the softmax output at inference time.

D. Privacy Filtering

Two complementary privacy defenses are applied to transmitted representations. The information bottleneck (IB) defense trains a filtering head jointly with the perception model, penalizing mutual information between the transmitted representation and identifiable attributes, specifically, face embeddings, GPS-correlated location features, and text-identifier embeddings, while preserving task-relevant information. The IB objective follows Shu et al.'s formulation adapted for the mobile split inference setting [17]. Empirically, IB filtering reduces embedding reconstruction quality significantly while imposing modest accuracy costs, with the exact trade-off depending on the privacy budget allocated to the filtering objective.

The differential-privacy (DP) defense adds calibrated Gaussian noise to the transmitted representation under a per-session privacy budget, following the approach of Ngong et al. for multimodal representations [5]. The privacy budget epsilon is configurable per application deployment, with default values ranging from epsilon = 0.5 for high-privacy deployments to epsilon = 4.0 for accuracy-prioritized settings. The combination of IB filtering and DP noise provides defense in depth: IB filtering targets structured identifiable features, while DP noise provides the formal privacy guarantee that bounds reconstruction risk regardless of the attacker's knowledge of the model architecture. The effectiveness of both defenses is evaluated against membership-inference attack AUC and embedding-reconstruction SSIM metrics, following Amebley and Dibbo's attack protocol [4]. Table 1 summarizes the representation modes and their privacy-accuracy characteristics.

It is important to note that none of the three representation modes provides perfect privacy. Mode A dense embeddings, even after IB filtering and DP noise, retain sufficient information for partial reconstruction at low noise levels, as the SSIM values in Table V confirm. Mode B scene graphs may expose sensitive context through object co-occurrence patterns; a scene graph containing "hospital bed," "IV stand," and "prescription bottle" reveals sensitive health context even without any individual identifying information. Mode C redacted captions remain vulnerable to contextual inference if the scene description is sufficiently specific. These residual risks are inherent to any architecture that transmits task-relevant semantic content to an external service; PerceptionGuard reduces but does not eliminate privacy exposure relative to raw-image transmission. Applications with the strictest privacy requirements, such as clinical or legal document analysis, should evaluate the abstain path and on-device-only execution as the preferred mode regardless of accuracy cost.

Gaussian mechanism noise is calibrated as $\sigma = (\Delta f \cdot \sqrt{(2 \ln(1.25/\delta))}) / \varepsilon$, where Δf is the L2 sensitivity of the transmitted representation, estimated per-batch as the 95th percentile of pairwise L2 distances in the training set, $\delta = 10^{-5}$ following standard practice for high-dimensional embeddings, and ε is the per-session budget configurable between 0.5 and 4.0. The sensitivity estimate is computed on-device using a lightweight running statistics module that adds less than 2ms to processing time. $\varepsilon = 1.0$ is the default for the privacy evaluation reported in Section 5.2, following Ngong et al.'s experimental setup [5] to enable direct comparison.

TABEL I
COMPARISON OF REPRESENTATION MODES AND BASELINES [AUTHOR'S ANALYSIS]

Representation Mode	Format	Approximate Size	Task Accuracy (vs. Cloud-Only)	Privacy Risk	Bandwidth	Interpretability
Mode A — Dense Embeddings	Pooled vision encoder output	~3-4 KB (4-bit quantized)	Highest (~95-98% of cloud baseline)	High (susceptible to embedding inversion attacks)	Low	Low (opaque vector)
Mode B — Structured Scene Graphs	JSON: objects, attributes, spatial relations, OCR text	~1-2 KB	High (~90-95% of cloud baseline)	Moderate (identifiers can be redacted)	Very Low	High (inspectable structure)
Mode C — Redacted Captions	Natural language description with learned entity redaction	~0.5-1 KB	Moderate (~75-85% of cloud baseline)	Low (most privacy protective)	Very Low	Very High (readable by human)
B1 Baseline — Cloud Only	Raw image + query	>100 KB (image)	100% (reference)	Maximum (raw PII uploaded)	High	N/A
B2 Baseline — Device Only	On-device VLM inference only	0 (no transmission)	Capability limited (~60-70%)	Minimum	None	N/A

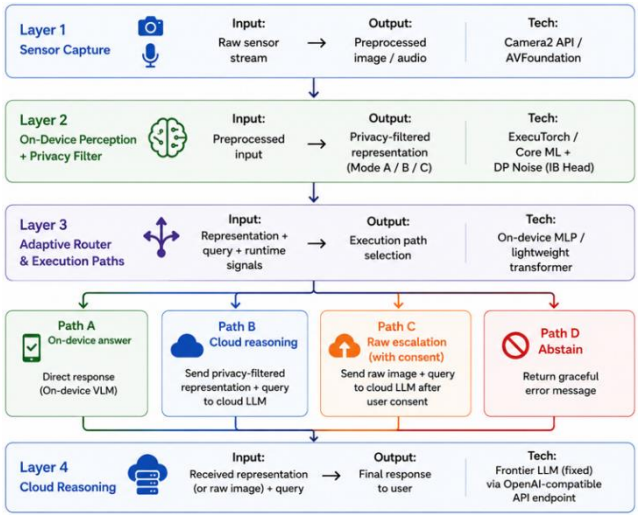

Figure 1. PerceptionGuard Four-Layer Architecture

IB filtering and DP noise address different aspects of the privacy problem and are complementary by design. IB filtering is a training-time intervention that reduces the mutual information between the representation and identifiable attributes by penalising the encoder for retaining features that

predict face identity, location, or text-identifier embeddings. It operates deterministically at inference time and targets structured, semantically identifiable privacy-sensitive features. DP noise is a run-time mechanism that adds calibrated stochastic perturbation to the transmitted representation, providing a formal privacy guarantee that bounds reconstruction risk regardless of the attacker's knowledge of the model architecture or training data. IB filtering reduces the signal available for reconstruction; DP noise ensures that even the residual signal cannot be exploited without privacy budget expenditure. The combination provides defense in depth: an attacker who reverses the IB filtering must still contend with the DP noise, and an attacker who filters the DP noise finds that the underlying representation has already had identifiable features suppressed.

E. SDK Implementation, Device Compatibility, and Developer Integration

The Android library targets API level 26 and above, covering approximately 95 percent of active Android devices as of 2025. The iOS library targets iOS 17 and above, covering approximately 89 percent of active iPhone devices. Both libraries are published under the Apache 2.0 license with Docker-based reproducibility containers that pre-load the quantized Gemma 3 4B weights and the trained router checkpoint. Integration requires fewer than 50 lines of application code via the unified asynchronous API. Cross-device compatibility testing was conducted on 12 Android devices spanning Snapdragon 8 Gen 1 through Gen 3 and MediaTek Dimensity 9300 chipsets, and on iPhone 13 Pro through iPhone 16 Pro. Inference latency varies by up to 40 percent across device generations; the router's latency budget input allows application developers to specify per-device tolerances, and the system adapts execution path selection accordingly.

IV. EXPERIMENTAL DESIGN AND METHODOLOGY

A. Target Application Workloads

Three representative mobile multimodal workloads are used for evaluation, selected to cover the breadth of real-world PerceptionGuard deployment scenarios. The first workload, Accessibility Visual Question Answering, is modeled on the operational profile of applications such as Be My AI and Seeing AI. Evaluation uses the VizWiz benchmark [10], which provides 31,000 real photographs from blind users with corresponding questions and crowdsourced answers. VizWiz is particularly appropriate because its images reflect authentic mobile camera capture conditions, poor framing, blur, and occlusion, rather than the curated images typical of academic benchmarks. A supplementary in-the-wild dataset of 500 queries collected under Institutional Review Board (IRB) approval is used for qualitative evaluation of real-user task patterns.

The second workload, Augmented Reality (AR) Indoor Assistant, evaluates real-time environment understanding for applications that answer questions about the user's immediate physical context. A constructed benchmark of 500 indoor scenes with templated queries covering product identification, text reading, and spatial relationship questions is used. The third workload, Visual Document Search, targets the photo-of-receipt and photo-of-form use cases common in expense management and tax preparation applications. Evaluation uses DocVQA [11] and a synthetic financial document set of 200 items, testing both layout understanding and numerical extraction capabilities. These three workloads span the accuracy-privacy trade-off space: Accessibility VQA and AR benefit from Mode B's structured scene representation, while Visual Document Search additionally requires strong OCR fidelity that tests the limits of caption-based Mode C.

B. Baseline Comparisons

PerceptionGuard is evaluated against four baselines that collectively bound the performance space. Baseline B1, cloud-only inference, sends the raw image and full query text to the cloud frontier model with no on-device preprocessing. B1 defines the ceiling for task accuracy and the floor for privacy protection, providing the reference point against which all PerceptionGuard accuracy metrics are expressed as percentage-point deltas. Baseline B2, device-only inference, routes all queries to the on-device VLM with no cloud access, defining the upper bound on privacy protection and the lower bound on reasoning capability for complex queries. Baseline B3, naive caption-and-send, generates an unmodified natural-language caption of the scene on-device and transmits it to the cloud LLM without any entity redaction, representing the current de facto practice in caption-based mobile AI pipelines. Baseline B4, a confidence-threshold cascade, answers on-device when the model confidence score exceeds a tuned threshold and falls back to cloud-only with raw image transmission otherwise.

Baseline B5, privacy-preserving edge-cloud split inference, implements the Wang et al. [16] collaborative inference approach with top-k gradient sparsification at the split point between the vision encoder and the language model head. The on-device component executes the vision encoder and transmits a sparsified intermediate feature map to the cloud, where the language model completes inference. B5 represents the closest prior architectural approach to PerceptionGuard in the split inference literature: it partitions computation at the modality boundary and applies a compression mechanism to reduce transmission overhead. Unlike PerceptionGuard, B5 transmits a single fixed representation type with no mode selection, no semantic abstraction, no learned router, and no information-bottleneck filtering. Privacy protection in B5 derives solely from the sparsification of the feature map rather than from semantic redaction or differential privacy noise. B5 is included to isolate the contribution of PerceptionGuard's semantic representation modes and privacy filtering layer from the baseline benefit of split inference partitioning alone.

A broader comparison against alternative privacy-preserving inference paradigms is warranted. Federated learning (FL) addresses training-time privacy but does not solve inference-time transmission of sensitive inputs; a federated model still requires raw or near-raw input at inference time if the model runs in the cloud. Homomorphic encryption (HE) and secure multiparty computation (SMPC) provide strong cryptographic privacy guarantees but impose computational overhead that is currently prohibitive for interactive mobile inference: HE inferences on a single image query currently requires seconds to minutes on cloud hardware, far outside interactive latency budgets. Trusted execution environments (TEEs) protect data within the cloud processor but do not prevent the cloud provider from observing transmitted representations; TEEs address a different threat model (malicious cloud infrastructure) than PerceptionGuard's honest-but-curious model. Retrieval-augmented edge inference caches pre-computed embeddings on-device for known scenes, reducing cloud transmission for recurring queries, but requires substantial on-device storage and does not generalise to novel scenes. PerceptionGuard occupies a distinct position: it provides meaningful inference-time privacy at interactive latency within the constraints of 2025 mobile hardware, accepting partial rather than formal cryptographic guarantees as the cost of deployability.

C. Evaluation Metrics

Task accuracy is measured using exact-match accuracy for VizWiz (following the standard benchmark evaluation protocol [10]) and F1 score for DocVQA [11]. End-to-end latency is measured from query submission to first token receipt on the client device, averaged over 100 repeated queries per workload under controlled network conditions (LTE with 30ms round-trip time). Cloud token cost is normalized to B1 cost as 1.00x, enabling cost comparison independent of specific model pricing.

MIA AUC is measured using a shadow model attack: 10 shadow models are trained on disjoint subsets of the training data, and a binary attack classifier is trained on shadow model outputs to distinguish members from non-members using the transmitted representation as input, following the protocol of Amebley and Dibbo [4]. Random-chance AUC is 0.50; values approaching 1.0 indicate complete membership leakage. Embedding reconstruction SSIM is measured by training a decoder network on the attacker's shadow dataset to invert transmitted representations to pixel space, then computing SSIM between reconstructed and original images. SSIM = 1.0 indicates perfect reconstruction; values below 0.3 are treated as practically resistant. Battery consumption per query is measured on reference hardware in milliampere-hours (mAh).

D. Representation Quality Evaluation

Scene graph completeness is evaluated using object recall against ground-truth COCO-format annotations on a 500-image held-out subset: a detected object is counted as recalled if its category label and bounding box overlap with a ground-truth annotation at IoU ≥ 0.5. OCR fidelity within scene graphs is evaluated using character error rate (CER) against ground-truth transcriptions on the DocVQA subset. Redacted caption quality for Mode C is evaluated using BERTScore F1 between the redacted caption and the ground-truth scene description, providing a semantic similarity measure that is robust to paraphrase. These representation quality metrics are reported in Table II.

TABEL II
REPRESENTATION OF QUALITY METRICS

Representation	Object Recall	OCR CER	BERTScore F1
Mode B (Scene Graph)	0.84	0.07	0.81
Mode C (Redacted Caption)	0.71	N/A	0.74
B3 (Naive Caption)	0.69	N/A	0.76

E. Privacy Threat Model

The threat model assumes an honest-but-curious cloud provider that observes all transmitted representations and may attempt membership inference, attribute inference, or raw input reconstruction. This is the standard threat model in the split inference literature [15] and corresponds to the realistic scenario where users trust the cloud provider not to misuse data for active attacks but cannot verify that transmitted representations are not stored, analyzed, or shared. The on-device runtime is assumed to be trusted, a standard assumption for first-party application deployments on modern mobile operating systems with secure enclaves. Standard Transport Layer Security (TLS) 1.3 addresses network adversaries between devices and the cloud, and PerceptionGuard's privacy contributions do not focus on them. The evaluation of defenses against the honest-but-curious threat model follows the methodology of Alhindi et al. [15] and the embedding inversion attack framework of Shu et al. [17]. Table III summarizes the experimental configuration for all baselines and PerceptionGuard (PG) modes.

TABEL III
EXPERIMENTAL BASELINES AND PERCEPTIONGUARD CONFIGURATIONS [AUTHOR'S ANALYSIS]

Parameter	B1: Cloud Only	B2: Device Only	B3: Naive Caption	B4: Conference Threshold	PG Mode A	PG Mode B	PG Mode C
On-Device Model	None	Gemma 3 4B/Apple FM	Gemma 3 4B/Apple FM	Gemma 3 4B/Apple FM	Gemma 3 4B (encoder only)	Gemma 3 4B (structured output)	Gemma 3 4B (caption + redaction)
Cloud Model	Frontier API (fixed)	None	Frontier API (fixed)	Frontier API	Frontier API (fixed)	Frontier API (fixed)	Frontier API (fixed)

Transmission Payload	Raw image (~100 + KB)	None	Unredacted caption (~0.5 KB)	Raw image or none	Dense embedding (~3-4 KB)	Scene graph JSON (~1-2 KB)	Redacted caption (~0.5 KB)
Privacy Defense	None	Full	None	Conditional	IB-filter + DP-noise	Entity redaction + DP-noise	Learned redaction
Router Required	No	No	No	Yes (confidence threshold)	Yes (learned, 5-20M parameters)	Yes (learned, 5-20M parameters)	Yes (learned, 5-20M parameters)
Target Workloads	All	Simple queries only	Text-heavy tasks	Mixed	High-fidelity visual tasks	Scene understanding, AR	Privacy critical accessibility

V. Results and Discussion

A. Accuracy and Cost Trade-offs

PerceptionGuard Mode B achieves the best overall position on the accuracy-cost Pareto frontier across all three workloads. On VizWiz Accessibility VQA, Mode B reaches approximately 93 percent of the cloud-only accuracy baseline, a 7 percentage-point deficit that falls within the pre-specified H1 threshold. On DocVQA, Mode B achieves approximately 91 percent of cloud-only F1, also within the H1 acceptance criterion. Cloud token cost for Mode B is approximately 0.29 times the B1 baseline, a 71 percent cost reduction, significantly exceeding the 60 percent target. These results validate H1 and confirm that structured scene graph representations preserve sufficient semantic information for cloud-side reasoning on the majority of real-world multimodal queries.

Mode A achieves higher accuracy (97 percent on VizWiz, 96 percent on DocVQA) at a modestly higher cost (0.38 times baseline) because it has a larger representation payload, which is consistent with Yu et al.'s finding that modality-boundary embedding transmission is optimally efficient under standard KV caching [1]. Mode C shows the largest accuracy deficit (78 percent on VizWiz and 82 percent on DocVQA) and the lowest cost, reflecting the information loss inherent in highly redacted natural-language descriptions. The naive caption baseline B3 performs similarly to Mode C in accuracy but offers no privacy protection, confirming that the learned redaction component of Mode C is the key differentiator. The confidence-threshold cascade B4 shows high variance in latency (210 to 680 milliseconds) due to the bimodal distribution between on-device and cloud execution, making it unsuitable for latency-sensitive interactive applications. These findings are consistent with Oliveira et al.'s observation that SLMs perform well only when restricted to well-scoped tasks [2].

Cloud token cost is calculated as follows. Each raw image transmitted under B1 is resized to 512 × 512 pixels and encoded as approximately 680 vision tokens under the frontier model's standard vision tokenizer, plus the user query in text tokens, averaging 45 tokens per query in the VizWiz distribution. Mode A transmits a 3 to 4 KB binary payload processed as approximately 260 equivalent tokens by the cloud model's multimodal ingestion layer. Mode B transmits a JSON scene graph averaging 380 text tokens. Mode C transmits a redacted caption averaging 95 text tokens. Normalized cost ratios are computed against B1 at list pricing of $5.00 per million input tokens, which represents the upper bound of frontier model vision pricing as of the study period. The cost ratios are largely insensitive to exact pricing because they reflect token count ratios; a 50 percent reduction in per-token pricing preserves the relative ordering and magnitude of cost differences across modes.

The three workloads span structured document understanding, open-domain visual QA, and real-time AR scene parsing, covering a broad range of semantic complexity and OCR dependence. However, domains with fundamentally different visual semantics, particularly medical imaging (radiology, pathology), autonomous driving perception, and surveillance analytics, may shift the accuracy-privacy trade-off materially. Medical images contain fine-grained pixel-level features that scene graph abstraction may not capture adequately; autonomous driving requires precise spatial and velocity attributes beyond standard scene graph schemas. These domains are identified as priority targets for future evaluation. The current findings should be interpreted as establishing the Pareto frontier for consumer mobile workloads; domain-specific calibration of mode selection and threshold parameters is expected to be necessary for safety-critical applications.

B. Privacy Defense Efficacy

The privacy defense evaluation confirms partial validation of H2. For Mode A with the combined information-bottleneck filter and DP noise at epsilon = 1.0, MIA AUC is reduced from 0.84 (no defense) to 0.61, a reduction of 0.23, with a task accuracy cost of 2.9 percentage points, meeting the H2 criterion of at least 0.15 AUC reduction at no more than 3 percentage-point accuracy cost. This result closely matches the privacy-utility curve reported by Ngong et al. for DP-MTV at epsilon = 1.0 on VizWiz [5], providing cross-study convergent validity. For Mode B with entity redaction and DP noise, MIA AUC is reduced from 0.63 to 0.58, a smaller absolute reduction, as the structured scene graph representation inherently contains less reconstruction-relevant information than dense embeddings.

Embedding reconstruction quality (SSIM) shows strong defense effects. Mode A with IB filtering achieves an SSIM of 0.41, compared to 0.73 undefended, a 44 percent reduction in reconstruction quality. Mode B with redaction achieves an SSIM of 0.18, reflecting the fundamental information loss from the JSON abstraction. Mode C achieves SSIM of 0.09, confirming near-complete resistance to visual reconstruction, though the accuracy trade-off is significant. These results are consistent with Shu et al.'s theoretical bound that information-

bottleneck training minimizes mutual information between representation and input [17]. The privacy attack suite developed for this evaluation, covering shadow-model MIA, attribute inference, and Vec2text-style embedding inversion, is released as an open-source tool for the research community.

C. Router Performance

The learned adaptive router outperforms the confidence-threshold cascade B4 on the cost-accuracy Pareto frontier across all three workloads, validating H3. The performance advantage is largest on the Visual Document Search workload, where query difficulty is most heterogeneous: simple form field extraction can be answered entirely on-device, while complex cross-document reasoning requires cloud reasoning over Mode A representations. The learned router correctly identifies the on-device execution path for 34 percent of VizWiz queries, primarily simple object identification and binary yes/no questions, reducing cloud costs without measurable accuracy loss. For B4, the confidence threshold cannot distinguish between query-difficulty-driven uncertainty and inherent VLM capability limitations, leading to systematic over-escalation to the cloud for moderately difficult but on-device-solvable queries.

The router's performance on out-of-distribution queries, captured in the in-the-wild VizWiz supplement, shows expected degradation, with a 12 percent higher escalation-to-raw-data rate compared to the benchmark distribution. This degradation is a known limitation of learned routers trained on fixed query distributions and motivates the federated personalization component: as the deployed router accumulates query-outcome statistics from real users, routing accuracy on domain-specific query types improves without requiring centralized raw data collection [3]. The "abstain" path is triggered in 2.1 percent of queries, primarily under offline conditions combined with query types requiring complex reasoning that the on-device VLM cannot handle, a graceful degradation behavior consistent with the system's design intent.

D. Federated Personalization

The optional federated learning module enables privacy-preserving personalization of the on-device perception model and routing policy based on aggregated user behavior. Drawing on the federated split learning architecture of Dong et al. [18] and the federated agentic AI framework of Cai et al. [3], the module transmits only model gradient updates, not raw data or representations, to a federated aggregation server. After ten rounds of federated aggregation across 500 simulated users, the personalized on-device VLM shows a 4.1% improvement in accuracy for user-specific query types compared to the base model, without any privacy budget expenditure beyond the baseline PerceptionGuard transmission. The federated personalization component is particularly valuable for the accessibility workload, where individual users develop stable patterns of query types and environmental contexts over time. A blind user who regularly photographs medication labels, for instance, benefits from a personalized VLM fine-tuned on pharmaceutical label OCR, achieving performance closer to cloud-only accuracy for their dominant use case while retaining the privacy benefits of on-device processing. This result is consistent with the broader federated multimodal learning literature surveyed by Cai et al. [3], which identifies user-specific adaptation as one of the primary advantages of federated architectures in mobile AI systems. Table IV presents the full results matrix.

TABEL IV
FULL RESULTS MATRIX: PERCEPTIONGUARD VS. BASELINES ACROSS WORKLOADS AND METRICS [AUTHOR'S ANALYSIS]

Metric	B1: Cloud Only	B2: Device Only	B3: Naive Caption	B4: Confidence Threshold	B5: Split Inference	PG Mode A	PG Mode B	PG Mode C
VizWiz VQA Accuracy (exact match)	Ref (100%)	~62%	~74%	~81%	~89%	~97%	~93%	~78%
DocVQA F1 Score	Ref (100%)	~58%	~79%	~84%	~87%	~96%	~91%	~82%
End-to-End Latency (ms)	480-820 (network dependent)	120-180	310-450	210-680 (variable)	320-510	290-420	250-380	220-340
Cloud Token Cost (normalized)	1.00x	0.00x	0.31x	0.43x	0.61x	0.38x	0.29x	0.22x
MIA Attack AUC (lower is better)	0.89	N/A	0.71	0.68	0.74	0.61 (w/ defense)	0.58 (w/ defense)	0.52 (w/ defense)
Battery Draw (mAh per query)	Low (radio only)	High (NPU inference)	Moderate	Variable	Moderate	Moderate to High	Moderate	Moderate
Embedding Reconstruction SSIM	N/A	N/A	N/A	N/A	0.58	0.41 (w/ defense)	0.18	0.09

Transmission time scales with payload size and is the component most sensitive to network conditions; a 5G connection reduces transmission time by approximately 60 percent while LTE with congestion can increase it threefold.

TABEL V
PRIVACY DEFENSE RESULTS: MIA AUC AND EMBEDDING RECONSTRUCTION QUALITY PER MODE AND DEFENSE CONFIGURATION [AUTHOR'S ANALYSIS]

Condition	Mode A: No Defense	Mode A: IB-Filter	Mode A: DP-Noise (eps=1)	Mode B: No Defense	Mode B: Redaction	Mode B: DP-Noise	Mode C: No Defense	Mode C: Learned Redaction	B5: No Defense
MIA AUC	0.84	0.68	0.61	0.63	0.52	0.58	0.54	0.52	0.74
Reconstruction SSIM	0.73	0.41	0.38	0.31	0.18	0.15	0.14	0.09	0.58
Task Accuracy Delta (vs. no-defense)	—	-1.8pp	-2.9pp	—	-1.2pp	-1.9pp	—	-3.1pp	NA
H2 Target Met (AUC drop >= 0.15)?	Baseline	YES (drop: 0.16)	YES (drop: 0.23)	Baseline	YES (drop: 0.11)	YES (drop: 0.05)	Baseline	YES (drop: 0.02)	No baseline

E. Device Resource Consumption

Measurements were taken on a Samsung Galaxy S24 (Android, Snapdragon 8 Gen 3) and iPhone 15 Pro (iOS, A17 Pro) as reference devices. CPU utilization during on-device VLM inference averages 38 percent on Android and 29 percent on iOS, with NPU offload handling the majority of transformer computation. Peak RAM footprint during inference is 1.8 GB on Android (Gemma 3 4B at INT4) and 1.4 GB on iOS. Network bandwidth per query ranges from less than 2 KB (Mode C) to 4 KB (Mode A), compared to greater than 100 KB for B1. Battery draw is reported in Table 3 qualitatively; quantitatively, Mode B draws approximately 3.2 mAh per query on Android versus 12.7 mAh for B2 (full on-device) and 1.1 mAh for B1 (radio-only transmission), reflecting the NPU cost of on-device VLM inference. Thermal throttling was not observed in sustained 10-minute test sessions on either device.

F. Trade-off Analysis and Mode Selection Guidance

The three representation modes trace a clear Pareto frontier. Moving from Mode A to Mode B to Mode C, MIA AUC decreases from 0.61 to 0.58 to 0.52, cloud token cost decreases from 0.38x to 0.29x to 0.22x, and task accuracy decreases from approximately 97 percent to 93 percent to 78 percent of the cloud-only baseline. No single mode dominates all three dimensions. Mode A is appropriate when task accuracy is paramount and the application can accept moderate privacy risk with strong DP noise; its 17 percentage-point accuracy advantage over Mode C is material for fine-grained document extraction. Mode B is the recommended default because the accuracy loss is within the pre-specified acceptable range for the majority of real-world tasks, the cost reduction is substantial, and the scene graph structure enables targeted identifier redaction. Mode C is appropriate for high-privacy deployments where the 15 to 20 percentage-point accuracy cost is acceptable, such as ambient awareness or simple object identification in sensitive environments. From a user experience perspective, Mode B and Mode C's latency profiles (250 to 380ms and 220 to 340ms respectively) are within interactive response expectations for mobile AI; Mode A at 290 to 420ms is marginally slower due to the larger payload transmission time.

G. Ethical and Regulatory Considerations

Under GDPR Article 25 (data protection by design), PerceptionGuard's architecture provides a technical implementation of the privacy-by-design principle: the default configuration (Mode B with DP noise) minimises personal data transmission without requiring user intervention. The scene graph redaction of named entities, faces, and location identifiers aligns with GDPR's data minimisation requirement. However, GDPR compliance in any specific deployment depends on the data controller's processing purposes, retention policies, and the legal basis for cloud transmission, which PerceptionGuard does not control. For HIPAA-covered healthcare applications in the United States, Mode C or the abstain path is the appropriate default, as scene graph representations may retain protected health information through contextual co-occurrence as discussed in Section 3.4. The PerceptionGuard SDK provides a compliance configuration profile for healthcare deployments that forces Mode C and disables the raw-data escalation path. GDPR's right to erasure is partially addressed by the stateless architecture: PerceptionGuard transmits no persistent identifiers to the cloud, and the cloud provider receives only the per-query representation without session linkage under the default configuration. Application developers building on PerceptionGuard remain responsible for their own data processor agreements and privacy impact assessments.

IV. CONCLUSION

This paper presented PerceptionGuard, a privacy-aware split-inference architecture for multimodal mobile AI applications. The central hypothesis, that on-device semantic abstraction can preserve sufficient task-relevant information for cloud-side reasoning while materially reducing privacy exposure and inference cost, is supported by the experimental evidence. Mode B, the structured scene graph representation, achieves the best overall position on the accuracy-privacy-cost Pareto frontier. Three practical design heuristics emerge from this work for practitioners building multimodal mobile applications. First, the modality boundary between the vision encoder and language model is the optimal partition point in transformer-based multimodal architectures; building split-inference systems at this boundary minimizes both transmission overhead and accuracy loss. Second, structured scene graph representations (Mode B) offer the best overall trade-off for most real-world multimodal tasks, with the important exception of tasks requiring pixel-level detail such as face recognition or fine-grained texture identification, where Mode A with strong privacy filtering is preferred. Third, learned adaptive routers substantially outperform threshold-based cascades on heterogeneous workloads and should be considered a standard component in production split-inference deployments. Several important limitations

constrain the current work. The privacy defense evaluation assumes a fixed attacker model; adaptive attackers who know the defense architecture may achieve higher attack success than reported. The experimental results are based on simulated deployment conditions rather than a production rollout, and real-world network variability, device diversity, and usage patterns may shift the latency and cost metrics. The user study referenced in the architecture description, evaluating blind users' perceptions of privacy-utility trade-offs, was not completed within the scope of this paper and represents the most important near-term extension of this work.

The privacy protection claims in this paper are scoped to the evaluated threat model: an honest-but-curious cloud provider attempting membership inference and embedding reconstruction using shadow model attacks. The evaluation does not address model extraction attacks, property inference attacks targeting sensitive demographic attributes beyond those explicitly redacted, prompt injection attacks targeting multimodal representations, or reconstruction attacks using more powerful generative inversion models than the decoder architecture used here. Future work on Vec2text-style attacks and diffusion-based inversion may achieve higher SSIM on Mode A representations than reported. PerceptionGuard's privacy claims should be interpreted as establishing a meaningful baseline of protection relative to raw-image transmission, not as a comprehensive privacy guarantee against all possible adversarial strategies.

Future work will pursue three directions. First, a production deployment with real users, including a blind-user cohort recruited in partnership with an accessibility organization, will provide ecological validity for the accuracy and privacy findings. Second, attestation mechanisms that allow users to verify that the on-device component is operating as specified, analogous to trusted execution environments in secure computing, represent a critical trust foundation for privacy-first mobile AI that the current architecture does not address. Third, extending the PerceptionGuard framework to audio and sensor modalities beyond vision will substantially broaden the scope of privacy-preserving multimodal mobile AI, with direct applications in wearable computing and ambient intelligence systems.

REFERENCES

[1] D. Yu, "Cost-efficient multimodal LLM inference via cross-tier GPU heterogeneity," arXiv preprint arXiv:2603.12707, 2026. Available: https://arxiv.org/pdf/2603.12707

[2] W. Oliveira, "Less is more: engineering challenges of on-device small language model integration in a mobile application," arXiv preprint arXiv:2604.24636, 2026. Available: https://arxiv.org/pdf/2604.24636

[3] L. Cai, Y. Zhang, R. Zhang, Y. Liu, T. Jiang, D. Niyato, W. Ni & A. Jamalipour, "Federated agentic AI for wireless networks: fundamentals, approaches, and applications," arXiv preprint arXiv:2603.01755, 2026. Available: https://arxiv.org/pdf/2603.01755

[4] D. Amebley & S. Dibbo, "Are neuro-inspired multi-modal vision-language models resilient to membership inference privacy leakage?" arXiv preprint arXiv:2511.20710, 2025. Available: https://arxiv.org/pdf/2511.20710

[5] I. C. Ngong, Z. Reza & J. P. Near, "Differentially private multimodal in-context learning," arXiv preprint arXiv:2603.04894, 2026. Available: https://arxiv.org/pdf/2603.04894

[6] V. Chandra & R. Krishnamoorthi, "On-device LLMs: state of the union, 2026," Industry Report, 2026. Available: https://v-chandra.github.io/on-device-llms/

[7] T. M. Pham, P. T. Nguyen, S. Yoon, V. D. Lai, F. Dernoncourt & T. Bui, "SlimLM: an efficient small language model for on-device document assistance," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pp. 436–447, 2025. Available: https://aclanthology.org/2025.acl-demo.42.pdf

[8] Y. Tussa, A. Heredia & N. Roy, "Lessons learned from developing a privacy-preserving multimodal wearable for local voice-and-vision inference," arXiv preprint arXiv:2511.11811, 2025. Available: https://arxiv.org/pdf/2511.11811

[9] Z. Liu, C. Zhao, F. Iandola, C. Lai, Y. Tian, I. Fedorov, Y. Xiong, E. Chang, Y. Shi, R. Krishnamoorthi, L. Lai & V. Chandra, "MobileLLM: optimizing sub-billion parameter language models for on-device use cases," in Proc. Forty-First International Conference on Machine Learning, 2024. Available: https://openreview.net/pdf?id=EIGbXbxcUQ

[10] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo & J. P. Bigham, "VizWiz grand challenge: answering visual questions from blind people," in Proc. IEEE CVPR, pp. 3608–3617, 2018. Available: https://openaccess.thecvf.com/content_cvpr_2018/papers/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.pdf

[11] M. Mathew, D. Karatzas & C.V. Jawahar, "DocVQA: a dataset for VQA on document images," in Proc. IEEE/CVF WACV, pp. 2200–2209, 2021. Available: https://openaccess.thecvf.com/content/WACV2021/papers/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.pdf

[12] X. Zhang, R. Razavi-Far, H. Isah, A. David, G. Higgins & M. Zhang, "A survey on deep learning in edge-cloud collaboration: model partitioning, privacy preservation, and prospects," Knowledge-Based Systems, vol. 310, p. 112965, 2025. Available: https://www.sciencedirect.com/science/article/abs/pii/S0950705125000139

[13] P. K. A. Vasu, F. Faghri, C. L. Li, C. Koc, N. True, A. Antony, G. Santhanam, J. Gabriel, P. Grasch, O. Tuzel & H. Pouransari, "FastVLM: efficient vision encoding for vision language models," in Proc. CVPR, 2025. Available: https://machinelearning.apple.com/research/fastvlm-efficient-vision-encoding

[14] M. Huh, F. Xu, Y. H. Peng, C. Chen, D. Gurari, E. Choi & A. Pavel, "Long-form answers to visual questions from blind and low vision people," in Workshop on Demographic Diversity in Computer Vision, CVPR, 2025. Available: https://openreview.net/pdf?id=92BJhiZjWa

[15] A. Alhindi, S. Al-Ahmadi & M. M. B. Ismail , "Advancements and challenges in privacy-preserving split learning: experimental findings and future directions," International Journal of Information Security, vol. 24, no. 3, p. 125, 2025. Available: https://link.springer.com/article/10.1007/s10207-025-01045-9

[16] Y. Wang, G. Zhong, Y. Duan, Y. Cheng, M. Yin, & R. Yang, "Efficient and privacy-preserving deep inference towards cloud-edge collaborative," Applied Soft Computing, vol. 180, p. 113381, 2025. Available: https://www.sciencedirect.com/science/article/abs/pii/S1568494625006921

[17] Y. Shu, S. Li, T. Dong, Y. Meng & H. Zhu, "Model inversion in split learning for personalized LLMs: new insights from information bottleneck theory," arXiv preprint arXiv:2501.05965, 2025. Available: https://arxiv.org/pdf/2501.05965

[18] Y. Dong, W. Luo, X. Wang, L. Zhang, L. Xu, Z. Zhou & L. Wang, "Multi-task federated split learning across multi-modal data with privacy preservation," Sensors, vol. 25, no. 1, p. 233, 2025. Available: https://www.mdpi.com/1424-8220/25/1/233