DIFFERENTIALLY PRIVATE MULTIMODAL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly deployed on sensitive data such as medical scans and financial documents, where in-context learning offers customization without retraining. Yet existing differentially private in-context learning methods are limited to few-shot, text-only settings because privacy cost scales with the number of demonstrations, and a single image can consume hundreds of tokens. We introduce Differentially Private Multimodal Task Vectors (DP-MTV), the first framework for many-shot multimodal in-context learning with differential privacy. Rather than protecting individual demonstrations, DP-MTV aggregates hundreds of private image-text examples into a single steering vector and privatizes the aggregate. Privacy cost is paid once at construction; the released vector then answers unlimited queries at zero additional cost. Across eight benchmarks and three VLM architectures, DP-MTV matches or exceeds non-private performance on the majority of tasks at ε=2, and on classification consistently outperforms the non-private baseline. Beyond this public-data setting, a *fully-private* variant that uses no auxiliary data for head selection matches it at ε=5, (for example, 96% on Flowers-102 with Qwen-VL versus 77% for non-private MTV)with a per-layer selection scheme extending competitive accuracy to tighter budgets. A novel membership inference attack validates these guarantees: 100% success without protection drops to the 50% random-guess baseline under DP-MTV.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly deployed on sensitive data: hospitals use them to interpret patient scans, tax services to process financial documents, legal firms to analyze case files. In-context learning (ICL) allows these organizations to customize models by providing demonstration examples at inference time, without finetuning (Brown et al., 2020; Dong et al., 2024). But when demonstrations contain private information, ICL creates a direct channel for leakage. Models can memorize and expose content from demonstrations through membership inference, data extraction, and prompt leaking (Wen et al., 2024; Duan et al., 2024; Hou et al., 2025), as illustrated in Figure 1.

The risk is especially acute for multimodal data. A single medical image loaded as a demonstration contains not just visual information but identifiable details that VLMs can extract through OCR (Liu et al., 2024): patient names, diagnoses, and identifiers. Beyond text, models can infer geolocation from scene details (Mendes et al., 2024), extract private attributes from environmental cues (Tömekçe et al., 2024), and reveal biometric characteristics despite anonymization (Caldarella et al., 2024; Kim et al., 2025). An adversary with API access can query for this information directly or perform membership inference to determine whether a specific individual's data was used as a demonstration.

Differential privacy (DP) provides formal guarantees that limit what an adversary can infer about any individual in the dataset (Dwork et al., 2006). Recent work has applied DP to text-based in-context learning by generating private synthetic demonstrations (Panda et al., 2023; Tang et al., 2023), aggregating outputs across disjoint exemplar sets (Wu et al., 2023; Duan et al., 2024), applying local differential privacy to labels (Zheng et al., 2024), leveraging auxiliary public data (Joo et al., 2025), and mixing few-shot with zero-shot outputs (Flemings et al., 2025). However, all existing methods are text-only and restricted to few-shot settings, because privacy cost accumulates with the number of demonstrations or tokens generated. For multimodal data, where a single image consumes hundreds of visual tokens (Bai et al., 2023), even a handful of demonstrations exhaust both context capacity

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

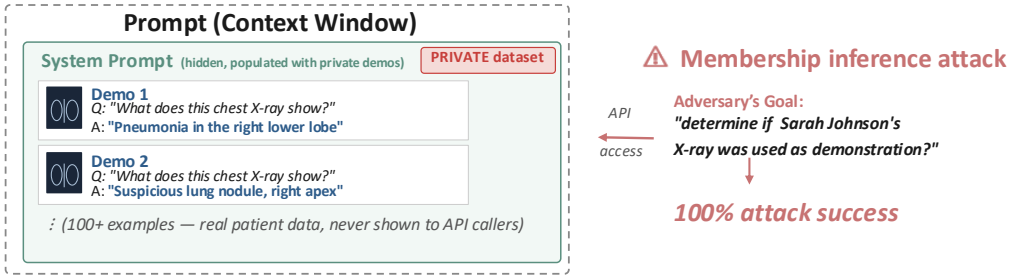


Figure 1: Many-shot in-context learning loads hundreds of private demonstrations, including patient X-rays, diagnoses, and identifiers, into the model's prompt. These demonstrations are hidden from the user interface but accessible through any API query. With a membership inference attack, an adversary can recover whether a given patient was in the demonstration set with 100% accuracy against unprotected in-context learning.

and privacy budget. . To our knowledge, no prior work addresses differentially private in-context learning for multimodal data.

One way to scale beyond context limits is Multimodal Task Vectors (MTV) (Huang et al., 2024), which aggregates activation patterns from hundreds of examples into compact steering vectors applied at inference time. This bypasses context limits entirely, but MTV provides no privacy guarantees. The vectors directly encode patterns from demonstrations, exposing them to potential inference attacks. Practitioners must therefore choose between privacy risks with many-shot learning or few-shot protection that does not extend to images.

We demonstrate the privacy risk in MTV by developing a novel membership inference attack against non-private MTV. In our empirical evaluation, our attack achieves a perfect success rate in determining membership in the set of demonstration examples for MTV, even when 100 demonstrations are used for each task vector. Our results suggest that additional defenses are needed to make MTV appropriate for settings that involve sensitive demonstrations.

To address this challenge, we introduce Differentially Private Multimodal Task Vectors (DP-MTV), the first framework enabling many-shot multimodal in-context learning with formal (ε, δ)-differential privacy. Operating in activation space fundamentally changes the privacy problem. Instead of protecting each token or demonstration individually, we aggregate activation patterns first and privatize the aggregate. DP-MTV partitions private data into disjoint chunks where each individual's data appears exactly once, clips each chunk's contribution to bound sensitivity, and adds calibrated Gaussian noise to the mean. This requires only a single private release regardless of dataset size, so privacy cost is paid once during construction. The released task vector then steers the model at inference time, answering unlimited queries at zero additional privacy cost. We propose a variant that uses a small amount of public auxiliary data for head selection, as well as a fully-private variant requiring no auxiliary data. The fully-private variant introduces a private head-selection mechanism, allowing both task-vector construction and head selection to operate entirely on private data. It recovers substantial utility at ε=5, matching or exceeding non-private MTV on several tasks (e.g., 96% vs. 77% on Flowers-102), while per-layer selection substantially improves utility at tighter privacy budgets. We evaluate DP-MTV on eight benchmarks spanning visual question answering and fine-grained classification using three VLM architectures. At $\varepsilon = 1.0$, DP-MTV achieves 49% accuracy on VizWiz compared to 55% for non-private MTV and 35% for zero-shot, preserving most of the gain from in-context learning under meaningful privacy constraints.

Contributions. In summary, we make the following contributions:

- We develop the first membership inference attack against multimodal task vectors, achieving 100% success rate across all datasets and demonstrating that unprotected many-shot ICL leaks membership information.
- We introduce DP-MTV, the first method for differentially private many-shot multimodal in-context learning. DP MTV uses disjoint data partitioning and per-layer activation clipping to bound sensitivity, producing a single private steering vector that enables unlimited inference at zero marginal privacy cost.

- We evaluate across eight benchmarks and three VLM architectures. DP-MTV preserves MTV gains at practical privacy budgets, and defeats our membership inference attack.

2 Background & Related Work

Differential Privacy. Differential privacy (DP) (Dwork et al., 2006; Dwork & Roth, 2014) provides formal guarantees that an algorithm's output reveals limited information about any individual input record. We use record-level add/remove adjacency throughout. Formally, a randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-differential privacy if for any datasets $\mathcal{D}, \mathcal{D}'$ differing by the addition or removal of one record, and any measurable set $\mathcal{O}$, $\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{O}] + \delta$. The privacy budget ε controls the strength of the guarantee: smaller values provide stronger protection. The parameter δ represents the probability of a privacy failure and is typically set to be cryptographically small (e.g., $\delta < 1/|\mathcal{D}|$). We rely on four standard results. The *Gaussian mechanism* (Dwork & Roth, 2014) achieves (ε, δ)-DP by adding noise calibrated to a query's sensitivity: for $f : \mathcal{D} \to \mathbb{R}^d$ with ℓ_2-sensitivity Δ_2, releasing $f(\mathcal{D}) + \mathcal{N}(0, \sigma^2 I)$ with $\sigma \geq \Delta_2 \sqrt{2\ln(1.25/\delta)}/\varepsilon$ satisfies (ε, δ)-DP. Our implementation uses the analytic Gaussian mechanism (Balle & Wang, 2018), which provides tighter bounds on privacy loss. The *exponential mechanism* (McSherry & Talwar, 2007) selects an output r with probability proportional to $\exp(\varepsilon u(\mathcal{D}, r)/(2\Delta u))$ for utility u with sensitivity Δu, and satisfies $(\varepsilon, 0)$-DP. The *composition theorem* states that releasing the outputs of both an $(\varepsilon_1, \delta_1)$-DP mechanism and an $(\varepsilon_2, \delta_2)$-DP mechanism satisfies $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP. The *post-processing property* guarantees that any computation on the output of a DP mechanism remains DP.

Multimodal Task Vectors. Multimodal Task Vectors (MTV) (Huang et al., 2024) enable many-shot in-context learning for vision-language models (VLMs) by encoding task knowledge into attention head activations rather than token sequences. Consider a VLM with L transformer layers, each containing H attention heads of dimension d. First, MTV processes multiple batches of in-context demonstrations through the VLM. For each forward pass, it extracts the attention head activations at the final token position, yielding a tensor in $\mathbb{R}^{L \times H \times d}$. These activations are averaged across all forward passes to produce the mean activation tensor $\bar{a} \in \mathbb{R}^{L \times H \times d}$. Second, MTV uses REINFORCE (Williams, 1992) to learn a binary mask $\mathbf{m} \in \{0, 1\}^{L \times H}$ identifying which attention heads benefit from activation injection. Third, at inference time, for each selected head where $\mathbf{m}_{l,h} = 1$, the model's original activation is replaced with the corresponding component of $\bar{a}$, steering behavior toward the demonstrated task. MTV's key property is that it aggregates information from hundreds of examples into a compact representation. This enables many-shot learning that would otherwise exceed context window limits, achieving substantial accuracy gains on visual question answering and image classification benchmarks.

Privacy in In-Context Learning. In-context learning poses inherent privacy risks. When demonstrations contain sensitive data, models exhibit higher confidence on those examples, enabling membership inference attacks that determine whether specific records were used (Wen et al., 2024; Duan et al., 2024). These risks extend to multimodal settings, where images may reveal sensitive attributes beyond their apparent content (Mendes et al., 2024; Tömekçe et al., 2024; Caldarella et al., 2024). To mitigate these risks, several frameworks have introduced Differential Privacy (DP) to In-Context Learning (ICL). Early approaches focused on text-only models, utilizing methods such as private synthetic demonstration generation (Panda et al., 2023; Tang et al., 2023) or ensemble-based aggregation like PATE (Wu et al., 2023). More recent works have explored Local Differential Privacy (LDP) applied to labels (Zheng et al., 2024) or leveraging public auxiliary data to reduce privacy costs (Joo et al., 2025). These methods share two limitations. First, they operate in token space, where privacy cost scales with sequence length. A single image corresponds to hundreds of visual tokens (Bai et al., 2023), making token-level protection prohibitively expensive for multimodal ICL. Second, all existing methods are restricted to few-shot settings due to context window constraints. Our work addresses both limitations by privatizing an activation aggregate once, so privacy cost is independent of the number of tokens processed at inference time.

DP Fine-Tuning for VLMs. DP-SGD (Abadi et al., 2016) is impractical in our setting. It requires per-sample gradients through the full VLM, which are memory-intensive and not yet broadly supported for multimodal architectures (concurrent work (Wei et al., 2025) takes initial steps). Moreover, with only hundreds of private examples, noise calibrated to an 8B-parameter model can overwhelm the learning signal (Tramer & Boneh, 2020).

Table 1: Comparison of ICL-style methods. Context-free† methods encode task knowledge outside the prompt, enabling unlimited queries without consuming context window tokens. Methods marked ‡ consume privacy budget per query, bounding the total queries answerable under a fixed budget.

Method	Modality	# Examples	DP Guarantee	Unlimited Queries	Context-Free†
Tang et al. (2023)	Text	≤ 8	✓	✓	✗
Wu et al. (2023)	Text	≤ 10	✓	✗‡	✗
Duan et al. (2024)	Text	≤ 10	✓	✗‡	✗
Zheng et al. (2024)	Text	small	✓(local)	✓	✗
Joo et al. (2025)	Text	few	✓	✓	✗
Flemings et al. (2025)	Text	few	✓	✓	✗
Jiang et al. (2024)	Multimodal	$\sim$2000	✗	✓	✗
Huang et al. (2024)	Multimodal	100+	✗	✓	✓
DP-MTV (Ours)	**Multimodal**	**100+**	✓	✓	✓

Algorithm 1: Activation Similarity Attack

Require: Target sample x^*, task vector τ, activation a^*, background samples $\mathcal{B} = \{x_1, \ldots, x_m\}$, model f_θ, DP parameters (σ, C)
Ensure: 1 if x^* was a demonstration for a^*; 0 otherwise
1: $\mathcal{S}_{\text{mem}} \leftarrow \{x^*\} \cup \textsc{Sample}(\mathcal{B}, k-1)$ ▷ Sample demonstrations with x^*
2: $\tau_{\text{mem}} \leftarrow \textsc{ComputeTaskVector}(\mathcal{S}_{\text{mem}}, \sigma, C)$ ▷ Compute its task vector
3: $\mathcal{S}_{\text{non}} \leftarrow \textsc{Sample}(\mathcal{B}, k)$ where $x^* \notin \mathcal{S}_{\text{non}}$ ▷ Sample demonstrations without x^*
4: $\tau_{\text{non}} \leftarrow \textsc{ComputeTaskVector}(\mathcal{S}_{\text{non}}, \sigma, C)$ ▷ Compute its task vector
5: $s_{\text{mem}} \leftarrow \textsc{CosineSim}(a^*, \tau_{\text{mem}})$ ▷ Compute similarity between a^* and "with x^*"
6: $s_{\text{non}} \leftarrow \textsc{CosineSim}(a^*, \tau_{\text{non}})$ ▷ Compute similarity between a^* and "without x^*"
7: **return** $\mathbf{1}[s_{\text{mem}} > s_{\text{non}}]$ ▷ Guess membership

3 Membership Inference for Multimodal Task Vectors

We propose a membership inference attack (MIA) against multimodal task vectors based on activation similarity. The key insight is that when a sample x^* is included in the computation of a task vector τ, its activation a^* directly contributes to the mean, resulting in higher cosine similarity between a^* and τ compared to a task vector that excludes x^*. We formalize this as a membership inference attack (Algorithm 1). Given a target sample x^* and access to one task vector a^*, the attacker must determine whether x^* was used in the demonstrations for a^*. We compute both a *member* task vector τ_{mem} (which includes x^*) and a *non-member* task vector τ_{non} (which excludes x^*). Then, we predict that x^* was a member if the cosine similarity between a^* and τ_{mem} is greater than the similarity between a^* and τ_{non}. Note that the attacker does not directly use the task vector τ under attack. Instead, the attacker constructs reference task vectors τ_{mem} and τ_{non} using background samples, then compares the target's activation similarity to each. This reference-based approach allows membership inference without requiring access to the original task vector's construction process.

Without differential privacy, this attack achieves near-perfect accuracy: the member similarity s_{mem} is consistently higher than s_{non}, allowing the attacker to reliably distinguish membership. See Section 5 for experimental results. Our proposed DP-MTV approach is an effective defense against this attack.

4 Differentially Private Multimodal Task Vectors

We consider a vision-language model with L transformer layers, each containing H attention heads of dimension d. Given a private dataset $\mathcal{D}_{\text{priv}}$ of N examples, the goal is to enable many-shot in-context learning from $\mathcal{D}_{\text{priv}}$ while satisfying (ε, δ)-differential privacy. For VQA tasks, each private record is an image-question-answer triple (x, q, a); for classification tasks, each private record is a labeled image example. We use record-level add/remove adjacency, where neighboring datasets differ by the addition or removal of one example. We protect each example in $\mathcal{D}_{\text{priv}}$ against adversaries who observe model outputs and attempt to infer whether specific examples were used in training

Algorithm 2: Private Mean Activation Computation

Input: $\mathcal{D}_{\text{priv}}$ (private dataset), F (VLM), $\mathcal{S}$ (selected layers), K (demonstrations per chunk), C (clipping threshold), $\varepsilon_{\text{tv}}, \delta_{\text{tv}}$ (privacy parameters)

Output: $\bar{a}^{\text{priv}} \in \mathbb{R}^{|\mathcal{S}| \times H \times d}$ (private mean activations)

Partition $\mathcal{D}_{\text{priv}}$ into m disjoint chunks $\mathcal{C}_i = \{\text{target}_i, \text{demo}_{i,1}, \ldots, \text{demo}_{i,K}\}$; ▷ Each example in exactly one chunk

for $i \leftarrow 1$ **to** m **do**

  $a_i \leftarrow \text{Extract}(F, \mathcal{C}_i, \mathcal{S})$; ▷ Forward pass chunk $\mathcal{C}_i$

  for $l \in \mathcal{S}$ **do**

    $\tilde{a}_i^{(l)} \leftarrow a_i^{(l)} \cdot \min\left(1, \frac{C}{\|a_i^{(l)}\|_2}\right)$; ▷ Per-layer clipping

$\bar{a} \leftarrow \frac{1}{m} \sum_{i=1}^{m} \tilde{a}_i$; ▷ Aggregate clipped activations

$\Delta_2 \leftarrow \frac{\sqrt{|\mathcal{S}| \cdot C}}{m}, \quad \sigma^2 \leftarrow \text{AnalyticGaussian}(\Delta_2, \varepsilon_{\text{tv}}, \delta_{\text{tv}})$; ▷ Sensitivity and noise scale

$\bar{a}^{\text{priv}} \leftarrow \bar{a} + \mathcal{N}(0, \sigma^2 I)$; ▷ Add Gaussian noise

return $\bar{a}^{priv}$

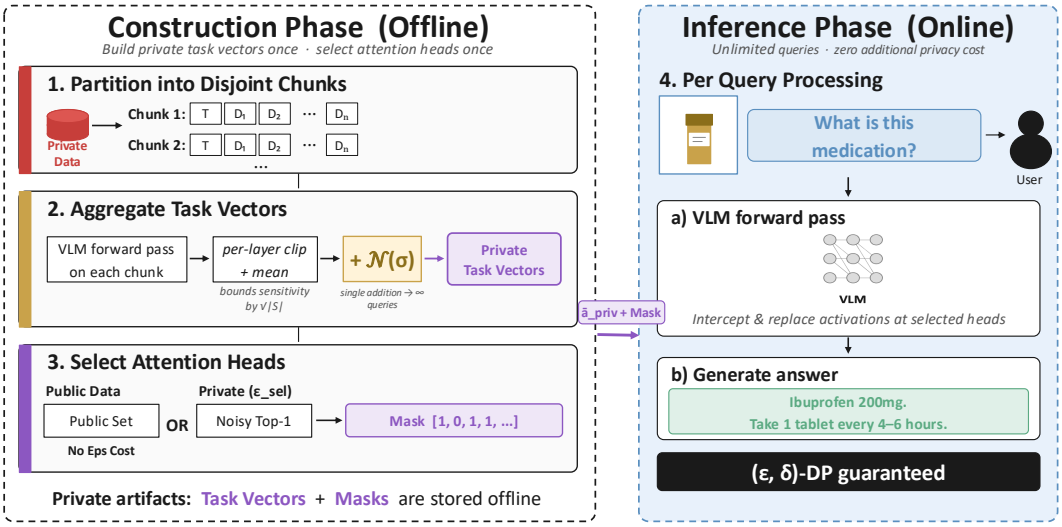


Figure 2: DP-MTV overview. **Construction (offline):** private demonstrations are partitioned into disjoint chunks; per-layer clipping bounds the sensitivity of mean activations, to which we add calibrated Gaussian noise. Attention heads are then selected via REINFORCE on public data or a private sample-and-aggregate mechanism followed by the exponential mechanism. **Inference (online):** the released task vector steers the VLM by replacing activations at the selected heads. Post-processing on the released vector allows unlimited queries at no additional privacy cost.

(membership inference) or extract sensitive attributes from them. The adversary can query the model arbitrarily but has no access to the training procedure or model internals.

Figure 2 illustrates our approach. MTV separates construction from inference, suggesting a natural privacy strategy: privatize the mean activations and head masks once during construction, then reuse them for unlimited queries without additional privacy cost. The key point is that we do not release per-example activations or per-query private computations. We release only a clipped aggregate and, optionally, a private head mask. Two technical challenges emerge: computing private mean activations with a tight sensitivity bound, and selecting heads privately.

Our method addresses both challenges through a two-phase approach (Figure 2). The **construction phase (offline)** partitions data into disjoint chunks so each example appears once, clips each chunk's activations to bounded norm, computes the mean, and adds calibrated noise to protect privacy. Head

selection either runs on public data at zero privacy cost or uses a private selection mechanism when no public data exists. The **inference phase (online)** processes queries by replacing activations at selected heads with the private mean, requiring no additional privacy cost.

Construction Step 1: Private Mean Activations. Computing private mean activations requires balancing two competing goals: tight sensitivity bounds that minimize noise, and preserving enough signal for effective learning. We achieve this balance through disjoint partitioning combined with per-layer clipping. We partition $\mathcal{D}_{\text{priv}}$ into m disjoint chunks where each example appears exactly once. Each chunk contains one target example plus K demonstration examples. For VQA tasks, demonstrations are image-question-answer triplets that provide context. For classification tasks, demonstrations are image-label pairs showing examples of the relevant classes. Each chunk undergoes a forward pass through the VLM to extract activation patterns at selected layers $\mathcal{S} \subseteq [L]$, producing activation tensor $a_i \in \mathbb{R}^{|\mathcal{S}| \times H \times d}$.

Before aggregation, we clip activations independently for each layer. For layer $l \in \mathcal{S}$, the layer-level activation $a_i^{(l)} \in \mathbb{R}^{H \times d}$ is clipped to norm at most C (line 5). After clipping, we compute the mean (line 6). Since each example appears in exactly one chunk and clipping bounds each chunk's contribution, adding or removing one example affects only one clipped tensor. The $\sqrt{|\mathcal{S}|}$ factor in the sensitivity Δ_2 (line 7) arises because clipping is applied independently per layer, and these contributions compose in ℓ_2 norm. With $|\mathcal{S}| = 32$, this yields $\Delta_2 \approx \frac{5.65C}{m}$. We add Gaussian noise calibrated to this sensitivity using the analytic Gaussian mechanism Balle & Wang (2018). We release $\bar{a}^{\text{priv}}$, which satisfies $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$-differential privacy (Theorem 4.1).

Construction Step 2: Attention Head Selection. Mean activations must be injected into task-relevant attention heads. We use REINFORCE (Williams, 1992) to learn which heads to modify by training Bernoulli distributions over attention heads and sampling binary masks that maximize task performance via policy gradients. When public examples from a related distribution exist, we run REINFORCE entirely on this public data at zero privacy cost. The public data need not match $\mathcal{D}_{\text{priv}}$ exactly–similar task structure suffices. For medical VQA, public examples might come from different hospitals or imaging modalities. For document understanding, public forms from different domains provide adequate signal. This variant concentrates the full privacy budget $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$ on privatizing mean activations. The final system satisfies $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$-DP since head selection with post-processing preserves privacy (Theorem 4.2).

When no public data exists, we introduce a differentially private counterpart to MTV's REINFORCE head selection (Algorithm 3). We privatize the REINFORCE *proposal distribution* once, making subsequent candidate-mask generation free by post-processing, and spend a small additional budget on selecting the final mask. Specifically, we run REINFORCE independently on disjoint private-data chunks and aggregate the resulting Bernoulli head-selection probabilities using the Gaussian mechanism. We then sample candidate masks from this private distribution and select among the most frequent candidates using the exponential mechanism (McSherry & Talwar, 2007) with clipped validation loss as utility. Full pseudocode and privacy accounting appear in Appendix C. The private-only variant satisfies $(\varepsilon_{\text{tv}} + \varepsilon_\theta + \varepsilon_{\text{mask}}, \delta_{\text{tv}} + \delta_\theta)$-DP by composition (Theorem 4.3).

Inference. During inference, the model processes queries using only the private mean activations $\bar{a}^{\text{priv}}$ and head mask $\mathbf{m}$ computed during construction. For each query, the model performs a standard forward pass. At each selected attention head where $\mathbf{m}_{l,h} = 1$, we intercept the computed activation and replace it with the corresponding component of $\bar{a}^{\text{priv}}$ before continuing the forward pass. The model then generates an answer using these modified activations. This inference procedure satisfies DP by the post-processing property, enabling the model to serve unlimited queries without accumulating additional privacy cost.

Privacy Analysis. We prove DP for both variants. Full proofs appear in Appendix B.

Theorem 4.1 (Private Mean Activations)**.** *Under record-level add/remove adjacency, data-independent chunk assignment ensures each record affects exactly one chunk, and per-layer clipping bounds each chunk's contribution. Algorithm 2 satisfies $(\varepsilon_{tv}, \delta_{tv})$-differential privacy with ℓ_2-sensitivity $\Delta_2 = \sqrt{|\mathcal{S}|} \cdot C/m$.*

Theorem 4.2 (Public-Data Variant)**.** *When head selection runs entirely on public data $\mathcal{D}_{pub}$, the complete system releasing $(\bar{a}^{priv}, \mathbf{m})$ satisfies $(\varepsilon_{tv}, \delta_{tv})$-differential privacy with respect to $\mathcal{D}_{priv}$.*

Head selection on public data incurs zero privacy cost. Inference applies a deterministic function to the released artifacts; by the post-processing property, this preserves the $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$ guarantee for unlimited queries.

Theorem 4.3 (Private-Only Variant)**.** *When head selection releases θ^{priv} using sample-and-aggregate and then uses the exponential mechanism over sampled candidate masks with loss clipping threshold C_{sel}, the complete system satisfies $(\varepsilon_{tv} + \varepsilon_{\theta} + \varepsilon_{mask}, \delta_{tv} + \delta_{\theta})$-differential privacy.*

Mean activations satisfy $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$-DP by Theorem 4.1. Releasing θ^{priv} satisfies $(\varepsilon_{\theta}, \delta_{\theta})$-DP, candidate sampling is post-processing, and the exponential mechanism provides $(\varepsilon_{\text{mask}}, 0)$-DP for the final selection. Sequential composition yields the total guarantee, and inference is DP by post-processing.

5 Experiments & Results

Datasets. We evaluate on eight vision-language benchmarks (see Table 4 in Appendix D.1). For VQA, we use VizWiz (Gurari et al., 2018) (visual questions from blind users), VQA-RAD (Lau et al., 2018) and PathVQA (He et al., 2020) (medical VQA on radiology and pathology images), OK-VQA (Marino et al., 2019) (knowledge-intensive questions), and TextVQA (Singh et al., 2019) (questions requiring reading text in images). For classification, we use Flowers102 (Nilsback & Zisserman, 2008), CUB-200 (Wah et al., 2011), and DTD (Cimpoi et al., 2014) for fine-grained visual recognition. Following MTV, classification tasks are formulated as 2-way 1-shot problems: given one example each of classes A and B, the model predicts the query image's class.

Models & Hyperparameters. We evaluate three vision-language models with different architectural foundations. **Qwen-VL** (Bai et al., 2023) is a LLaMA-based VLM pretrained on interleaved image-text data. **LLaMA3-ViLA-1.5-8B** (Lin et al., 2024) (ViLA-1.5) uses LLaMA-3 as its language backbone. **Idefics2-8B** (Laurençon et al., 2024) is a Mistral-based model pretrained on web-scraped multimodal documents. All three models are designed for interleaved multimodal ICL, making them suitable testbeds for evaluating DP-MTV. We extract activations from all 32 layers, partition each training set into $m = 100$ disjoint chunks with clipping threshold $C = 1.0$, and evaluate $\varepsilon \in \{0.1, 0.5, 1, 2, 5\}$. Each chunk contains $K = 8$ demonstrations for VQA (900 examples per dataset) or $K = 2$ for classification (100 examples), representing 2–50% of each training set.

Baselines & Evaluation. We compare against three baselines. **0-Shot** performs zero-shot inference without in-context examples. **Clean MTV** applies standard MTV without privacy constraints, establishing an upper bound. The gap between these baselines indicates the headroom available for DP-MTV. We focus our evaluation on the variant of DP-MTV that uses auxiliary public data for head selection, since it provides the most practical path to deployment. We measure accuracy using the VQA evaluation metric (Antol et al., 2015) for VQA tasks and exact match for classification. All results are averaged over 3 random seeds. Privacy guarantees are reported as (ε, δ)-differential privacy with $\delta = 10^{-5}$.

5.1 Results

We evaluate DP-MTV on five VQA benchmarks and three classification datasets using three VLM architectures. Our experiments address three questions: *(1) Does DP-MTV enable effective private in-context learning? (2) When does DP-MTV work well? (3) Does DP-MTV generalize across model architectures?* Table 2 presents our utility results; Figure 4 in Appendix E shows privacy-utility tradeoffs across the full range of privacy budgets.

DP-MTV preserves in-context learning gains under privacy. At ε=2, DP-MTV matches or exceeds non-private MTV on the majority of dataset-model combinations (**bold** entries in Table 2). On VizWiz, where MTV provides the largest gain over zero-shot (+20% for Qwen-VL), DP-MTV achieves 50.4% at ε=1, retaining 92% of MTV's 54.6%. Performance improves as the privacy budget increases (Figure 4a), and at $\varepsilon \geq 2$ the privacy cost is negligible across most tasks.

The baseline gap predicts when DP-MTV succeeds. Across all architectures, DP-MTV effectiveness depends on the *baseline gap* (MTV − Zero-shot). When the gap is large, task vectors encode useful signal that survives noise addition. VizWiz exemplifies this: Qwen-VL's +20% gap provides headroom for DP-MTV to recover substantial gains. When the gap is small or negative, as with TextVQA on Qwen-VL (-4%), there is no signal to preserve, and DP-MTV performs comparably to

Table 2: DP-MTV preserves in-context learning gains under formal differential privacy across three VLM architectures and eight benchmarks. At ϵ=2, DP-MTV matches or exceeds the non-private baseline on the majority of tasks (**bold**), demonstrating that meaningful privacy can be achieved at negligible utility cost. Results are consistent across architectures, confirming that the approach is model-agnostic. Full results with standard deviations across seeds appear in Table 6 (Appendix E.3).

Model	Method	VizWiz	OK-VQA	TextVQA	VQA-RAD	PathVQA	Flowers	CUB	DTD
Qwen-VL	Zero-shot	35	58	68	38	25	54	57	54
	MTV (no DP)	55	60	64	40	28	77	78	76
	DP-MTV ϵ=5	51	56	65	41	29	**97**	**93**	**93**
	DP-MTV ϵ=2	50	57	65	42	30	**96**	**93**	**93**
	DP-MTV ϵ=1	50	56	**64**	**43**	**30**	**96**	**93**	**93**
ViLA	Zero-shot	55	39	49	39	29	74	70	82
	MTV (no DP)	61	51	53	43	31	84	81	84
	DP-MTV ϵ=5	**62**	48	54	42	31	**90**	**92**	**90**
	DP-MTV ϵ=2	**62**	48	54	42	31	**90**	**92**	**90**
	DP-MTV ϵ=1	**61**	40	52	19	30	50	**89**	**92**
idefics2	Zero-shot	39	51	69	52	32	85	88	90
	MTV (no DP)	57	52	65	52	32	87	89	94
	DP-MTV ϵ=5	**59**	**52**	**66**	**52**	**32**	**87**	**90**	**95**
	DP-MTV ϵ=2	**59**	**52**	**66**	**52**	**32**	**87**	**90**	**95**
	DP-MTV ϵ=1	**59**	47	57	0	0	**87**	**90**	**94**

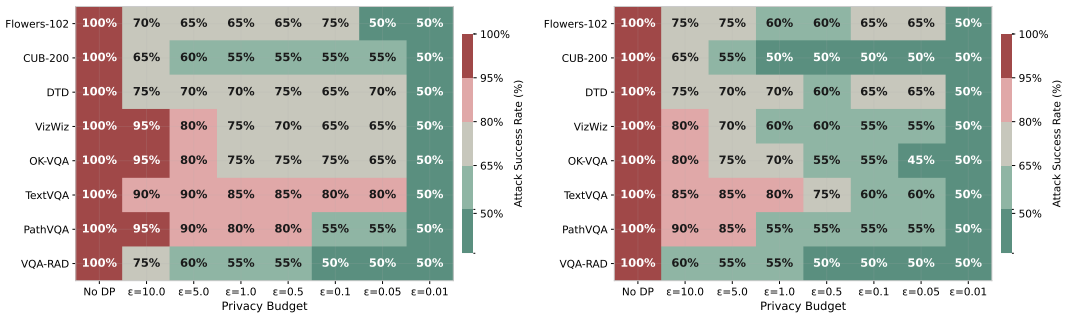
Table 3: Fully-private DP-MTV (Algorithm 3). Head selection requires *no* auxiliary public data and allocates the total privacy budget across ε_{tv}, ε_θ, and $\varepsilon_{\text{mask}}$ using a 60/30/10 split. Zero-shot and non-private MTV are repeated for reference. **Bold** entries match or exceed non-private MTV.

Model	Method	VizWiz	OK-VQA	TextVQA	VQA-RAD	PathVQA	Flowers	CUB	DTD
Qwen-VL	Zero-shot	35	58	68	38	25	54	57	54
	MTV (no DP)	55	60	64	40	28	77	78	76
	DP-MTV (Private) ϵ=5	49	55	58	**42**	**28**	**96**	**90**	**76**
	DP-MTV (Private) ϵ=2	11	24	13	26	18	45	40	49
	DP-MTV (Private) ϵ=1	0	0	0	0	0	0	0	0
ViLA	Zero-shot	55	39	49	39	29	74	70	82
	MTV (no DP)	61	51	53	43	31	84	81	84
	DP-MTV (Private) ϵ=5	57	25	33	39	**31**	82	78	**91**
	DP-MTV (Private) ϵ=2	0	0	2	2	2	0	0	0
	DP-MTV (Private) ϵ=1	0	0	0	0	0	0	0	0
idefics2	Zero-shot	39	51	69	52	32	85	88	90
	MTV (no DP)	57	52	65	52	32	87	89	94
	DP-MTV (Private) ϵ=5	6	21	12	46	30	68	49	82
	DP-MTV (Private) ϵ=2	2	3	1	0	1	0	0	0
	DP-MTV (Private) ϵ=1	0	0	0	0	0	4	12	1

both MTV and zero-shot. This also explains the zero-valued entries in Table 2. For instance, idefics2 achieves identical MTV and zero-shot accuracy on VQA-RAD (52%) and PathVQA (32%), so the task vector carries no signal and noise can only hurt. At $\varepsilon \geq 2$, these entries recover to baseline. This gives practitioners a simple diagnostic: estimate the baseline gap on public data before committing to private deployment.

DP mechanisms can improve performance. On all three classification benchmarks, DP-MTV consistently *exceeds* non-private MTV, sometimes substantially (Qwen-VL Flowers: 96% vs. 77%; ViLA CUB: 92% vs. 81%; idefics2 DTD: 95% vs. 94%). This holds across all architectures and privacy budgets (Figure 4b), ruling out random variation. We attribute this to per-layer clipping removing outlier activations that dominate transformer representations (Chen et al., 2020), combined with Gaussian noise acting as regularization (Bishop, 1995).

The fully-private variant removes the public-data assumption at a utility cost. Table 3 evaluates fully-private DP-MTV (Algorithm 3), which spends part of the privacy budget on head selection rather than using auxiliary public data. Utility degrades sharply at $\varepsilon \leq 2$, but recovers substantially at ε=5, matching or exceeding non-private MTV on several tasks (e.g., Qwen-VL Flowers: **96**). The degradation stems from the higher-dimensional private mask (p=$|\mathcal{S}|H$; Theorem 4.3); selecting



(a) 5 demonstrations per task vector

(b) 100 demonstrations per task vector

Figure 3: **Empirical privacy validation.** Membership inference attack success rate (Algorithm 1); 50% is random guessing. Without DP, the attack succeeds on every dataset. With DP-MTV, attack success drops to 50% as ε decreases. Many-shot strengthens privacy as task vectors built from 100 demonstrations (b) reach random-guess at less restrictive budgets than those built from 5 demonstrations (a).

whole layers instead substantially improves performance at ε=2 (Appendix E.5). Thus, the fully-private variant provides an alternative when no auxiliary public data is available, with layer-level selection narrowing the utility gap at tighter privacy budgets.

DP-MTV generalizes across architectures. Table 2 spans three VLMs with distinct language backbones (LLaMA for Qwen-VL, LLaMA-3 for ViLA, and Mistral for idefics2), and all show the same behavior: classification preserved or improved, VQA results following the baseline gap pattern. ViLA shows sharper sensitivity at ε=1 on some tasks (VQA-RAD: 19%) but recovers fully at ε=2 (42%), suggesting moderate budgets suffice for stable deployment.

Empirical privacy validation. Formal DP guarantees bound worst-case leakage but do not reveal whether concrete attacks succeed. We design an activation-space membership inference attack (Algorithm 1) targeting DP-MTV directly: the attacker receives a task vector that either includes or excludes a target sample and must distinguish the two cases; random guessing achieves 50%. Figure 3 evaluates whether DP-MTV successfully defends against this attack. We run the attack against non-private MTV and against DP-MTV at several privacy budgets. Without DP, the attack succeeds on every dataset in both the 5- and 100-demonstration settings. DP-MTV significantly reduces the attack's success rate. At moderate privacy budgets, attack success already drops well below the non-private baseline, and as ε decreases, many datasets approach the 50% random-guess rate. These empirical results are consistent with the formal DP guarantees and suggest that DP-MTV successfully defends against strong privacy attacks like membership inference.

Hyperparameters & Efficiency. Figure 5 (Appendix E.4) shows accuracy as a function of clipping threshold C on VizWiz (Qwen-VL, ε=1). C=1.0 provides a favorable tradeoff; performance degrades gradually at higher thresholds. We also vary the number of chunks $m \in \{50, 100, 500, 1000\}$ and demonstrations $K \in \{2, 8, 16\}$; performance remains stable across both (Table 7, Appendix E.4). Once constructed, the private steering vector adds no inference overhead: per-query latency matches zero-shot (180 ms vs. 1071 ms for K=8 few-shot ICL). Full timing details appear in Appendix E.2.

6 Conclusion

We introduced DP-MTV, the first framework for differentially private many-shot multimodal in-context learning. By aggregating hundreds of private demonstrations into a single steering vector and privatizing the aggregate, DP-MTV pays all privacy cost once at construction and answers unlimited queries at zero additional cost. Across eight benchmarks and three VLM architectures, DP-MTV matches or exceeds non-private performance on the majority of tasks at ε=2, and on classification consistently outperforms the non-private baseline. A novel membership inference attack validates these guarantees, dropping from 100% success to the 50% random-guess baseline under DP-MTV. Future directions include adaptive clipping strategies, fully private head selection with tighter composition, and extensions to other activation editing methods.

7 Ethics Statement

This work aims to enable vision-language models to learn from sensitive multimodal data while providing formal differential privacy guarantees. Our evaluation includes medical VQA datasets (VQA-RAD and PathVQA) and VizWiz, which contains images and questions from blind and low-vision users. Differential privacy protects the contribution of individual examples but does not prevent population-level inference or mitigate biases inherited from the underlying VLM.

8 AI Assistance Disclosure

Generative AI tools were used to assist with research execution, literature retrieval and discovery, and editing the manuscript for clarity and presentation. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the methods, experiments, results, claims, and final content of the paper.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International conference on machine learning*, pp. 394–403. PMLR, 2018.

Chris M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Simone Caldarella, Massimiliano Mancini, Elisa Ricci, and Rahaf Aljundi. The phantom menace: unmasking privacy leakages in vision-language models. In *European Conference on Computer Vision*, pp. 435–451. Springer, 2024.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13773–13782, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the privacy risk of in-context learning. *arXiv preprint arXiv:2411.10512*, 2024.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

James Flemings, Haosheng Gan, Hongyi Li, Meisam Razaviyayn, and Murali Annavaram. Differentially private in-context learning via sampling few-shot mixed with zero-shot outputs. *arXiv preprint arXiv:2501.19287*, 2025.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Dai Hou, Zhenkai Yang, Lei Zheng, Bo Jin, Huan Xu, Ying Li, Bo Xu, and Kai Peng. Neighborhood deviation attack against in-context learning. *Applied Sciences*, 15(8):4177, 2025.

Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. *Advances in Neural Information Processing Systems*, 37:22124–22153, 2024.

Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024.

Seongho Joo, Hyukhun Koh, and Kyomin Jung. Public data assisted differentially private in-context learning. *arXiv preprint arXiv:2509.10932*, 2025.

Younggun Kim, Sirnam Swetha, Fazil Kagdi, and Mubarak Shah. Safe-llava: A privacy-preserving vision-language dataset and benchmark for biometric safety. *arXiv preprint arXiv:2509.00192*, 2025.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199, 2019. doi: 10.1109/CVPR.2019.00331.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007)*, pp. 94–103. IEEE, 2007.

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Granular privacy control for geolocation with vision language models. *arXiv preprint arXiv:2407.04952*, 2024.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Ashwinee Panda, Tong Wu, Jiachen Wang, and Prateek Mittal. Differentially private in-context learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. *arXiv preprint arXiv:2309.11765*, 2023.

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models. *Advances in Neural Information Processing Systems*, 37: 103619–103651, 2024.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Qianshan Wei, Jiaqi Li, Zihan You, Yi Zhan, Kecen Li, Jialin Wu, Xinfeng Li Hengjun Liu, Yi Yu, Bin Cao, Yiwen Xu, Yang Liu, and Guilin Qi. Dual-priv pruning : Efficient differential private fine-tuning in multimodal large language models, 2025. URL https://arxiv.org/abs/2506.07077.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against in-context learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3481–3495, 2024.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Tong Wu, Ashwinee Panda, Jiachen T Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. *arXiv preprint arXiv:2305.01639*, 2023.

Chunyan Zheng, Keke Sun, Wenhao Zhao, Haibo Zhou, Lixin Jiang, Shaoyang Song, and Chunlai Zhou. Locally differentially private in-context learning. *arXiv preprint arXiv:2405.04032*, 2024.

A LIMITATIONS

Our work has several limitations. First, DP-MTV's effectiveness depends on the baseline gap between MTV and zero-shot performance. When MTV provides little improvement over zero-shot (e.g., TextVQA on Qwen-VL), there is no signal to preserve, and DP-MTV cannot outperform the zero-shot baseline. Practitioners should estimate this gap on public data before committing to private deployment.

Second, the fully-private variant (Section 9, Algorithm 3) spends additional privacy budget on head selection ($\varepsilon_\theta + \varepsilon_{\text{mask}}$), reducing the budget available for mean-activation privatization relative to the public-data variant. As Table 3 shows, this cost is steep at tight budgets ($\varepsilon \leq 2$) and only becomes competitive with the public variant at $\varepsilon = 5$. Section E.5 shows that selecting whole layers rather than individual heads recovers much of this gap at $\varepsilon = 2$. When public auxiliary data is available, the public-data variant reaches higher accuracy for the same budget; the fully-private variant trades some of that accuracy to remove the public-data assumption entirely.

Third, the clipping threshold C requires tuning. While $C = 1.0$ works well across our benchmarks (Figure 5), optimal values may vary for different model architectures or task domains. Adaptive clipping strategies could address this but remain unexplored.

Fourth, our analysis uses basic composition; tighter bounds via Rényi differential privacy (Mironov, 2017) or the moments accountant (Abadi et al., 2016) could further improve the privacy-utility tradeoff, particularly for the fully-private variant.

Finally, our MIA evaluation assumes the attacker has access to the same model and can compute activations for arbitrary samples. This white-box assumption may overestimate attack success in deployment scenarios with restricted API access.

B PRIVACY ANALYSIS PROOFS

Proof of Theorem 4.1. We analyze privacy under record-level add/remove adjacency, where neighboring datasets differ by adding or removing one example.

Sensitivity bound. Each example appears in exactly one chunk. Per-layer clipping ensures $\|\tilde{a}_i^{(l)}\|_2 \leq C$ for each layer $l \in \mathcal{S}$. The full clipped tensor thus satisfies:

$$\|\tilde{a}_i\|_2 = \sqrt{\sum_{l \in \mathcal{S}} \|\tilde{a}_i^{(l)}\|_2^2} \leq \sqrt{|\mathcal{S}| \cdot C^2} = \sqrt{|\mathcal{S}|} \cdot C$$

Adding or removing one example affects one chunk's contribution to the mean $\bar{a} = \frac{1}{m} \sum_i \tilde{a}_i$, yielding the ℓ_2-sensitivity $\Delta_2 = \sqrt{|\mathcal{S}|} \cdot C/m$.

Privacy guarantee. The Gaussian mechanism with $\sigma = \Delta_2 \sqrt{2 \ln(1.25/\delta_{\text{tv}})}/\varepsilon_{\text{tv}}$ satisfies $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$-DP by standard results (Dwork & Roth, 2014). Our implementation uses the analytic Gaussian mechanism with the same sensitivity. □

Proof of Theorem 4.2. Head selection operates entirely on $\mathcal{D}_{\text{pub}}$, which is independent of $\mathcal{D}_{\text{priv}}$, contributing $(0, 0)$ to privacy cost. The released artifacts $(\bar{a}^{\text{priv}}, \mathbf{m})$ thus inherit the $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$ guarantee from Theorem 4.1. Inference applies deterministic functions to these artifacts; by post-processing, this preserves $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$-DP for unlimited queries. □

Proof of Theorem 4.3. Mean activations satisfy $(\varepsilon_{\text{tv}}, \delta_{\text{tv}})$-DP by Theorem 4.1.

DP proposal distribution. Condition on the previously released $\bar{a}^{\text{priv}}$. Let $p = |\mathcal{S}|H$. For any fixed REINFORCE randomness, adding or removing one record changes at most one chunk-wise probability vector $\theta_r \in [0, 1]^p$. Since the ℓ_2 diameter of $[0, 1]^p$ is $\sqrt{p}$, the average $\bar{\theta} = \frac{1}{c} \sum_{r=1}^{c} \theta_r$ has sensitivity $\Delta_\theta = \sqrt{p}/c$. The Gaussian mechanism therefore releases θ^{priv} with $(\varepsilon_\theta, \delta_\theta)$-DP. Projection to $[0, 1]^p$ and sampling masks from θ^{priv} are post-processing.

DP final mask. Condition on $\bar{a}^{\text{priv}}$, θ^{priv}, and the sampled candidate masks. The candidate set is then fixed. For a candidate mask $\mathbf{m}$, define utility

$$u(\mathcal{D}, \mathbf{m}) = -\sum_{(x,q,a)\in\mathcal{D}_{\text{val}}} \min(\text{CrossEntropy}(F_{\mathbf{m}}(x,q),a), C_{\text{sel}}).$$

For fixed validation-sampling randomness, adding or removing one validation record changes at most one clipped loss term in $[0, C_{\text{sel}}]$, so the utility sensitivity is C_{sel}. The exponential mechanism with probabilities proportional to $\exp(\varepsilon_{\text{mask}} u/(2C_{\text{sel}}))$ satisfies $(\varepsilon_{\text{mask}}, 0)$-DP (McSherry & Talwar, 2007).

By adaptive sequential composition, releasing $\bar{a}^{\text{priv}}$, θ^{priv}, and $\mathbf{m}$ satisfies $(\varepsilon_{\text{tv}}+\varepsilon_\theta+\varepsilon_{\text{mask}}, \delta_{\text{tv}}+\delta_\theta)$-DP. Inference is post-processing. □

C Private Head Selection

Like the non-private head-selection approach in MTV, our private-only head-selection algorithm separates proposal generation from final validation-based selection. The algorithm appears in Algorithm 3. First, sample-and-aggregate privatizes the REINFORCE proposal distribution itself: each chunk produces a probability vector over heads, the vectors are averaged, and Gaussian noise is added before any masks are sampled. The sampled candidate set can therefore depend on private data only through θ^{priv}, making it post-processing. Second, once this candidate set is fixed, we select among candidates with the exponential mechanism using clipped validation loss. Sequential composition gives the total privacy cost for the private-only head-selection pipeline.

D Experimental Details

D.1 Dataset Statistics

Table 4 provides complete statistics for all datasets used in our experiments.

Table 4: Dataset Statistics. All datasets use standard train/test splits from original papers.

Dataset	Domain	Train	Test	Task
VizWiz	General VQA	20,523	8,000	Open-ended
VQA-RAD	Medical VQA	1,793	451	Open-ended
PathVQA	Medical VQA	19,654	6,719	Open-ended
OK-VQA	Knowledge VQA	9,009	5,046	Open-ended
TextVQA	Text VQA	34,602	5,000	Open-ended
Flowers102	Classification	1,020	6,149	2-way 1-shot
CUB-200	Classification	5,994	5,794	2-way 1-shot
DTD	Classification	3,760	1,880	2-way 1-shot

With $m = 100$ chunks and $K = 8$ demonstrations per chunk (VQA) or $K = 2$ (classification), we use approximately 500-900 training examples per dataset for private activation computation, corresponding to 2-20% of each training set depending on dataset size.

D.2 Implementation Details

Computational Resources. All experiments run on NVIDIA V100, A100 and H100 GPUs with 30GB and 80GB memory. Mean activation computation takes 1-2 hours per dataset depending on size. REINFORCE training for head selection takes 2-4 hours for the public variant.

Public Data for Head Selection. For the public-data variant, we use validation splits from related datasets when available. For VizWiz, we use VQA v2 validation data; for medical VQA tasks (VQA-RAD, PathVQA), we use samples from each other as public data since they share similar task structure despite different imaging modalities.

Algorithm 3: Private Head Selection

Input: $\mathcal{D}_{\text{priv}}$ (private dataset), F (VLM), $\bar{a}^{\text{priv}}$ (private mean activations), $\mathcal{S}$ (selected layers), c (REINFORCE chunks), T (mask samples), $\bar{k}$ (candidates), $\varepsilon_\theta, \delta_\theta, \varepsilon_{\text{mask}}$ (privacy parameters), C_{sel} (loss clipping threshold)

Output: $\mathbf{m} \in \{0,1\}^{|\mathcal{S}| \times H}$ (binary mask indicating which heads to modify)

; ▷ Phase 1: Learn chunk-wise Bernoulli probabilities

$p \leftarrow |\mathcal{S}| \cdot H$;

Partition $\mathcal{D}_{\text{priv}}$ into disjoint chunks $\mathcal{D}_1, \ldots, \mathcal{D}_c$;

for $r \leftarrow 1$ **to** c **do**

  $\eta_r \leftarrow \text{REINFORCE}(\mathcal{D}_r, F, \bar{a}^{\text{priv}})$; ▷ Train policy on one chunk

  $\theta_r \leftarrow \sigma(\eta_r)$; ▷ $\theta_r \in [0,1]^p$

; ▷ Phase 2: Release DP aggregate probabilities

$\bar{\theta} \leftarrow \frac{1}{c} \sum_{r=1}^{c} \theta_r$;

$\Delta_\theta \leftarrow \frac{\sqrt{p}}{c}, \quad \sigma_\theta^2 \leftarrow \text{AnalyticGaussian}(\Delta_\theta, \varepsilon_\theta, \delta_\theta)$;

$\theta^{\text{priv}} \leftarrow \Pi_{[0,1]^p}\left(\bar{\theta} + \mathcal{N}(0, \sigma_\theta^2 I_p)\right)$; ▷ Project to valid probabilities

; ▷ Phase 3: Sample candidate masks from DP probabilities

$\mathcal{M} \leftarrow \emptyset$; ▷ Store sampled masks

for $j \leftarrow 1$ **to** T **do**

  $\lambda_j \sim \text{Bernoulli}(\theta^{\text{priv}})$; ▷ Independent Bernoulli draw per head

  $\mathcal{M} \leftarrow \mathcal{M} \cup \{\lambda_j\}$

$\{\mathbf{m}_1, \ldots, \mathbf{m}_{\bar{k}}\} \leftarrow \text{Top-}\bar{k}\text{-FrequentUnique}(\mathcal{M})$; ▷ Retain most frequent unique masks

; ▷ Phase 4: Compute aggregate losses with clipping

$\mathcal{D}_{\text{val}} \leftarrow$ Sample $B = 100$ examples from $\mathcal{D}_{\text{priv}}$ using public randomness ; ▷ Validation set

for $i \leftarrow 1$ **to** $\bar{k}$ **do**

  $s_i \leftarrow 0$;

  for $(x, q, a) \in \mathcal{D}_{val}$ **do**

    Replace activations in F at heads indicated by $\mathbf{m}_i$ with $\bar{a}^{\text{priv}}$;

    $\ell \leftarrow \text{CrossEntropy}(F(x, q), a)$; ▷ Per-example loss

    $s_i \leftarrow s_i + \min(\ell, C_{\text{sel}})$; ▷ Clip and accumulate

; ▷ Phase 5: Apply exponential mechanism for private selection

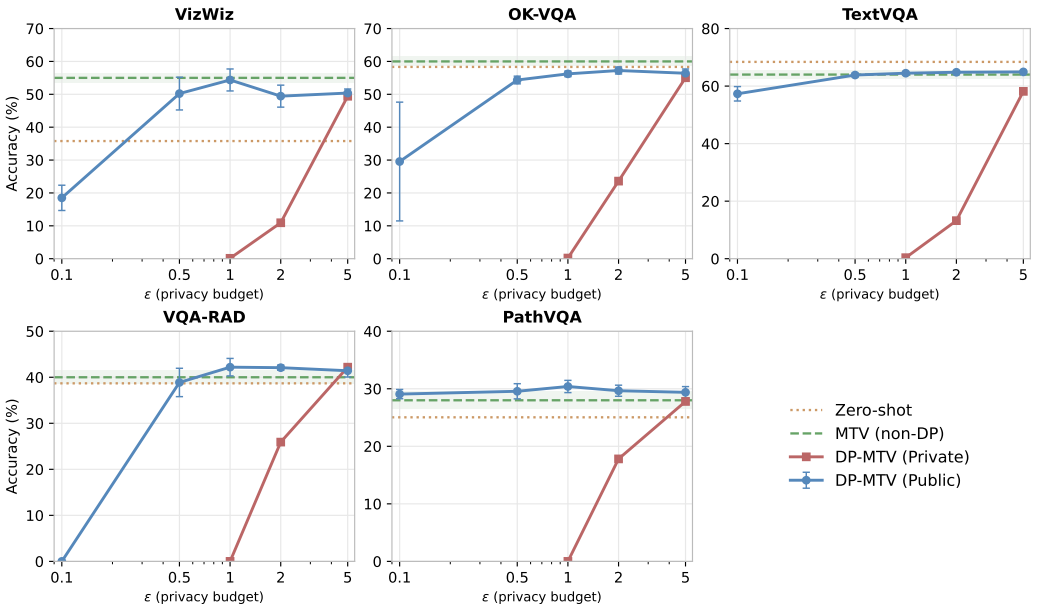
$i^* \sim \Pr[i] \propto \exp\left(-\frac{\varepsilon_{\text{mask}} s_i}{2C_{\text{sel}}}\right)$; ▷ Select mask with lowest noisy loss

return $\mathbf{m}_{i^*}$

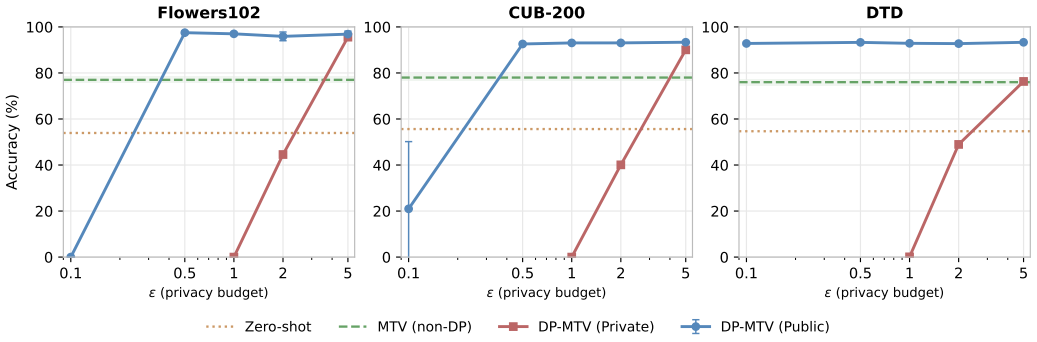
E Additional Results

E.1 Privacy-Utility Curves

Figure 4 shows the full privacy-utility tradeoff for Qwen-VL across all privacy budgets. Table 2 in the main text reports results at specific epsilon values; these curves show the complete picture.



(a) VQA Tasks



(b) Classification Tasks

Figure 4: Privacy-utility tradeoffs for Qwen-VL across privacy budgets. Dashed lines: non-private MTV; dotted lines: zero-shot. (a) On VQA, DP-MTV performs best on datasets where MTV provides the largest gains over zero-shot (e.g., VizWiz); performance varies by architecture (Table 2). (b) On classification, DP-MTV matches or exceeds MTV at practical privacy budgets.

E.2 Efficiency Analysis

Table 5 compares deployment costs. DP-MTV's per-query inference is identical to zero-shot (180 ms) because the task vector is precomputed. Few-shot ICL with $K{=}8$ is 5.96× slower per query. DP-MTV's one-time cost consists of task vector construction (<90 s) plus REINFORCE head selection (∼79 min).

Table 5: Deployment cost (Qwen-VL, VizWiz, single GPU). DP-MTV matches zero-shot latency while providing (ε, δ)-DP.

Method	One-time	Per-query	Weights	Privacy
Zero-shot	—	180 ms	unchanged	None
Few-shot ($K{=}8$)	—	1071 ms	unchanged	None
DP-MTV	81 min	180 ms	unchanged	(ε, δ)-DP

E.3 Full Results with Error Bars

Table 6 reports the complete results from Table 2 with means and standard deviations across three random seeds. Most cells exhibit low variance (std $< 2\%$), indicating stable performance across data permutations. Two notable exceptions occur with ViLA at ε=1: Flowers (49.9 ± 45.9) and VQA-RAD (18.9 ± 16.8) show bimodal behavior where some seeds recover near-full accuracy while others collapse, indicating that this model operates at the noise threshold for these tasks. This instability resolves entirely at $\varepsilon \geq 2$. The two zero-valued cells for idefics2 (VQA-RAD and PathVQA at ε=1) are deterministic: every seed collapses identically, consistent with the zero baseline gap discussed in the main text. Cells without standard deviations have $n < 3$ seeds available.

Table 6: Full results with error bars corresponding to Table 2. Entries are mean $\pm$ std across seeds (typically n=3). **Bold** indicates DP-MTV matches or exceeds non-private MTV.

Model	Method	VizWiz	OK-VQA	TextVQA	VQA-RAD	PathVQA	Flowers	CUB	DTD
Qwen-VL	Zero-shot	35.4±0.1	58.4±0.1	68.4±0.2	38.2±2.1	25.1±0.1	53.8±0.2	56.9±0.6	54.4±0.2
	MTV (no DP)	54.6±1.4	60.4±0.9	63.9±1.5	40.0±2.5	28.2±0.7	76.8±1.3	77.6±1.6	76.3±2.1
	DP-MTV ϵ=5	50.4±1.5	56.4±1.6	64.9±0.1	41.4±1.7	29.4±1.2	96.9±1.6	93.4	93.3±0.2
	DP-MTV ϵ=2	49.4±4.1	57.2±1.3	64.8±0.4	42.1±0.7	29.7±1.2	95.9±2.3	93.1±0.7	92.7±1.0
	DP-MTV ϵ=1	50.4±5.4	56.2±1.0	64.1±0.8	42.9±2.3	29.9±1.2	96.2±1.0	93.1±0.6	92.9±0.1
ViLA	Zero-shot	55.3±0.2	38.8±0.2	48.7±0.0	39.3±1.2	29.4±0.1	74.2±0.4	70.4±1.1	82.3±0.3
	MTV (no DP)	61.2±1.8	51.1±0.4	53.0±1.6	42.7±1.3	30.6±1.5	83.8±1.4	80.8±1.8	83.9±0.7
	DP-MTV ϵ=5	61.9±0.5	47.8±1.0	54.0±1.3	41.6±1.6	31.2±0.8	89.9±0.3	91.8±0.6	90.1±0.3
	DP-MTV ϵ=2	60.8±1.6	46.6±0.5	54.5±0.2	39.2±0.3	31.2±1.0	90.4±1.5	92.1±1.2	89.8±1.2
	DP-MTV ϵ=1	61.3±1.0	40.1±6.2	51.5±2.7	18.9±16.8	30.1±1.5	49.9±45.9	88.8±1.0	91.9±1.1
idefics2	Zero-shot	39.0±0.1	51.0±0.2	68.7±0.1	52.2±0.8	31.7±0.1	84.6±0.1	88.0	89.7
	MTV (no DP)	56.5±0.9	52.1±0.5	65.0±1.6	52.2±1.2	32.4±0.6	86.7±0.4	88.7	94.1
	DP-MTV ϵ=5	59.4±1.4	51.9±0.5	66.0±0.8	51.9±0.4	32.0±0.4	87.0±0.5	90.1±0.5	94.7±0.5
	DP-MTV ϵ=2	57.4±2.9	50.5±1.0	65.7±0.7	49.2±1.4	31.6±0.5	87.4±0.3	90.0±0.1	94.7±0.6
	DP-MTV ϵ=1	59.0±2.5	47.3±0.6	56.9±1.4	0.0±0.0	0.0±0.0	87.0±0.4	89.9±0.6	94.5±0.2

E.4 Robustness to Hyperparameters

We evaluate sensitivity to key hyperparameters on VizWiz (Qwen-VL, ε=1). Figure 5 shows the effect of clipping threshold C; C=1.0 provides a favorable tradeoff. Table 7 varies the number of chunks m and demonstrations K; performance is stable across both.

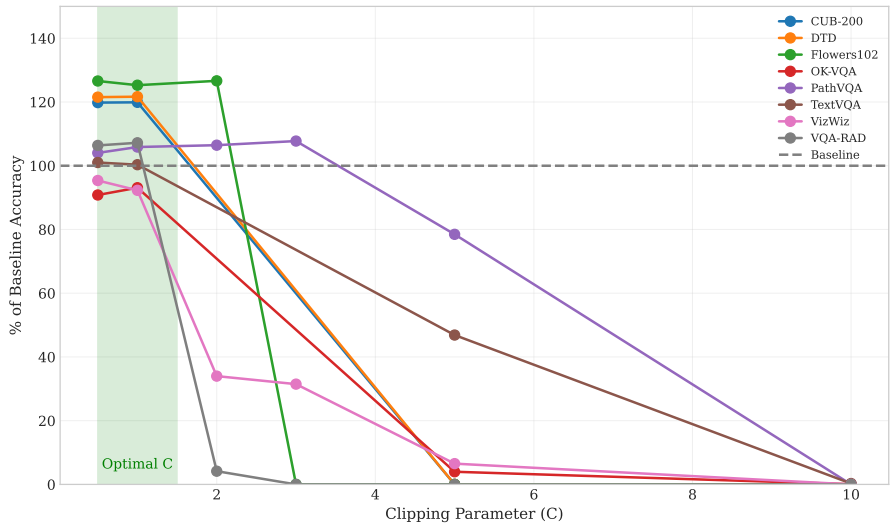


Figure 5: Effect of clipping threshold C on VizWiz accuracy (Qwen-VL, $\varepsilon = 1.0$). $C = 1.0$ balances signal preservation with noise calibration. Performance degrades gradually at higher thresholds.

Table 7: Sensitivity to hyperparameters on VizWiz (Qwen-VL, $\varepsilon = 1.0$, $C = 1.0$). Performance is stable across both number of chunks m and number of shots K, confirming the method is not brittle to these choices.

	Number of Chunks m			
	50	100	500	1000
DP-MTV	49.8±8.6	49.1±10.3	50.6±3.9	48.7±5.9

	Number of Shots K		
	2	8	16
DP-MTV	51.1±2.6	52.0±1.2	50.6±3.3

E.5 Head-Selection Granularity: Per-Head vs. Per-Layer

The fully-private variant releases one selection probability per (layer, head) pair, giving $p = |\mathcal{S}|H$. Since its sensitivity scales as $\sqrt{p}/c$ (Theorem 4.3), we evaluate a coarser variant that selects entire layers, reducing p from 1024 to 32 and sensitivity by $\sqrt{32} \approx 5.7\times$.

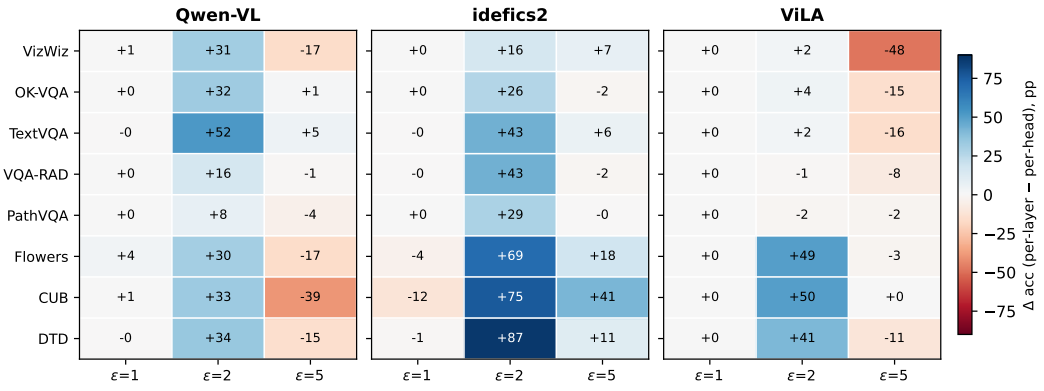


Figure 6: Effect of head-selection granularity for fully-private DP-MTV. Each cell reports the accuracy difference between per-layer and per-head selection (per-layer − per-head), in percentage points. Positive values favor per-layer selection and negative values favor per-head selection. Per-layer selection substantially improves utility at ε=2 for Qwen-VL and idefics2 and on ViLA classification, while the preferred granularity becomes model-dependent at ε=5.

These results show that coarser layer-level selection can substantially improve the fully-private variant at tighter privacy budgets, while fine-grained head selection becomes advantageous for some models as the privacy budget increases.