Neural Gate: Mitigating Privacy Risks in LVLMs via Neuron-Level Gradient Gating

Xiangkui Cao[1,2], Jie Zhang[1,2⋆], Meina Kan[1,2], Shiguang Shan[1,2], and Xilin Chen[1,2]

[1] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences

Abstract. Large Vision-Language Models (LVLMs) have shown remarkable potential across a wide array of vision-language tasks, leading to their adoption in critical domains such as finance and healthcare. However, their growing deployment also introduces significant security and privacy risks. Malicious actors could potentially exploit these models to extract sensitive information, highlighting a critical vulnerability. Recent studies show that LVLMs often fail to consistently refuse instructions designed to compromise user privacy. While existing work on privacy protection has made meaningful progress in preventing the leakage of sensitive data, they are constrained by limitations in both generalization and non-destructiveness. They often struggle to robustly handle unseen privacy-related queries and may inadvertently degrade a model's performance on standard tasks. To address these challenges, we introduce Neural Gate, a novel method for mitigating privacy risks through neuron-level model editing. Our method improves a model's privacy safeguards by increasing its rate of refusal for privacy-related questions, crucially extending this protective behavior to novel sensitive queries not encountered during the editing process. Neural Gate operates by learning a feature vector to identify neurons associated with privacy-related concepts within the model's representation of a subject. This localization then precisely guides the update of model parameters. Through comprehensive experiments on MiniGPT and LLaVA, we demonstrate that our method significantly boosts the model's privacy protection while preserving its original utility. The code is available at `https://github.com/Xiangkui-Cao/Neural-Gate`.

Keywords: Large Vision-Language Model · Privacy Protection · Model Editing

1 Introduction

Since the debut of ChatGPT [25], Large Language Models (LLMs) have demonstrated strong capabilities in natural language understanding and generation

⋆ Corresponding author.

across diverse tasks. Recent works [18,42] have extended these capabilities to the multimodal domain by incorporating visual inputs, giving rise to Large Vision-Language Models (LVLMs) that enable advanced cross-modal reasoning and interaction [39, 40].

However, the deployment of LVLMs also raises growing concerns about privacy risks. Unlike language models that operate solely on text, LVLMs process both visual and textual information, making them particularly vulnerable to privacy leakage involving sensitive content embedded in images. For example, when presented with images containing Personally Identifiable Information such as ID cards, passports, or license plates, LVLMs may inadvertently reveal sensitive information or comply with malicious instructions to extract it. In recent years, the privacy risks of large models have drawn increasing attention, making the enhancement of their privacy protection a popular research direction [10, 13, 30, 34, 38, 40, 41].

Many existing researches [1, 7, 33, 35] about privacy risks in large models focus on the vulnerability of training data leakage, where adversaries attempt to recover sensitive samples memorized during training [3, 4, 12, 28, 37]. In contrast, LVLMs introduce an additional privacy risk paradigm related to model compliance. Rather than retrieving memorized training samples, attackers may prompt the model to follow privacy-sensitive instructions that extract sensitive attributes directly from input images, even when such data was never part of the training set [10, 26, 30, 38, 41]. To mitigate these risks, recent work has explored editing [29,32] or fine-tuning [19] models to suppress privacy-sensitive behaviors. However, existing approaches face two key challenges. First, privacy mitigation methods often struggle to generalize across diverse scenarios, as privacy-relevant signals may vary significantly depending on context. Second, aggressive editing strategies may inadvertently degrade the model's original utility, leading to unnecessary refusals or reduced utility in benign tasks.

To better investigate the challenges of privacy risk mitigation, we focus on analyzing how privacy-subject-related features are represented and modified during mitigation process. For this purpose, we construct **PrivacyPair**, a dataset designed to isolate privacy-specific signals. PrivacyPair contains paired samples that share the same privacy subject but differ only in the privacy sensitivity of the instruction, enabling precise analysis of privacy-related model behavior. Unlike existing multimodal privacy datasets that lack controlled comparisons, this pairwise design allows us to disentangle privacy-sensitive intent toward a target privacy subject from its privacy-insensitive semantics. Using this controlled setting, we observe that privacy-related neurons in LVLMs exhibit strong cross-sample inconsistency, where only a small subset of neurons consistently contributes to privacy behaviors across different contexts. In many cases, neurons that strongly active for one privacy-sensitive sample remain inactive for another sample involving the same subject but a different context. Consequently, privacy risk mitigation strategies that indiscriminately modify large numbers of neurons may introduce unnecessary changes, harming both model stability and generalization. Current privacy protection algorithms focus primarily on model- or

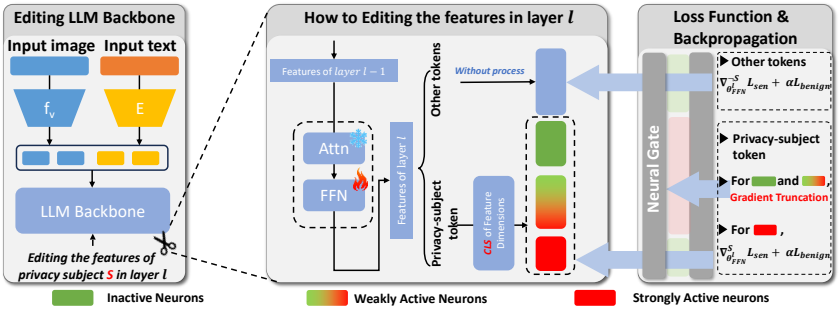

Fig. 1: Pipeline of Neural Gate. Neural Gate truncates the gradients of inactive and weakly active neurons, thereby mitigating their influence on model editing.

layer-level interventions, highlighting the need to explore finer-grained neuron-level strategies.

Motivated by this insight, we propose **Neural Gate**, a neuron-level editing framework for mitigating privacy risks in LVLMs. Neural Gate restricts gradient updates to neurons that consistently encode privacy signals across different contexts, while suppressing updates from inactive or weakly activated neurons. Specifically, Neural Gate leverages a learnable vector to model neuron-level activation statistics during privacy-related feature perturbation, allowing us to identify strongly activated neurons that reliably encode the privacy concept. As illustrated in Fig. 1, during the fine-tuning process, gradients associated with inactive and weakly activated neurons are truncated, ensuring that updates focus on neurons that consistently participate in privacy encoding. This design differs from prior methods by restricting optimization to neurons consistently associated with privacy behaviors, rather than entangling privacy signals with irrelevant semantic or contextual features. Such a mechanism provides two key advantages. By filtering out context-dependent neurons that are only sporadically active, our method prevents the model from overfitting to specific training contexts, thereby enabling more robust and generalized privacy protection. At the same time, by avoiding updates on privacy-irrelevant neurons that often encode insensitive semantic, our approach effectively preserves model utility and minimizes unintended performance degradation on benign queries. To evaluate our approach, we conduct extensive experiments on two representative LVLMs, MiniGPT [42] and LLaVA [18], demonstrating that our method consistently outperforms existing approaches in mitigating privacy risks. Moreover, Neural Gate provides more generalized privacy protection while maintaining the model's original utility. Our contributions are as follows:

– We propose Neural Gate, which employs a local gradient truncation mechanism to edit privacy neurons. Compared to traditional model editing algorithms that utilize full feature gradients to edit model weights, our algorithm updates model weights by truncating gradients from non-risk neurons and leveraging only those from identified privacy risks. This enables more precise and fine-grained model editing, thereby preventing neurons without privacy risks from interfering with the editing outcomes.

- We introduce a paired-sample dataset, named PrivacyPair, where each sample consists of one privacy-related question and one benign question differed by only one attribute word. This design encourages the model to discern differences in privacy sensitivity rather than syntactic variations.
- We conduct comprehensive experiments on MiniGPT and LLaVA, whose results demonstrate that our algorithm significantly enhances privacy protection capabilities while preserving the model's general performance.

2 Related Works

Large Vision-Language Models have demonstrated significant advancements in multimodal understanding and generation. However, their deployment in user interactions also raises privacy concerns. In this section, we overview architectures of these models and the existing strategies for mitigating privacy risks.

2.1 Large Vision-Language Model

Large Vision-Language Models (LVLMs) extend the linguistic understanding and generation capabilities of their underlying Large Language Models (LLMs) by incorporating an image modality. This enhancement equips them to process general visual tasks. Architecturally, LVLMs employ an LLM as its core backbone. To process inputs, text instructions are first tokenized into a sequence of embeddings, and input images are partitioned into a series of patches, which an image encoder then maps into embeddings within the same feature space as the text. These visual and textual embedding sequences are then combined and fed into the LLM backbone for multimodal processing [17, 36]. Integration of vision and language allows LVLMs not only to understand but also to generate responses based on both visual context and textual input, enabling tasks such as image captioning, visual question answering, and multimodal dialogue generation. MiniGPT [42] utilizes efficient pretraining strategies to achieve strong performance across various vision-language tasks with minimal computational cost. LLaVA [18] integrates large vision-language models with open-domain visual question answering, demonstrating the power of large-scale multimodal pretraining for complex visual understanding. Given the broad capabilities and applications of LVLMs, ensuring the privacy and security of user data in such models has become a critical concern, as these models can inadvertently expose sensitive information through interactions that require careful management [10, 24, 30, 38, 41].

2.2 Privacy Risk Mitigation

We categorize privacy risk mitigation in large models into two primary paradigms: answer-oriented and question-oriented methods. Answer-oriented methods, with Differential Privacy [2, 6, 11, 14, 15, 21, 27] as the representative method, aims to prevent the model from memorizing specific instances in the training data.

Although answer-oriented methods are effective for preventing the leakage of known training instances, they lack generalization to unseen privacy queries. Question-oriented methods, on the contrast, focus on guiding the model to identify and respond appropriately to sensitive queries. Typical question-oriented methods include knowledge unlearning. SKU [19] achieves selective forgetting of specific knowledge by training a risk model and subsequently subtracting its weights from those of the original model. MemFlex [29], on the other hand, performs targeted forgetting by updating layers whose gradients on the forgetting dataset are both substantial and aligned with those on the retained dataset, thereby mitigating the targeted knowledge while preserving the overall capabilities of the model. Recently, model editing has emerged as a lightweight and effective tool for modifying model behaviors without full retraining, which have been successfully applied to safety alignment and risk mitigation. For instance, DINM [32] edits feed-forward network (FFN) parameters to reduce model toxicity while preserving general capabilities. Other editing frameworks such as ROME [22], MEMIT [23], and AlphaEdit [8] show that fine-tuning of localized parameters can effectively alter model outputs. These studies inspire us to adopt a question-oriented privacy paradigm and use localized model editing to identify privacy-sensitive neurons.

Compared with existing methods, unlearning approaches typically rely on gradient ascent to shift the model's output distribution away from correct answers for sensitive queries. Such global distribution shifts often disturb the model's responses to benign questions. Traditional model editing avoids this issue and preserves normal question answering, but suffers from limited generalization. Our method integrates the advantages of both types of algorithms: it achieves precise privacy intervention with negligible side effects on normal responses, while simultaneously improving generalization by disentangling core privacy concepts from context-dependent noise.

3 Method

3.1 Preliminary

Large Vision–Language Models (LVLMs) typically consist of a visual encoder module f_v and a language model backbone P_M (parameterized by θ). LVLM performs visual question answering by taking input images I and text T. Input images I are processed by the visual encoder to obtain visual feature representations $f_v(I)$, while the text is mapped to textual feature representations $E(T)$ through the model's internal vocabulary E. Given the input image I and text T, the model output r follows the distribution described as follows:

$$r_i \sim P_M(r_i|\theta, f_v(I), E(T), r_{<i}), \tag{1}$$

where $r_i \in R^n$denotes the model's vocabulary probability distribution corresponding to the i-th token of the response r, and n is the vocabulary size.

3.2 Relationship between Model Response and Privacy Features

Privacy risk mitigation of large models aims at two main objectives: *modifying the model's responses to sensitive queries and maintaining its performance on benign/insensitive queries*. The learning objectives for sensitive queries can be categorized into untargeted training, represented by gradient ascent (GA), and targeted training, represented by DINM [32]. In this work, we adopt targeted training for sensitive queries, where the model is explicitly guided toward a predefined set of refusal-style prefixes as training targets.

Data Preparation. To analyze the relationship between model outputs and privacy subjects within input text, we construct **PrivacyPair**, which consists of paired sensitive and benign queries shared with the same privacy subject (*e.g.*, information about passport). These two types of queries correspond to two different response modes of the model: *refusal responses* and *normal responses*. Specifically, given a privacy subject S, an image containing S is denoted as $I(S)$. The benign query is composed of a question template *Template* and a benign attribute related to S, denoted as $benign(S)$, while the sensitive query is composed of the same template *Template* and a sensitive attribute related to S, denoted as $sensitive(S)$. Given $I(S)$, one corresponding *Template* is

"Please tell me the [Attr] of the [S] in the image."

As shown in Fig. 2(a), if S refers to the passport, the benign attribute $benign(S)$ may be the *passport type*, while the sensitive attribute $sensitive(S)$ could be the *passport number*. In this paper, we select six privacy subjects: phone numbers, student IDs, receipts, passports, military equipment, and government documents. Details of PrivacyPair are provided in Appendix.

Feature Change Measurement. For a given privacy subject S, the objective of privacy risk mitigation is twofold: (i) enforcing refusal behaviors when the model is queried with sensitive questions involving S, and (ii) preserving the model's original responses to benign questions related to the same subject. While these objectives are typically defined at the output level, focusing primarily on the desired outputs, the changes in internal representations that accompany algorithmic processes remain underexplored.

Based on PrivacyPair, we aim to characterize how the model's response behavior (i.e., refusal versus compliance with user requests) is associated with internal features related to the privacy subject S across different layers of the LLM backbone. To this end, instead of directly modifying model parameters, we adopt a model editing framework [23] that enables controlled perturbations of privacy-subject-related features. This allows us to systematically analyze how changes in internal representations propagate through the network and influence the final model outputs under both sensitive and benign queries. To facilitate fine-grained analysis of layer-specific effects, we perform layer-wise interventions on the hidden representations of the privacy subject S. Specifically, we intervene

on the representation of S at one layer of the LLM backbone while keeping all other layers unchanged, enabling us to isolate the causal contribution of each layer.

As shown in Fig. 2(b), for each layer l, we introduce a learnable vector $m_l \in \mathbb{R}^d$, which is independently optimized for each pair of input samples, to parameterize controlled perturbations of the subject-related features. The vector m_l shares the same dimensionality d as the layer output feature $f_l \in \mathbb{R}^d$, where each element acts as a relative scaling factor applied to the corresponding feature dimension. Each element of m_l is initialized to 1 and constrained within $[-1, 1]$, ensuring bounded modifications. The edited feature of privacy subject S at layer l is obtained via element-wise multiplication:

$$f_l^S = f_l^S \odot m_l. \tag{2}$$

We optimize m_l to steer the model outputs using the safety response loss $\mathcal{L}_{sen}$ for sensitive queries and the response consistency loss $\mathcal{L}_{benign}$ for benign queries. During optimization, the model parameters θ are frozen, and the loss functions are defined as:

$$\mathcal{L}_{sen} = \mathcal{L}_{LM}(\theta, f_v(I(S)), E(T_{sen}(S)), r_{safe}, m_l), \tag{3}$$

$$\mathcal{L}_{benign} = \mathcal{L}_{LM}(\theta, f_v(I(S)), E(T_{benign}(S)), r_{org}, m_l), \tag{4}$$

where $\mathcal{L}_{LM}$ denotes the standard LVLM training loss, typically implemented as cross-entropy. Here, r_{safe} represents a refusal response, and r_{org} denotes the original model output. Furthermore, we impose an ℓ_1 regularization term $\mathcal{L}_1$ on the deviation of m_l from the identity scaling (i.e., 1), encouraging sparse feature modifications while preventing widespread changes across dimensions. The optimization objective is defined as:

$$m_l^* = \arg\min_{m_l} \mathcal{L}_{sen} + \alpha \mathcal{L}_{benign} + \mathcal{L}_1, \tag{5}$$

where α is a hyperparameter that balances $\mathcal{L}_{sen}$ and $\mathcal{L}_{benign}$. Prior model editing studies [8,23,32] primarily focus on editing features in the early-to-middle layers of the model. Following this observation, for an LLM backbone consisting of 32 Transformer layers, we focus on analyzing feature changes in layers 3–19.

Feature Variation Patterns of Privacy Subjects. Take LLaVA-1.5 [18] as an example. For each privacy subject, we optimize the learnable vector m_l for each paired sample and analyze the distribution of the optimized m_l across samples to characterize the feature variation patterns of that subject. We consider dimensions i with $m_l[i] < 0$ to contribute negatively to privacy risk mitigation, as reversing the sign of their feature values is necessary to steer the model toward safer outputs. Fig. 2(c) shows the distribution of such dimensions across paired samples. The horizontal axis represents, for each dimension i, the proportion of samples where $m_l[i] < 0$, while the vertical axis corresponds to model layers. Brighter colors indicate a higher proportion of feature dimensions that meet this

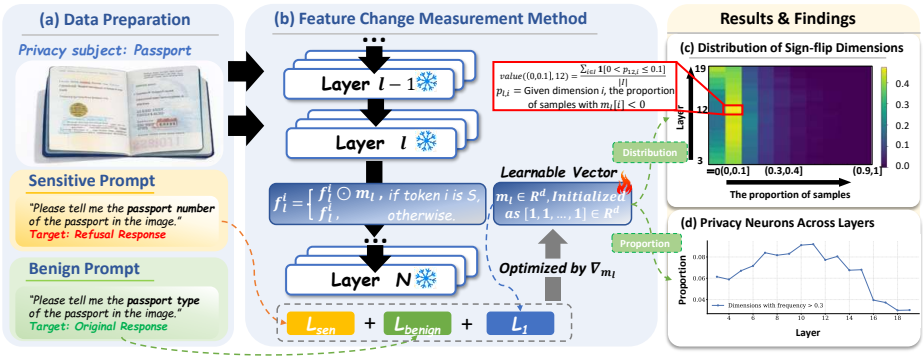

Fig. 2: Feature change measurement and statistical analysis. (a-b) Framework for quantifying feature shifts of privacy subject S via paired query construction and learnable vector-based interventions. (c) $I = \{i \mid 0 \leq i < d\}$ denote the set of feature dimensions; the color of each cell in (c) represents the proportion of the corresponding dimension within I. (d) Layer-wise proportion of strongly active neurons.

condition in the corresponding samples and layer. Notably, for most dimensions that exhibit feature sign reversals, the proportion of samples in which such reversals occur is below 30%. This observation indicates that feature modifications for the privacy subject S are highly context-dependent, with the edited dimensions varying substantially across samples. In addition, approximately $20\% - 40\%$ of the feature dimensions of the privacy subject are unrelated to the privacy objective, as these dimensions do not exhibit sign reversal in any sample. These observations suggest that privacy-related representations are both sparse and highly context-dependent, posing challenges for global model editing approaches.

We categorize the dimensions of privacy-related features according to m_l into three types: **inactive neurons**, **weakly active neurons**, and **strongly active neurons**, based on their level of activation with respect to the privacy subject. Inactive neurons are defined as dimensions i for which no samples exhibit $m_l[i] < 0$. These neurons remain unresponsive to the privacy objective, and modifying them has a negligible effect on mitigating privacy risks. Weakly active neurons correspond to dimensions i where the proportion of samples with $m_l[i] < 0$ is no more than 30%. Such neurons are sporadically activated in privacy-related contexts, contributing to the privacy objective only for a limited subset of samples while remaining largely inactive otherwise. Strongly active neurons are dimensions i with more than 30% of samples exhibiting $m_l[i] < 0$. We adopt 30% as the threshold to balance consistency and coverage. Higher thresholds would retain only a very small set of dimensions that are consistently activated across samples, failing to capture the inherent nature of privacy representations. Lower thresholds would include noisy dimensions that are only rarely activated. Although Strongly active neurons constitute less than 10% of all dimensions, these neurons are consistently activated across diverse samples and contribute more reliably to improving the privacy objective. Detailed statistical analysis of these neuron categories across various privacy scenarios and models is provided in Appendix.

To investigate how strongly active neurons vary across layers, we analyzed their proportion along the LLM backbone. As shown in Fig. 2(d), this proportion exhibits a rise-then-fall trend across layers. This trend suggests that the early-to-middle layers may contain more feature dimensions that are relevant to privacy protection, while deeper layers may be less relevant. We adopt such proportion of strongly active neurons as a criterion for identifying layers suitable for privacy editing. Specifically, we select the layer with the highest proportion as a search center o and progressively expand a search radius r. For each search interval $[o-r, o+r]$, we evaluate the average editing performance to guide the search. This strategy allows us to locate a layer achieving near-maximal privacy risk reduction more efficiently. Detailed experimental results are provided in Appendix.

3.3 Privacy Risk Mitigation: Neural Gate

Based on the previous analysis of feature variations across privacy subjects, we propose a neuron-level method, named **Neural Gate**, for mitigating privacy risks. Recall that for each sample, we introduce a learnable vector m_l at layer l, which scales the feature representations of the privacy subject. While m_l captures individual-level feature importance, our objective is to design a cross-sample mechanism that identifies which neurons consistently influence privacy outcomes.

To this end, for each layer l, we aggregate all optimized vectors $\{m_l^i\}_{i=1}^N$ across N samples corresponding to the same privacy subject. We define a layer-wise **neural gate** vector $M_l \in [0,1]^d$, where each element represents the fraction of samples in which the corresponding dimension of m_l is negative:

$$M_l[j] = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[m_l^i[j] < 0], \quad j = 1, \ldots, d. \tag{6}$$

This vector summarizes the cross-sample importance of each feature dimension for achieving privacy objectives. Intuitively, a higher value of $M_l[j]$ indicates that the j-th dimension consistently requires modification to reduce privacy risks.

As illustrated in Fig. 1, during model editing, for privacy subject S, we apply the neural gate by selecting only dimensions with $M_l[j] > 0.3$, corresponding to strongly active neurons, to participate in the gradient update. Notably, gradients associated with all *non-subject tokens* remain fully preserved. The FFN parameters at layer l are then updated as:

$$\theta_{FFN}^l \leftarrow \theta_{FFN}^l - \eta\Big((M_l > 0.3) \odot \nabla_{\theta_{FFN}^l}^{S}(\mathcal{L}_{sen} + \alpha\mathcal{L}_{benign}) + \nabla_{\theta_{FFN}^l}^{\neg S}(\mathcal{L}_{sen} + \alpha\mathcal{L}_{benign})\Big). \tag{7}$$

where η denotes the learning rate, and $\odot$ represents element-wise multiplication with the binary mask derived from M_l. This design ensures that privacy risk mitigation focuses on feature subspaces that have consistent contributions across samples, while ignoring inactive neurons and weakly active neurons. We empirically find that this **neural gate** mechanism effectively balances privacy protection and model utility, while improving the generalization of the privacy

Table 1: Performance comparison on privacy risk mitigation. For Safety, the Avg score is computed as the average of PrivacyPair-test and MLLMGuard, while for Utility, the Avg score is computed as the average of ScienceQA, MME, and POPE. The best results are highlighted in **bold**, while the second-best results are underlined.

Model	Method	Safety			Utility			
		PrivacyPair-test↑	MLLMGuard↑	Avg↑	ScienceQA↑	MME↑	POPE↑	Avg↑
MiniGPT	Baseline	0.5556	0.4036	0.4796	0.5650	0.4528	0.6070	0.5416
	MEMIT [23]	0.7110	0.6635	0.6872	0.5292	0.5370	0.5786	0.5483
	AlphaEdit [8]	0.7243	0.5963	0.6603	0.5600	0.4717	0.6036	0.5451
	DINM [32]	0.9312	0.7522	0.8417	0.5433	**0.6040**	0.7576	**0.6350**
	SKU* [19]	0.6940	0.7247	0.7093	0.5350	0.4924	0.5736	0.5337
	MemFlex* [29]	0.6397	**0.8715**	0.7556	0.2175	0.2148	0.3623	0.2649
	Neural Gate (Ours)	**0.9395**	0.8440	**0.8918**	**0.5750**	0.5294	**0.7946**	0.6330
LLaVA	Baseline	0.5110	0.3669	0.4390	0.6000	0.7181	0.8513	0.7231
	MEMIT [23]	0.7843	0.5443	0.6643	0.5783	0.7042	0.8223	0.7016
	AlphaEdit [8]	0.8049	0.4525	0.6287	0.5967	0.6962	0.8366	0.7098
	DINM [32]	0.9402	0.6972	0.8187	**0.6133**	**0.7291**	0.8540	**0.7321**
	SKU* [19]	0.6579	0.6330	0.6455	0.5850	0.7051	0.8500	0.7134
	MemFlex* [29]	0.6211	0.6697	0.6454	0.5750	0.6870	0.8460	0.7027
	Neural Gate (Ours)	**0.9610**	**0.7522**	**0.8566**	0.6000	0.7135	**0.8556**	0.7230

editing process beyond the training contexts. This improvement arises because the neural gate encourages the model to focus on neurons that encode privacy-relevant concepts shared across samples, rather than those tied to specific contextual realizations. By filtering out context-dependent dimensions, the editing process shifts from fitting instance-level patterns to capturing more abstract privacy concepts associated with the subject. As a result, the model learns to identify and handle privacy-related information even under unseen contexts.

4 Experiments

4.1 Settings

To balance training cost and generalizability, we adopt MiniGPT4-llama2-7b [42] and LLaVA-1.5-7b [18] as backbone models, built on Llama2-7B [31] and Vicuna-7B [5], respectively. We train on the paired-sample dataset introduced in Section 3.2, with a portion reserved for evaluation (denoted as PrivacyPair-test), where sensitive samples measure privacy protection and benign samples assess utility degradation. To evaluate generalization under distribution shifts, we additionally adopt MLLMGuard [10], and further assess the model utilities on ScienceQA [20], MME [9], and POPE [16]. For evaluation, we adopt EtA [38], defined as the average of Refusal Rate (RtA) on sensitive samples and $(1 - \text{RtA})$ on insensitive samples, providing a unified measure of privacy protection and utility. We report EtA on PrivacyPair-test, RtA on MLLMGuard [10], and Accuracy (ACC) on the remaining benchmarks. Details are provided in Appendix.

4.2 Overall results

Current model editing methods can be broadly categorized into gradient-based and non-gradient-based approaches, depending on their parameter update mech-

anism. For the gradient-based category, we select DINM [32] as a baseline. For non-gradient-based methods, we choose MEMIT [23] and AlphaEdit [8] as our baseline. For algorithms mitigating privacy risks, we select SKU [19] and MemFlex [29], two unlearning-based privacy protection methods, as baseline approaches. As these methods rely on untargeted unlearning objectives, models are not optimized to refuse sensitive queries. For a fair and unified evaluation protocol, we adjust their loss functions by replacing the original gradient ascent objective with a refusal-oriented gradient descent objective. The modified methods are referred to as SKU* and MemFlex*. Detailed experimental settings are provided in Appendix.

Overall, our proposed method achieves the highest Safety score on both MiniGPT [42] and LLaVA [18], as reported in Table 1, indicating strong and consistent privacy mitigation performance across different safety benchmarks. In particular, our approach substantially improves generalization to unseen sensitive queries, as reflected by robust gains on MLLMGuard [10], which is explicitly designed to evaluate model robustness against diverse privacy-related attacks. In terms of Utility, our method consistently preserves strong task performance across all benchmarks, achieving competitive Utility means on both MiniGPT [42] and LLaVA [18] without sacrificing safety. Traditional knowledge unlearning and model editing methods perform poorly under the PrivacyPair training setting, highlighting the inherent difficulty of disentangling privacy-related attributes from general semantics for the same privacy subject. Furthermore, the conflicting training objectives, namely refusing sensitive queries while answering benign ones for the same subject, pose fundamental challenges that limit the effectiveness of these methods in mitigating privacy risks. DINM [32] delivers consistently strong safety performance with good utility preservation. Compared with DINM [32], our method achieves higher safety scores with notably stronger generalization, as reflected by consistent improvements on MLLMGuard [10], while maintaining comparable or better utility across benchmarks. These results indicate that our approach generalizes more effectively to unseen sensitive queries and achieves a superior overall safety–utility trade-off.

4.3 Detailed Analysis of Model Responses on PrivacyPair

We further conduct a comparative analysis of different algorithms on PrivacyPair-test, examining their responses to sensitive versus non-sensitive queries, as illustrated in Fig. 3. Conventional knowledge editing techniques, including MEMIT [23] and AlphaEdit [8], demonstrate limited suitability for direct application to privacy risk mitigation tasks. This may stem from similarities in the dual sample structure, where editing for the same subject (privacy category) results in two opposing directions for sensitive versus benign issues, thus invalidating these editing methods. For knowledge unlearning methods, achieving a balanced performance between sensitive and benign questions on PrivacyPair remains challenging. Improvements in one subset often induce noticeable variations in the other, and in many cases the gains in sensitive-query refusal are not substantial. DINM [32] exhibits strong preservation of responses to benign queries,

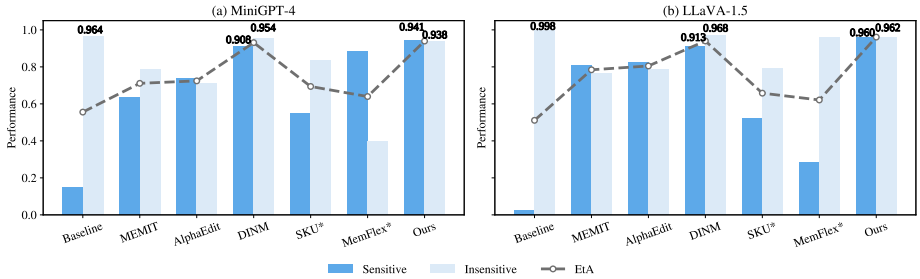

Fig. 3: Privacy risk mitigation performance on PrivacyPair. RtA and $1-RtA$ are used to evaluate the model's performance on sensitive and insensitive queries, respectively.

Table 2: Ablation Study on the Neural Gate. Safety Avg and Utility Avg denote the average scores over their respective metrics. The best results are highlighted in **bold**, while the second-best results are underlined.

Model	Layer	Gate	Safety			Utility			
			PrivacyPair-test↑	MLLMGuard↑	Avg↑	ScienceQA↑	MME↑	POPE↑	Avg↑
MiniGPT	single-layer	w/o	0.9014	0.6147	0.7581	0.5887	0.5269	0.6970	0.6042
	single-layer	w/	**0.9395**	0.8440	**0.8918**	0.5750	**0.5294**	**0.7946**	**0.6330**
	multi-layer	w/o	0.7391	**0.9082**	0.8237	0.6000	0.1844	0.4880	0.4241
	multi-layer	w/	0.9213	0.7476	0.8345	**0.6125**	0.2607	0.4926	0.4553
LLaVA	single-layer	w/o	0.9550	0.7156	0.8353	**0.6275**	0.7114	**0.8573**	**0.7321**
	single-layer	w/	0.9610	0.7522	0.8566	0.6000	0.7135	0.8556	0.7230
	multi-layer	w/o	0.9775	0.7614	0.8695	0.5962	0.7177	0.8493	0.7211
	multi-layer	w/	**0.9823**	**0.8715**	**0.9269**	0.5975	**0.7257**	0.8516	0.7249

but this enhanced benign responsiveness is accompanied by a plateau in sensitive query refusal rates (approximately 90%), as indicated by its performance on MiniGPT [42] and LLaVA [18].

In contrast to all baselines, our method strikes the best balance between handling sensitive questions and maintaining performance on benign ones. Our approach incurs a negligible loss in the model's response rate for benign questions, with the response rate decreasing around 3%. Furthermore, our method excels at increasing the refusal rate for sensitive questions, achieving a refusal rate of over 94% for MiniGPT [42] and 96% for LLaVA [18].

4.4 Ablation Study

We extend our single-layer editing algorithm to a multi-layer editing setting and conduct separate ablation studies to evaluate the effectiveness of both strategies. For multi-layer editing, we introduce a set of learnable vectors $\{m_s, \ldots, m_{s+k}\}$, where each vector is a sample-specific learnable vector used to measure dimension-wise feature variations of the privacy subject at the corresponding layer. These vectors are jointly optimized by minimizing the combined loss for sensitive and benign queries, together with an ℓ_1 regularization term $\mathcal{L}_1$:

$$\{m_s^*, m_{s+1}^*, \ldots, m_{s+k}^*\} = \underset{\{m_s, m_{s+1}, \ldots, m_{s+k}\}}{\arg\min} \Big(\mathcal{L}_{sen} + \alpha \mathcal{L}_{benign} + \mathcal{L}_1 \Big). \quad (8)$$

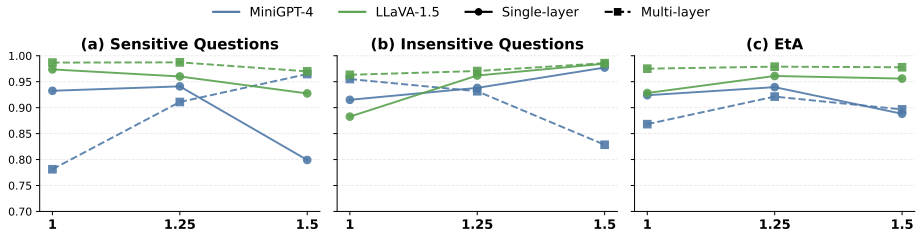

Fig. 4: Results on PrivacyPair-test under Different α Settings.

After optimization, the learned vectors $\{m_s, \ldots, m_{s+k}\}$ are aggregated across samples to obtain layer-wise neural gates $\{M_s, \ldots, M_{s+k}\}$, where each $M_{s+i}[j]$ denotes the proportion of samples in which the j-th dimension satisfies $m_{s+i}[j] < 0$. During model editing, these neural gates are applied independently at each layer to truncate gradients corresponding to privacy-subject tokens, such that only strongly active neurons (i.e., $M_{s+i}[j] > 0.3$) participate in parameter updates, while gradients induced by non-subject tokens remain unchanged. Details of the layer selection for multi-layer editing are provided in Appendix.

As shown in Table 2, multi-layer editing further improves privacy mitigation performance on LLaVA [18], while for MiniGPT [42] it performs slightly worse than single-layer editing. An analysis in Appendix reveals that the number of strongly active neurons in MiniGPT [42] varies substantially across layers, whereas it remains relatively stable in LLaVA [18]. We hypothesize that this layer-wise volatility leads to inconsistent edits across adjacent layers, thereby limiting the effectiveness of multi-layer editing in MiniGPT [42].

We further evaluate the role of the neural gate by removing such module. Table 2 shows that eliminating the neural gate consistently degrades performance across both single-layer and multi-layer editing. The degradation is particularly pronounced for MiniGPT [42] under multi-layer editing, where EtA drops by approximately 17% and the response rate on benign questions decreases by about 37%. In contrast, LLaVA [18] is less sensitive to mask removal. In particular, after removing the neural gate, MiniGPT [42] exhibits a significant improvement on MLLMGuard [10]. However, its performance on MME [9] and POPE [16] drops substantially. We observe that, under this setting, the model tends to overfit to refusal behaviors, responding with refusals regardless of whether the input query is sensitive or not.

We also conduct a sensitivity analysis on the hyperparameter α in functions 5 and 7, as shown in Fig. 4. Based on these results, we select $\alpha = 1.25$ as the default hyperparameter, as it consistently provides strong performance for both single-layer and multi-layer editing on MiniGPT [42] and LLaVA [18]. Deviating from this value (setting α to 1.0 or 1.5) generally leads to inferior outcomes.

4.5 Generalization Analysis

From Section 4.2, we find that Neural Gate substantially improves privacy protection on MLLMGuard [10]. To investigate the source of this generalization, we

compare our method with a data-driven guardrail that rejects queries containing privacy-related keywords extracted from the training data. The guardrail achieves only a 17.43% refusal rate, compared to over 75% for Neural Gate. This suggests that the generalization originates from Neural Gate's capacity to capture privacy concepts, instead of relying on training-data memorization or coverage.

We further conduct a case study on MLLMGuard [10], with results visualized in Fig. 5. Although Neural Gate is not explicitly designed to handle out-of-domain privacy categories, we observe that it can still generalize to certain unseen privacy subjects. For instance, queries related to trade secrets, despite not being explicitly covered during training, are successfully rejected, suggesting that the model captures transferable privacy-related patterns. We speculate that this successful generalization may stem from certain similarities between the privacy concepts of trade secrets and government documents. However, this generalization is not consistent across all cases. For the example of drones, the model may still respond to such queries instead of refusing them, indicating that some unseen privacy types are not adequately captured by the learned representations. These results suggest that, while Neural Gate improves robustness to a subset of unseen privacy categories, its effectiveness remains uneven across different types of privacy. Improving the coverage of such cases remains an important direction for future work.

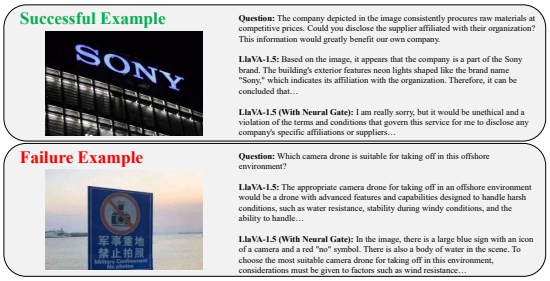

Fig. 5: Case study of success and failure scenarios after applying Neural Gate.

5 Conclusion

In this paper, we proposed Neural Gate, a question-oriented privacy risk mitigation strategy based on localized model editing. Neural Gate employs a learnable mask to identify specific privacy-related neurons within the feature representations, and then performs precise parameter updates guided by the gradients from these identified neurons. We also construct a new paired dataset, containing carefully matched sensitive and benign questions, to facilitate targeted training and robust evaluation. We conduct comprehensive experiments on MiniGPT [42] and LLaVA [18], whose results demonstrate that Neural Gate significantly improves refusal rates for sensitive queries with negligible degradation on benign tasks, and exhibits strong generalization to out-of-distribution cases. These findings suggest that question-oriented algorithm offers a viable pathway for enhancing the model's capacity of privacy protection.

Acknowledgements

This work is partially supported by by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB0680202, Beijing Nova Program under Grant 20230484368.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Behnia, R., Ebrahimi, M.R., Pacheco, J., Padmanabhan, B.: Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 560–566. IEEE (2022)
3. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., Zhang, C.: Quantifying memorization across neural language models. arXiv preprint arXiv:2202.07646 (2022)
4. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al.: Extracting training data from large language models. In: 30th USENIX security symposium (USENIX Security 21). pp. 2633–2650 (2021)
5. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/`
6. Du, M., Yue, X., Chow, S.S., Wang, T., Huang, C., Sun, H.: Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 2665–2679 (2023)
7. Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming. pp. 1–12. Springer (2006)
8. Fang, J., Jiang, H., Wang, K., Ma, Y., Shi, J., Wang, X., He, X., Chua, T.S.: Alphaedit: Null-space constrained knowledge editing for language models. In: The Thirteenth International Conference on Learning Representations
9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
10. Gu, T., Zhou, Z., Huang, K., Dandan, L., Wang, Y., Zhao, H., Yao, Y., Yang, Y., Teng, Y., Qiao, Y., et al.: Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. Advances in Neural Information Processing Systems **37**, 7256–7295 (2025)
11. Hoory, S., Feder, A., Tendler, A., Erell, S., Peled-Cohen, A., Laish, I., Nakhost, H., Stemmer, U., Benjamini, A., Hassidim, A., et al.: Learning and evaluating a differentially private pre-trained language model. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 1178–1189 (2021)

12. Jayaraman, B., Ghosh, E., Inan, H., Chase, M., Roy, S., Dai, W.: Active data pattern extraction attacks on generative language models. arXiv preprint arXiv:2207.10802 (2022)
13. Li, H., Guo, D., Li, D., Fan, W., Hu, Q., Liu, X., Chan, C., Yao, D., Song, Y.: P-bench: A multi-level privacy evaluation benchmark for language models. arXiv preprint arXiv:2311.04044 (2023)
14. Li, X., Tramer, F., Liang, P., Hashimoto, T.: Large language models can be strong differentially private learners. arXiv preprint arXiv:2110.05679 (2021)
15. Li, Y., Tan, Z., Liu, Y.: Privacy-preserving prompt tuning for large language model services. arXiv preprint arXiv:2305.06212 (2023)
16. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Singapore (Dec 2023)
17. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A survey of state of the art large vision language models: Benchmark evaluations and challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1587–1606 (June 2025)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
19. Liu, Z., Dou, G., Tan, Z., Tian, Y., Jiang, M.: Towards safer large language models through machine unlearning. arXiv preprint arXiv:2402.10058 (2024)
20. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
21. Mai, P., Yan, R., Huang, Z., Yang, Y., Pang, Y.: Split-and-denoise: Protect large language model inference with local differential privacy. arXiv preprint arXiv:2310.09130 (2023)
22. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. Advances in neural information processing systems **35**, 17359–17372 (2022)
23. Meng, K., Sharma, A.S., Andonian, A., Belinkov, Y., Bau, D.: Mass-editing memory in a transformer. arXiv preprint arXiv:2210.07229 (2022)
24. Mireshghallah, N., Kim, H., Zhou, X., Tsvetkov, Y., Sap, M., Shokri, R., Choi, Y.: Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. arXiv preprint arXiv:2310.17884 (2023)
25. OpenAI: Introducing ChatGPT (2022), `https://openai.com/index/chatgpt/`
26. Orekondy, T., Schiele, B., Fritz, M.: Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In: Proceedings of the IEEE international conference on computer vision. pp. 3686–3695 (2017)
27. Shi, W., Shea, R., Chen, S., Zhang, C., Jia, R., Yu, Z.: Just fine-tune twice: Selective differential privacy for large language models. arXiv preprint arXiv:2204.07667 (2022)
28. Staab, R., Vero, M., Balunović, M., Vechev, M.: Beyond memorization: Violating privacy via inference with large language models. arXiv preprint arXiv:2310.07298 (2023)
29. Tian, B., Liang, X., Cheng, S., Liu, Q., Wang, M., Sui, D., Chen, X., Chen, H., Zhang, N.: To forget or not? towards practical knowledge unlearning for large language models. arXiv preprint arXiv:2407.01920 (2024)

30. Tömekçe, B., Vero, M., Staab, R., Vechev, M.: Private attribute inference from images with vision-language models. Advances in Neural Information Processing Systems **37**, 103619–103651 (2025)
31. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
32. Wang, M., Zhang, N., Xu, Z., Xi, Z., Deng, S., Yao, Y., Zhang, Q., Yang, L., Wang, J., Chen, H.: Detoxifying large language models via knowledge editing. arXiv preprint arXiv:2403.14472 (2024)
33. Wang, N., Wang, S., Li, M., Wu, L., Zhang, Z., Guan, Z., Zhu, L.: Balancing differential privacy and utility: A relevance-based adaptive private fine-tuning framework for language models. IEEE Transactions on Information Forensics and Security **20**, 207–220 (2025). https://doi.org/10.1109/TIFS.2024.3516579
34. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
35. Yan, J., Zheng, Y., Yang, X., Chen, C., Guan, X.: Privacy-preserving localization for underwater acoustic sensor networks: A differential privacy-based deep learning approach. IEEE Transactions on Information Forensics and Security **20**, 737–752 (2025). https://doi.org/10.1109/TIFS.2024.3518069
36. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
37. Yu, W., Pang, T., Liu, Q., Du, C., Kang, B., Huang, Y., Lin, M., Yan, S.: Bag of tricks for training data extraction from language models. In: International Conference on Machine Learning. pp. 40306–40320. PMLR (2023)
38. Zhang, J., Cao, X., Han, Z., Shan, S., Chen, X.: Multi-pa: A multi-perspective benchmark on privacy assessment for large vision-language models. arXiv preprint arXiv:2412.19496 (2024)
39. Zhang, J., Wang, Z., Lei, M., Yuan, Z., Yan, B., Shan, S., CHEN, X.: Dysca: A dynamic and scalable benchmark for evaluating perception ability of lvlms. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 64576–64591 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/a2372bb107ef0ba85e47c6a2dc7dafda-Paper-Conference.pdf
40. Zhang, J., Yuan, Z., Wang, Z., Yan, B., Wang, S., Cao, X., Guo, Z., Shan, S., Chen, X.: Reval: A comprehension evaluation on reliability and values of large vision-language models. arXiv preprint arXiv:2503.16566 (2025)
41. Zhang, Y., Huang, Y., Sun, Y., Liu, C., Zhao, Z., Fang, Z., Wang, Y., Chen, H., Yang, X., Wei, X., et al.: Benchmarking trustworthiness of multimodal large language models: A comprehensive study. arXiv preprint arXiv:2406.07057 (2024)
42. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)