

When Privacy Meets Recovery: The Overlooked Half of Surrogate-Driven Privacy Preservation for MLLM Editing

Siyuan Xu¹, Yibing Liu^{1*}, Peilin Chen¹, Yung-Hui Li², Shiqi Wang¹, Sam Kwong³

¹City University of Hong Kong

²Hon Hai Research Institute

³Lingnan University

siyuanxu333@gmail.com, lyibing112@gmail.com, plchen3@cityu.edu.hk,
yunghui.li@foxconn.com, shiqwang@cityu.edu.hk, samkwong@ln.edu.hk

Abstract

Privacy leakage in Multimodal Large Language Models (MLLMs) has long been an intractable problem. Existing studies, though effectively obscure private information in MLLMs, often overlook the evaluation of the authenticity and recovery quality of user privacy. To this end, this work uniquely focuses on the critical challenge of how to restore surrogate-driven protected data in diverse MLLM scenarios. We first bridge this research gap by contributing the SPPE (Surrogate Privacy Protected Editable) dataset, which includes a wide range of privacy categories and user instructions to simulate real MLLM applications. This dataset offers protected surrogates alongside their various MLLM-edited versions, thus enabling the direct assessment of privacy recovery quality. By formulating privacy recovery as a guided generation task conditioned on complementary multimodal signals, we further introduce a unified approach that reliably reconstructs private content while preserving the fidelity of MLLM-generated edits. The experiments on both SPPE and InstructPix2Pix further show that our approach generalizes well across diverse visual content and editing tasks, achieving a strong balance between privacy protection and MLLM usability.

Introduction

The rapid development of Multimodal Large Language Models (MLLMs) has unlocked powerful reasoning over complex visual inputs. However, this strength is often accompanied by serious privacy risks, as diverse and subtle private cues embedded in visual input can be easily captured and exposed. While prior work (Mishra et al. 2025) has leveraged textual surrogates (e.g., descriptions or summaries of visual content) to bypass direct image uploads for tasks such as visual question answering, these approaches fall short in scenarios that require direct manipulation of visual inputs, such as image editing. This motivates the use of visual surrogates—synthetic substitutes for sensitive regions—which have demonstrated strong privacy protection through perceptual concealment (Xu et al. 2024a), while retaining high utility for downstream analysis (Abdulaziz,

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

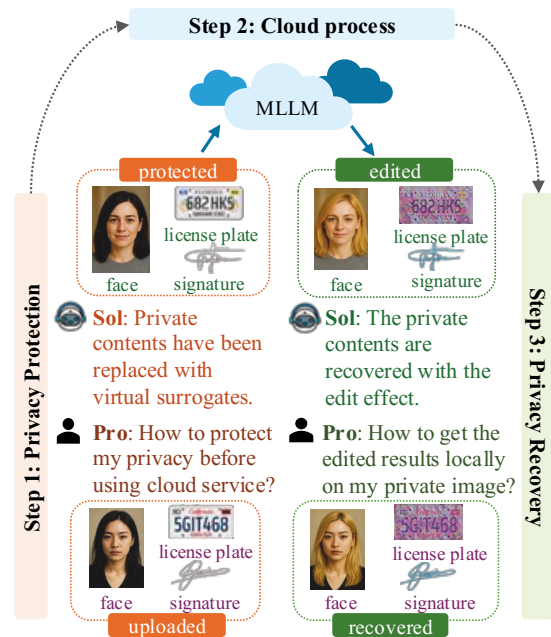


Figure 1: Demonstration of our Edit-Compatible Surrogate-Driven Privacy Protection paradigm. Sensitive regions in the original image are locally replaced with synthetic content to create a surrogate, which is sent to the cloud for editing by MLLMs. The surrogate’s edits are locally combined with the original image to produce a privacy-preserving output that faithfully reflects MLLM-intended modifications.

D’Amicantonio, and Bondarev 2025). However, edits performed on surrogates often deviate from those on the original image, which makes it difficult to preserve the visual-semantic consistency with the original. To address this, our work focuses on the post-surrogate stage, defining a recovery process that faithfully reconstructs outputs while preserving both the consistency of the intended edits and the integrity of the original content.

As illustrated in Figure 1, the Edit-Compatible Surrogate-Driven Privacy Protection paradigm leverages surrogates to


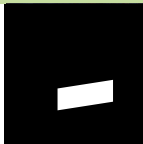




Input						Results	
Privacy Type	Edit Prompt	Image I	Mask area M	Surrogate S	Edited Surrogate S'	Recovered \hat{I}	GT I'
License plate	Turn to penciled style						

Figure 2: The sensitive category (C) is “license plate,” and the edit prompt is “Turn to penciled style.” The original image I contains the private content “SUBARU,” which is replaced by a synthetic one, “908ABD,” in the surrogate image S . However, the MLLM-edited surrogate output S' retains the synthetic plate “908ABD” rather than reflecting the original content, necessitating recovery of the surrogate output to better approximate the edited original image I' .

overcome the challenges in privacy-preserving editing. This process begins locally, where privacy-sensitive regions are replaced with synthesized surrogates containing no original private information. These surrogates are then edited by the MLLM in the cloud while the MLLM is restricted to interacting solely with the surrogate image. Subsequently, the edited surrogates serve as semantic references to guide the integration of the editing intent with the original image locally. The final result is then locally recovered as an edited version of the original image by simulating the MLLM’s editing behavior based on the transformation observed in the edited surrogate.

To provide a unified standard for this paradigm, we introduce **SPPE-Bench**. This benchmark is designed to evaluate recovery quality with respect to both MLLM editing fidelity and privacy preservation under surrogate-based protection. The dataset covers a broad range of privacy-sensitive categories and diverse editing instructions, simulating realistic scenarios. Building on our dataset, we propose Surrogate-to-Original Editable Recovery (**SOER**), a unified framework that reconstructs MLLM-style edits on original images without exposing them to the MLLM. We leverage a Diffusion Transformer (DiT) to jointly model semantic, visual, and spatial cues, guiding the generation process to faithfully reproduce the intended edits while maintaining consistency with the original visual content. Extensive evaluations on the SPPE benchmark and the InstructPix2Pix dataset demonstrate strong performance in edit fidelity, privacy preservation, and generalization. Our main contributions are summarized as follows:

- We center on an under-explored privacy problem in MLLM services, emphasizing the challenge of enabling effective image editing under surrogate-driven privacy constraints.
- We introduce **SPPE**, to the best of our knowledge, the first benchmark dataset explicitly designed to evaluate edit fidelity of utilizing MLLMs under privacy-preserving conditions.
- We propose **SOER**, a multimodal generative framework that achieves privacy-aware recovery of MLLM outputs across diverse sensitive content and editing instructions using a single model.

Related Work

Multimodal Large Language Model

Multimodal Large Language Models (MLLMs) have shown impressive capabilities in aligning and reasoning over multimodal inputs. Models like LLaVA and MiniGPT-4 (Liu et al. 2023; Zhu et al. 2023) leverage instruction tuning on top of LLaMA (Touvron et al. 2023) to enable strong performance in tasks such as VQA, grounded reasoning, and human-AI interaction. To extend MLLMs to image synthesis, GILL (Koh, Fried, and Salakhutdinov 2023) integrates LLMs with diffusion models, while SEED and SEED-2 (Ge et al. 2023a,b) propose visual tokenizers that align image and text embeddings for coherent generation. Recently, SmartEdit (Huang et al. 2024) introduces a Bidirectional Interaction Module to enhance instruction understanding and editing precision. Together, these advances highlight the shift toward generalist MLLMs that unify vision-language understanding and editing.

MLLM Privacy

Prior works on privacy-preserving learning for MLLMs include Differential Privacy (DP) (Chien et al. 2023; Yang et al. 2019; Wei et al. 2025) and inference-time protection methods like ReVision and MARRS (Mishra et al. 2025; Ates et al. 2023). While effective for general protection, these approaches fall short when tasks involve direct manipulation of visual content (e.g., image editing), as they lack concrete visual input. Surrogate-driven protection (Gafni, Wolf, and Taigman 2019; Hukkelås et al. 2023; Maximov, Elezi, and Leal-Taixé 2022; Xu et al. 2024a) addresses this by replacing sensitive regions with synthetic content, but existing methods prioritize privacy over content consistency, often resulting in inconsistent outputs. Consequently, the crucial task of recovering intended outputs after editing has been largely overlooked—a significant challenge given the diverse and flexible nature of MLLM editing scenarios.

SPPE Dataset

Dataset Overview

We propose SPPE, a new benchmark dataset specifically designed for evaluating the image editing quality of Multimodal Large Language Models (MLLMs) under surrogate-driven, privacy-preserving scenarios. Table 1 compares



Figure 3: Performance of surrogate generation. This example shows an image containing a sensitive student ID card region (leftmost panel). On the right, the top row compares our surrogate method with traditional privacy protection techniques, demonstrating superior concealment of private content while maintaining semantic coherence. The bottom row presents surrogates generated with varying protection strengths, where a higher strength indicates less influence from the original image. We select a middle strength that achieves a flexible trade-off between privacy protection and semantic fidelity.

Dataset	A.	P.	E.	R.	MLLM	Scale
VISPR						12,000
VizWiz-Priv	✓	✓				5,537
Redactions	✓					8,473
DIPA	✓					1,495
DIPA2	✓					1,304
BIV-Priv-Seg	✓					728*
Multi-P2A					✓	31,962
ReVision		✓		✓	✓	1,700
HR-VISPR		✓				10,110
Ours	✓	✓	✓(65)	✓	✓	55,696

Table 1: Benchmark Comparison. **A.**: region-level private content annotation; **P.**: availability of protected versions; **E.**: suitability for image editing tasks; **R.**: recovery requirement; **MLLM**: relevance to MLLM use cases; **Scale**: dataset size.

SPPE with existing privacy protection datasets (Orekondy, Schiele, and Fritz 2017; Xu et al. 2023; Gurari et al. 2019; Orekondy, Fritz, and Schiele 2018; Xu et al. 2024b; Tseng et al. 2025; Zhang et al. 2025; Mishra et al. 2025; Abdulaziz, D’Amicantonio, and Bondarev 2025), showing that SPPE uniquely provides large-scale coverage and complete data support across protection, editing, and recovery stages, specifically tailored for evaluating privacy-aware MLLM editing. Figure 2 illustrates a representative sample. Each instance in SPPE is represented as a 7-tuple: the sensitive category C , the edit instruction P , the original image I , the privacy mask M , the surrogate image S , the edited S' and I' obtained by applying prompt P with MLLM.

SPPE is constructed through the following four stages: (1) Data Collection, (2) Surrogate Generation, (3) Prompt Definition, (4) MLLM Editing.

Data Collection We collect original images and object-level annotations from widely used privacy datasets, including VISPR (Orekondy, Schiele, and Fritz 2017), Visual Redaction (Orekondy, Fritz, and Schiele 2018), DIPA (Xu et al. 2023), and DIPA2 (Xu et al. 2024b). A total of 36 fine-grained categories are unified into 10 high-level privacy types and further grouped into three modalities—textual, vi-

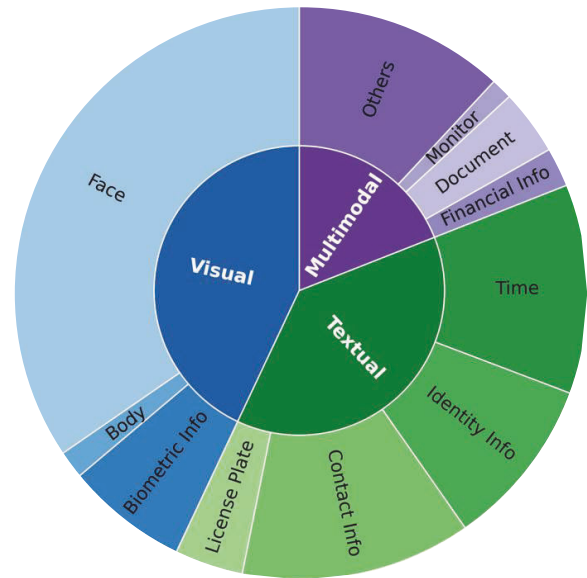


Figure 4: Distribution of privacy-related categories across three modality types (Textual, Visual, and Multimodal). The outer ring represents fine-grained categories, while the inner ring shows their grouping into higher-level modality classes.

sual, and multimodal—based on their appearance. Figure 4 illustrates the distribution, showcasing broad coverage of privacy types to support realistic and robust evaluation.

Surrogate Generation We generate privacy-preserving surrogates by inpainting masked sensitive regions using the SDXL model (Podell et al. 2024). As shown in the top right of Figure 3, this approach preserves image utility better than traditional methods. The protection strength can be adjusted by varying the influence of the original image (see bottom right); we adopt a moderate guidance scale of 0.5 to balance privacy and utility.

Prompt definition To simulate flexible real-world user operations, we defined 65 prompts covering a variety of editing types, including style transfer, local feature modification,

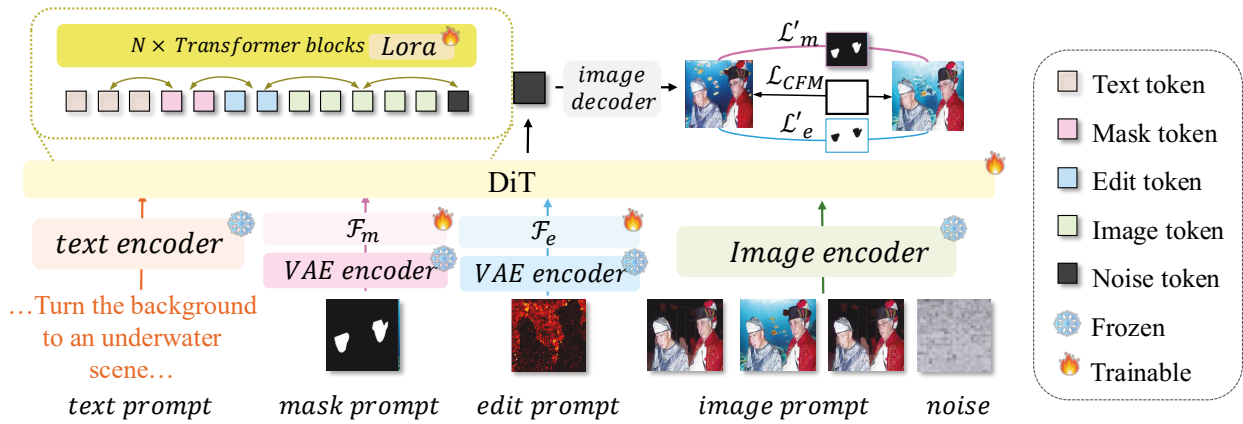


Figure 5: Overview of the proposed Surrogate-to-Original Editable Recovery (SOER). SOER processes semantic, visual, restoration, and edit cues through dedicated encoders to extract rich spatial and directional information. The resulting embeddings are combined and fed into a DiT-based transformer for multimodal interaction, enabling the generation of an edited image that faithfully reflects MLLM-driven edits while remaining consistent with the original content.

object addition or removal, etc. Some examples are shown in Table 2, and the full set of edits can be found in the appendix.

Type	Example
Style	Turn to a pencil sketch style...
Concept	Turn the person into a vampire...
Addition	Let the person wear sunglasses...
Removal	Remove the text...
Replace	Change the background to a forest...
Appearance	Make the person look older,...
...	...

Table 2: Examples of editing prompts.

MLLM editing We employ SmartEdit (Huang et al. 2024) to generate MLLM-style edits for both the original image I and its surrogate S . Two prompts are randomly selected per instance to produce (S'_1, I'_1) and (S'_2, I'_2) , where I'_1 and I'_2 serve as ground-truth references for evaluating the fidelity and accuracy of local recovery.

Surrogate-to-Original Editable Recovery

Task Definition

To address the challenge of privacy-preserving MLLM image editing, we define a new recovery task. This task aims to reconstruct the MLLM output \hat{I} locally without uploading the original image I to the MLLM. The primary objective is to ensure that the reconstructed image \hat{I} closely matches the output generated by directly applying the MLLM to the original image. Formally, given a sample as defined in the dataset section, the goal is to develop a recovery model \mathcal{R} that generates \hat{I} by :

$$\hat{I} = \mathcal{R}(I, S, S', C, M, P),$$

The goal is to minimize the difference between \hat{I} and I' .

Surrogate-to-Original Editable Recovery

We propose **SOER**, a DiT-based generative framework for privacy-aware output recovery that preserves the MLLM editing effects. Our method integrates four types of guidance—semantic, visual, restore, and edit—to generate a reconstructed MLLM output. These guidance cues enable effective control over spatial localization and precise perception of editing intent, allowing for fine-grained and controllable reconstruction that closely approximates the original MLLM-edited result.

Semantic guidance. The user’s editing intent is encoded using a pretrained text encoder applied to the instruction prompt p_{text} (e.g., “add glasses” or “change background to sunset”). This semantic embedding serves as a high-level guidance signal that directs the model to apply structural and stylistic changes aligned with the intended modification.

Visual Guidance. To support faithful and privacy-compliant reconstruction, we utilize visual references that encode both content priors and editing context. The original image I conveys detailed private visual content, while the surrogate S and its edited version S' capture the transformation behavior induced by the MLLM. These three images are processed by a shared VAE-based encoder to produce embeddings $p_I, p_S, p_{S'}$, providing the recovery model with cues about the source content and MLLM editing effects.

Restore & Edit Guidance. To enhance localization accuracy and improve disentanglement between preservation and modification, we further introduce two spatial priors. The privacy mask M , defined by differences between I and S , highlights regions where private content has been replaced and must be faithfully preserved. The edit map, computed from the difference between S and S' , identifies areas modified by the MLLM. While such information is partially encoded in the visual embeddings, these explicit region-level signals offer more precise guidance on where to recover original content and where to apply editing transformations.

Both M and the edit map are encoded using a VAE encoder \mathcal{E} and passed through guidance-specific Transformers \mathcal{F}_m and \mathcal{F}_e . Their outputs are averaged via module \mathcal{A} to obtain compact embeddings p_m and p_e :

$$p_m = \mathcal{A}(\mathcal{F}_m(\mathcal{E}(M))), \quad p_e = \mathcal{A}(\mathcal{F}_e(\mathcal{E}(|S' - S|^2))) \quad (1)$$

Multimodal Control Generation All guidance embeddings and the noisy latent representation are first projected into a shared embedding space and concatenated into a composite input sequence:

$$G = [p_e, p_m, p_{\text{text}}, p_s, p_{s'}, p_l, n_{\hat{f}}]. \quad (2)$$

Here, $n_{\hat{f}}$ represents the noisy latent token to be progressively denoised into the target image \hat{I} . The denoising process is performed by a DiT backbone enhanced with Multimodal Attention (MMA) modules, which enable dynamic interaction and information exchange among editing intent, content representation, and spatial localization cues. Formally, MMA computes:

$$Q = GW_Q, \quad K = GW_K, \quad V = GW_V,$$

$$\text{MMA}(G) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices, and d_k is the dimension of the key vectors.

Loss Computation. We optimize the local reconstruction module via a Conditional Flow Matching (CFM) loss:

$$\mathcal{L}_{\text{all}} = \mathbb{E}_{t, p_t(x|z), q(z)} \|v_\theta(x, t) - u_t(x|z)\|^2 \quad (3)$$

To jointly optimize both guidance pathways, we adopt an alternating training strategy: at each step, gradients are propagated through either the mask encoder \mathcal{F}_m or the edit encoder \mathcal{F}_e , while freezing the other. Correspondingly, we apply region-aware supervision by re-weighting \mathcal{L}_{all} based on the binary mask M :

$$\mathcal{L}_m = \mathcal{L}_{\text{all}} + \mathcal{L}_{\text{all}} \odot M, \quad \mathcal{L}_e = \mathcal{L}_{\text{all}} + \mathcal{L}_{\text{all}} \odot (1 - M) \quad (4)$$

This strategy guides the model to maintain original content consistency in sensitive regions while ensuring faithful edit reproduction elsewhere. By applying region-weighted loss and alternately training the mask and edit encoders, the model learns to balance preservation and editing, achieving stable and effective guidance.

Experiment

Datasets. We evaluate our method on both the proposed SPPE dataset and the public **InstructPix2Pix** dataset. For fair comparison, all methods (VISII (Nguyen et al. 2023), Prompt-Diffusion (Wang et al. 2023), and Edit-Transfer (Chen et al. 2025)) are trained on the SPPE training split. Evaluation is conducted on the SPPE test set and a subset of the InstructPix2Pix dataset, covering 934 unique prompts. Since InstructPix2Pix lacks sensitive region annotations, we generate full-image surrogates for its samples. While InstructPix2Pix contains synthetic images and

TEXTUAL					
	SSIM \uparrow	PSNR \uparrow	CSIM \uparrow	DirS \uparrow	DirI \uparrow
GT	1.0000	inf	1.0000	0.7571	1.0000
P-diff	0.4437	14.40	0.7031	0.5605	0.6269
VISII	0.6778	20.32	0.7567	0.5435	0.6245
E-tran	<u>0.7483</u>	<u>20.69</u>	0.8126	<u>0.6717</u>	0.7223
Ours	0.8049	23.00	0.8482	0.7139	0.7584
MULTIMODAL					
	SSIM \uparrow	PSNR \uparrow	CSIM \uparrow	DirS \uparrow	DirI \uparrow
GT	1.0000	inf	1.0000	0.5899	1.0000
P-diff	0.4400	14.19	0.7124	0.4541	0.6145
VISII	0.6669	19.58	0.7535	0.4181	0.6025
E-tran	0.7144	19.69	0.7961	0.5283	0.6867
Ours	0.7608	21.62	0.8267	0.5631	0.7195
VISUAL					
	SSIM \uparrow	PSNR \uparrow	CSIM \uparrow	DirS \uparrow	DirI \uparrow
GT	1.0000	inf	1.0000	0.7281	1.0000
P-diff	0.5350	15.24	0.7614	0.5360	0.6114
VISII	0.5715	16.54	0.7292	0.4360	0.5270
E-tran	0.7600	20.42	0.8350	0.6143	0.6828
Ours	0.8166	22.92	0.8693	0.6656	0.7274

Table 3: *Edit Consistency*. Evaluations on the SPPE dataset demonstrate that our SOER model consistently outperforms existing methods, including VISII, P-Diff, and E-Transfer. The best and second-best results are highlighted in **bold** and underlined, respectively. Arrows (\uparrow) indicate that higher values are preferable. SOER achieves best performance across all metrics, showcasing its superior ability to preserve MLLM-intended edits under privacy-preserving conditions.

prompts absent from SPPE, it offers a comprehensive benchmark for assessing model generalization to real-world scenarios where users may provide novel edit instructions.

Evaluation Metrics. We evaluate privacy-aware MLLM editing from two perspectives: *Edit Consistency* and *Source Integrity*. *Edit Consistency* measures how well the edited image \hat{I} aligns with the intended edits I' by the MLLM, using metrics such as CLIP Similarity (CLIP-Sim), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Directional Similarity (DirS/DirI) (Nguyen et al. 2023; Lai et al. 2025), which compares editing directions between image pairs like $(I \rightarrow \hat{I})$ and $(S \rightarrow S')$ or $(I \rightarrow I')$. *Source Integrity* evaluates how well \hat{I} preserves the original content by comparing it with I using CLIP-Sim, SSIM, and PSNR. Since minimizing edits can artificially inflate these metrics, Source Integrity is considered jointly with edit-related metrics for balanced evaluation. To this end, we use the corresponding metric values computed on the ground-truth edited image I' as a reference to assess how closely \hat{I} matches the expected editing outcomes.

Qualitative Results

Fig. 3 presents qualitative results across various privacy scenarios (visual, textual, and multimodal), demonstrating three key strengths of our method. First, we achieve superior *in-*

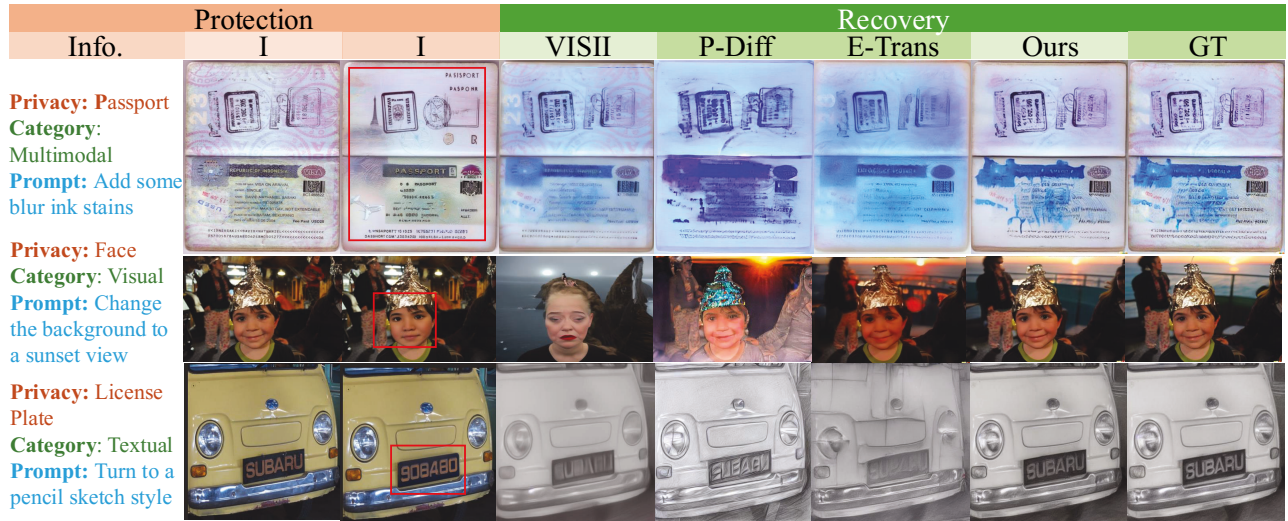


Figure 6: Qualitative comparison across textual, visual, and multimodal privacy scenarios. The examples highlight our method’s ability to accurately reproduce MLLM-driven edits, maintain semantic consistency with the original prompts, and effectively recover sensitive content in the source image.

TEXTUAL			
	SSIM	PSNR	CSIM
GT	0.627	16.48	0.651
P-diff	0.375 (-0.252)	11.42 (-5.06)	0.507 (-0.145)
VISII	0.689 (+0.062)	17.41 (+0.93)	0.680 (+0.029)
E-tran	0.615 (-0.012)	16.00 (-0.48)	0.616 (-0.035)
Ours	0.636 (+0.009)	16.41 (-0.07)	0.646 (-0.006)
MULTIMODAL			
	SSIM	PSNR	CSIM
GT	0.631	16.14	0.671
P-diff	0.386 (-0.245)	11.31 (-4.83)	0.540 (-0.131)
VISII	0.686 (+0.055)	17.33 (+1.19)	0.697 (+0.026)
E-tran	0.616 (-0.015)	15.72 (-0.42)	0.636 (-0.035)
Ours	0.637 (+0.006)	16.07 (-0.07)	0.664 (-0.006)
VISUAL			
	SSIM	PSNR	CSIM
GT	0.650	16.49	0.740
P-diff	0.457 (-0.193)	12.23 (-4.25)	0.621 (-0.119)
VISII	0.608 (-0.042)	16.33 (-0.15)	0.683 (-0.057)
E-tran	0.637 (-0.013)	16.02 (-0.47)	0.710 (-0.030)
Ours	0.656 (+0.006)	16.50 (+0.01)	0.733 (-0.007)

Table 4: *Source Integrity* Evaluations on the SPPE dataset show that our SOER model consistently achieves results closest to the ground-truth (GT) reference. The best and second-best results are marked in **bold** and underlined.

struction fidelity, as seen in the top example where our approach accurately realizes the intended modification (e.g., adding blue ink stains), in contrast to other methods that produce incomplete or imprecise edits. Second, we achieve superior *MLLM consistency*. In the middle example, although existing methods like Prompt-Diff and Edit-Transfer

successfully perform the requested edit (e.g., changing the background to a sunset view), our output aligns more faithfully with the actual MLLM’s style and details. Finally, the bottom example highlights our robust *privacy preservation*. Despite global transformations such as style transfer, our model successfully disentangles edited regions from areas that must remain consistent with the original image, thereby accurately simulating MLLM edits on the original content while retaining sensitive details (e.g., the license plate).

Quantitative Results

EDIT CONSISTENCY					
	SSIM \uparrow	PSNR \uparrow	CSIM \uparrow	DirS \uparrow	DirI \uparrow
GT	1.0000	inf	1.0000	0.2705	1.0000
P-diff	0.4111	12.2141	0.7417	0.1568	0.3699
E-tran	<u>0.5251</u>	<u>15.9315</u>	<u>0.7661</u>	<u>0.1717</u>	<u>0.3446</u>
Ours	0.5340	16.0415	0.7857	0.2077	0.3817
SOURCE INTEGRITY					
	SSIM	PSNR	CSIM		
GT	0.6914	17.7684	0.8769		
P-diff	0.4300 (-0.26)	11.7821 (-5.99)	0.7248 (-0.15)		
E-tran	0.5320 (-0.16)	15.1055 (-2.66)	0.7684 (-0.11)		
Ours	0.5413 (-0.15)	15.1643 (-2.60)	0.7783 (-0.10)		

Table 5: Quantitative comparison on the **InstructP2P** dataset. We adopt similar evaluation protocols to measure *Edit Consistency* and *Source Integrity*, assessing the model’s generalization ability on unseen images and prompts.

Note: VISII is not included in this comparison since it uses a model pretrained on InstructP2P during inference, giving it an unfair advantage in this specific evaluation.

Our method achieves the highest Edit Consistency across

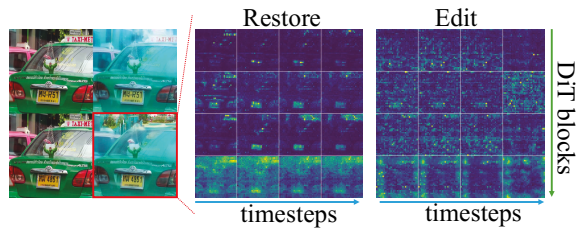


Figure 7: We visualize attention maps corresponding to the restore and edit guidance tokens, revealing distinct attention patterns—restore tokens focus on sensitive regions requiring preservation (e.g., license plates), while edit tokens highlight areas of semantic modification (background), reflecting the spatial roles of the two guidance signals.

a wide range of privacy categories and edit types, significantly outperforming baselines and consistently producing metrics that are closest to the ground truth. For instance, in the Multimodal category, our method not only surpasses all competitors with a DirS of 0.5631 but also achieves the smallest divergence from the GT value 0.5899, demonstrating faithful capturing of MLLM edit hints. Our approach also consistently yields Source Integrity scores that are closest to the ground truth, striking a strong balance between faithful editing and content preservation. For example, while methods like VISII may occasionally score highest in Source Integrity (e.g., an SSIM of 0.6891 in the Textual category compared to our 0.6358), they often exhibit poor Edit Consistency (with an SSIM of only 0.6778 against our 0.8049), suggesting a tendency toward under-editing. These results highlight the advantage of our approach in achieving privacy-aware edits that are both semantically aligned and visually coherent.

Table 5 presents results on the Instruct-p2p dataset, which requires generalization across unseen prompts and a full-image surrogate. The use of a full-image surrogate naturally introduces greater discrepancies with the original image, which in turn makes the environment more challenging and leads to a notably low ground-truth alignment (DirS of only 0.2705). Despite these challenging conditions, our approach demonstrates robust performance. Our method achieves particularly strong results on the DirS (0.2077) and DirI (0.3817) metrics, significantly outperforming competitors. The demonstrated robustness highlights our approach’s potential as a privacy-preserving editing solution that can be effectively integrated into diverse real-world applications with minimal need for costly retraining.

Ablation study

Impact of Multimodal Guidance Signals. We conduct an ablation study (Table 6) to evaluate the contribution of each guidance component. Removing the text prompt leads to significant degradation, highlighting the importance of semantic alignment with user intent. Excluding spatial cues (e.g., mask or edit prompt) also results in performance drops, confirming their role in localizing edits. We further examine the spatial influence patterns in the Visualization section.

	SSIM \uparrow	PSNR \uparrow	CSIM \uparrow	DirS \uparrow	DirI \uparrow
Ours	0.802	22.71	0.853	0.666	0.738
- mask	0.789	21.99	0.843	0.650	0.724
- diff	0.787	21.91	0.843	0.651	0.726
- text	0.714	19.11	0.814	0.631	0.707
+ CNN	0.718	19.77	0.793	0.592	0.673
+ CLS	0.770	21.70	0.827	0.620	0.702

Table 6: Ablation study on multimodal guidance components and encoding strategies. We evaluate the impact of removing (‘-’) individual guidance signals, excluding the no-visual-guidance setting due to its severe degradation. We also compare different encoders (‘+’) for mask and edit map representations, where the VAE-based encoder demonstrates superior to CNN based and CLS-based methods.

These results demonstrate the complementary nature of all guidance signals in supporting accurate generation. The no-visual-guidance setting is excluded due to performance collapse caused by the lack of reference inputs.

Representation Strategies for Spatial Guidance. We assess the impact of different representation strategies for restore and edit guidance. Substituting our transformer-based encoder with a CNN, or summarizing spatial inputs using a single CLS token leads to notable performance drops, highlighting the importance of expressive, context-aware encoders in capturing edit intent and spatial cues.

Visualization

To better understand how restore and edit guidance signals influence generation, we visualize attention maps across DiT blocks and denoising steps (Figure 7). The restore guidance exhibits focused attention on privacy-sensitive regions (e.g., license plates), facilitating accurate content preservation. In contrast, the edit guidance attends to regions aligned with the intended visual effect—for example, emphasizing a more global focus in this case, where the prompt requires transforming the scene into an underwater style.

Conclusion

This work focused on a critical yet overlooked challenge: enabling MLLM image editing under surrogate-driven privacy protection. We introduced SPPE, a novel benchmark for evaluating edit fidelity in privacy-aware settings, and proposed a multimodal generative framework, SOER, that locally recovered edited images consistent with both the original content and MLLM edits. Extensive experiments demonstrated that SOER outperformed existing methods in simulating MLLM effects while preserving private content, even in challenging scenarios. Our work established a new standard in the field and provided a practical solution for real-world MLLM applications under privacy constraints.

Acknowledgements

This work acknowledges the Hon Hai-CityU Joint Research Center and Hon Hai Research Institute for their financial and technical support.

References

- Abdulaziz, S.; D'Amicantonio, G.; and Bondarev, E. 2025. Evaluation of Human Visual Privacy Protection: A Three-Dimensional Framework and Benchmark Dataset. arXiv:2507.13981.
- Ates, H. C.; Bhargava, S.; Li, S.; Lu, J.; Maddula, S.; Moniz, J. R. A.; Nalamalapu, A. K.; Nguyen, R. H.; Ozyildirim, M.; Patel, A.; Piraviperumal, D.; Renkens, V.; Samal, A.; Tran, T.; Tseng, B.-H.; Yu, H.; Zhang, Y.; and Zou, S. 2023. MARRS: Multimodal Reference Resolution System. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, 51–58. Association for Computational Linguistics.
- Chen, L.; Mao, Q.; Gu, Y.; and Shou, M. Z. 2025. Edit Transfer: Learning Image Editing via Vision In-Context Relations. *arXiv preprint arXiv:2503.13327*.
- Chien, E.; Chen, W.-N.; Pan, C.; Li, P.; Ozgur, A.; and Milenkovic, O. 2023. Differentially private decoupled graph convolutions for multigranular topology protection. *Advances in Neural Information Processing Systems*, 36: 45381–45401.
- Gafni, O.; Wolf, L.; and Taigman, Y. 2019. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9378–9387.
- Ge, Y.; Ge, Y.; Zeng, Z.; Wang, X.; and Shan, Y. 2023a. Planting a SEED of Vision in Large Language Model. arXiv:2307.08041.
- Ge, Y.; Zhao, S.; Zeng, Z.; Ge, Y.; Li, C.; Wang, X.; and Shan, Y. 2023b. Making LLaMA SEE and Draw with SEED Tokenizer. arXiv:2310.01218.
- Gurari, D.; Li, Q.; Lin, C.; Zhao, Y.; Guo, A.; Stangl, A.; and Bigham, J. P. 2019. VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Huang, Y.; Xie, L.; Wang, X.; Yuan, Z.; Cun, X.; Ge, Y.; Zhou, J.; Dong, C.; Huang, R.; Zhang, R.; et al. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8362–8371.
- Hukkelås, H.; Smebye, M.; Mester, R.; and Lindseth, F. 2023. Realistic full-body anonymization with surface-guided GANs. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, 1430–1440.
- Koh, J. Y.; Fried, D.; and Salakhutdinov, R. 2023. Generating Images with Multimodal Language Models. *NeurIPS*.
- Lai, B.; Juefei-Xu, F.; Liu, M.; Dai, X.; Mehta, N.; Zhu, C.; Huang, Z.; Rehg, J. M.; Lee, S.; Zhang, N.; and Xiao, T. 2025. Unleashing in-context learning of autoregressive models for few-shot image manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18346–18357.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.
- Maximov, M.; Elezi, I.; and Leal-Taixé, L. 2022. Decoupling identity and visual quality for image and video anonymization. In *Proceedings of the Asian Conference on Computer Vision*, 3637–3653.
- Mishra, A.; Noh, R.; Fu, H.; Li, M.; and Kim, M. 2025. ReVision: A Dataset and Baseline VLM for Privacy-Preserving Task-Oriented Visual Instruction Rewriting. arXiv:2502.14780.
- Nguyen, T.; Li, Y.; Ojha, U.; and Lee, Y. J. 2023. Visual Instruction Inversion: Image Editing via Image Prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Orekondy, T.; Fritz, M.; and Schiele, B. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8466–8475.
- Orekondy, T.; Schiele, B.; and Fritz, M. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*, 3686–3695.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.
- Tseng, Y.; Sharma, T.; Zhang, L.; Stangl, A.; Findlater, L.; Wang, Y.; and Gurari, D. 2025. BIV-Priv-Seg: Locating Private Content in Images Taken by People With Visual Impairments. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 430–440.
- Wang, Z.; Jiang, Y.; Lu, Y.; yelong shen; He, P.; Chen, W.; Wang, Z.; and Zhou, M. 2023. In-Context Learning Unlocked for Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Wei, Q.; Li, J.; You, Z.; Zhan, Y.; Li, K.; Wu, J.; Liu, X. L. H.; Yu, Y.; Cao, B.; Xu, Y.; et al. 2025. Dual-Priv Pruning: Efficient Differential Private Fine-Tuning in Multimodal Large Language Models. *arXiv preprint arXiv:2506.07077*.
- Xu, A.; Fang, S.; Yang, H.; Hosio, S.; and Yatani, K. 2024a. Examining Human Perception of Generative Content Replacement in Image Privacy Protection. In *CHI*, 777:1–777:16.
- Xu, A.; Zhou, Z.; Miyazaki, K.; Yoshikawa, R.; Hosio, S.; and Yatani, K. 2023. DIPA: An Image Dataset with Cross-cultural Privacy Concern Annotations. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 259–266.
- Xu, A.; Zhou, Z.; Miyazaki, K.; Yoshikawa, R.; Hosio, S.; and Yatani, K. 2024b. DIPA2: An Image Dataset with Cross-cultural Privacy Perception Annotations. *Proceedings of the*

ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7(4): 1–30.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhang, J.; Cao, X.; Han, Z.; Shan, S.; and Chen, X. 2025. Multi-P²A: A Multi-perspective Benchmark on Privacy Assessment for Large Vision-Language Models. arXiv:2412.19496.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.