Unsafe2Safe: Controllable Image Anonymization for Downstream Utility

Supplementary Material

Minh Dinh SouYoung Jin
Dartmouth College
{Minh.T.Dinh.GR, SouYoung.Jin}@dartmouth.edu

S1. Dataset Construction and Preprocessing

Data usage. All images are sourced from publicly available datasets (e.g., Caltech101 [13], MIT Indoor Scenes [37], and MS-COCO [29]) and used in accordance with their respective research usage policies.

S1.1. Dataset Choice

We select MS-COCO, Caltech101, and MIT Indoor67 because they provide standardized downstream utility labels while containing diverse real-world identity and contextual cues that require whole-image anonymization beyond facial regions.

Many commonly used computer vision datasets, such as CIFAR or iNaturalist, contain limited privacy-sensitive content and therefore do not meaningfully test anonymization methods. In contrast, datasets used by prior anonymization works [22, 25] (e.g., CelebA-HQ, FFHQ) primarily evaluate re-identification rates and focus heavily on faces. While these benchmarks are suitable for measuring facial identity removal, they lack downstream task labels, making it difficult to assess the utility–privacy trade-off.

Web-scale datasets such as LAION or Flickr are highly representative of privacy-prone internet imagery and often contain caption annotations. However, their scale makes large-scale controlled anonymization experiments computationally prohibitive, particularly when multi-stage editing and evaluation are required.

We refrain from using ImageNet because it is likely included in the pretraining data of the VLMs, diffusion models, and utility backbones used in our pipeline. Such overlap could confound evaluation by introducing unintended memorization or distribution leakage. Our chosen datasets allow controlled evaluation of anonymization quality while minimizing potential pretraining bias.

S1.2. Filtering of MS-COCO for Content Preservation

Because the editing step may sometimes produce results that deviate from the original scene semantics, we apply a CLIP-based filtering step to retain only high-quality edited images. For each edited image x', we compute its CLIP similarity to the corresponding public caption and normalize it by the CLIP similarity between the original private image x and the same caption:

$$s_{\text{norm}} = \frac{\text{CLIP}(c^{\text{public}}, x')}{\text{CLIP}(c^{\text{public}}, x)}. \tag{S1}$$

This normalized score measures semantic preservation relative to the original image-caption alignment. A value close to 1 indicates that the edited image remains as semantically aligned with the public caption as the original image, whereas lower values suggest semantic degradation or unintended content changes.

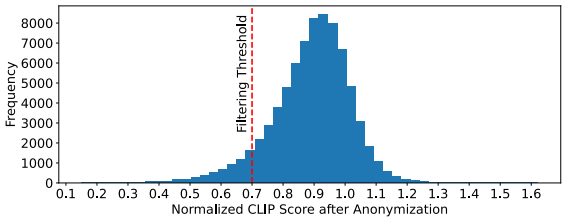

Figure S1. **Distribution of normalized CLIP scores for private images after anonymization.** Most samples cluster around 0.9, indicating strong preservation of original semantics. The red line marks the filtering threshold at 0.7, below which edited samples are discarded during editing model training.

Figure S1 shows that the majority of edited samples cluster tightly around 0.9, demonstrating that our anonymization procedure largely preserves global semantics. The small left-tail corresponds to failure cases where the edit significantly alters scene composition or weakens alignment with the public caption. Such samples can introduce noisy supervision signals during training, potentially harming both downstream utility and generative stability.

We retain only edited images with $s_{\text{norm}} > 0.7$. As shown in the distribution, this threshold falls in the lower tail and removes only a small fraction of samples—approximately 7.35% of the original *train2014* split (reducing it from 51,401 to 47,623). As a result, semantically degraded edits are excluded while preserving the vast majority of anonymized data. This filtering step improves the consistency of the training signal without meaningfully reducing dataset diversity, thereby balancing semantic fidelity and robustness in the anonymized training set. Notably, the resulting 47,623 anonymized image pairs constitute one of the largest publicly constructed before–and–after datasets for image editing [53], and, to our knowledge, the largest

that is explicitly privacy-aware.

S2. Implementation Details

S2.1. Language models

Through out all experiments, we used InternVL2.5 [8] as the VLM and Qwen3-4B with non-thinking mode (Qwen3-4B-Instruct-2507) [52] as the LLM. This choice is mainly empiracal towards a fast but robust model. Once obtaining the answers from the model, we automatically parse the sections following the structure in the prompt.

S2.2. Diffusion model training and generating

For **FreePrompt** [30], following the official implementation, we used an empty string as the source prompt and set the `SELF_REPLACE_STEPS` ratio to 0.4 to encourage substantial appearance changes. Images were generated at a resolution of 512×512 using 50 diffusion steps with a guidance scale of 7.5.

For **FlowEdit** [24], we adopted the SD3 backbone with default hyperparameters provided in the official release. Across all FlowEdit-based experiments, we used c^{priv} as the caption describing the source image.

For **InstructPix2Pix** [7], we trained on 4 GPUs with a batch size of 64 and a learning rate of 1×10^{-5}. Training was conducted for 200 epochs with gradient accumulation to an effective batch size of 256. We initialized all modules with MagicBrush [53] weights and updated only the UNet parameters. Following the original pipeline, all training images were resized to 256×256, while inference was performed at 512×512 using classifier-free guidance scales of 1.5 for the image and 7.5 for the text prompt, with 100 denoising steps.

We used an identical training and inference setup for the modified **InstructPix2Pix** equipped with **Safe Attention**. For initialization, the MagicBrush cross-attention weights were copied into the query, key, value, and output projection layers of the Safe Attention module to ensure compatibility and stable convergence.

For **OminiControl** [44], we adopted the *subject* configuration with a batch size of 4. We used a dummy positional offset of $(0, 0)$ to disable spatial displacement and trained for 12,000 intervals directly from the FLUX.1 dev [26] checkpoint. To the best of our knowledge, this makes our work among the first to introduce a privacy-aware, FLUX-based image editing model.

For **DeepPrivacy2** [22] and **FaceAnon** [25], we directly applied the default implementations released by the authors to our dataset without modification.

S2.3. Evaluation metrics for Stage 1

VISPR attribute grouping. VISPR [34] provides privacy annotations using attribute identifiers a_i. In our evaluation, privacy inspection is treated as a binary classification task: an image is labeled *safe* if it has attribute `a0_safe`, and *unsafe* otherwise.

To analyze performance under specific privacy risks, we further evaluate several attribute groups:

- **Face:** `a9`, `a10` (complete and partial face)
- **Health Indicators:** `a39`, `a41`, `a43` (physical disability, injury, medicine)
- **Vehicles:** `a102`, `a103`, `a104` (vehicle ownership, license plate complete/partial)
- **Personal Opinion:** `a61`, `a62` (general and political opinions)

An image is considered privacy-sensitive for a given group if any attribute in the corresponding set is present.

Evaluation protocol. For each image, the VLM detector predicts whether privacy-sensitive content is present under the specified criterion. Predictions are compared against VISPR annotations, and we report recall, precision, and F1 score. Recall is emphasized because missed detections would allow sensitive images to enter the training pipeline. All results are computed on the VISPR test split.

S2.4. Evaluation metrics for Stage 2

S2.4.1. Cheating Scores

To measure unintended preservation of perceptual cues from the original image, we compute the Learned Perceptual Image Patch Similarity (LPIPS) [55] with a VGG-16 backbone [42]. Higher LPIPS indicates stronger deviation from the source image, suggesting that the anonymization process does not simply reconstruct or copy identity-related features.

S2.4.2. Privacy Scores

Face similarity (**FaceSim**) is computed using the Antevelop-v2 face encoder [10]. For each detected face in the original image, we compute cosine similarity to its nearest counterpart in the anonymized image and retain only the closest match to avoid bias from unmatched pairs.

To evaluate textual leakage (**TextSim**), we use InternVL2.5-8B [8] to extract text from anonymized images and compute the token-set similarity using the `rapidfuzz` package. Lower similarity indicates more effective removal or distortion of identifiable textual content.

For demographic analysis (**Race Entropy**), the same VLM predicts demographic attributes (White, Black, Asian, Hispanic, Other), from which we compute the race entropy metric described in Section 4.3.

To obtain the **VLMScore**, we employ InternVL3.5-8B [48] as a judge model. The raw and anonymized images are jointly provided to the VLM with a structured prompt asking it to assign a score from 0–100 reflecting how effec-

Table S1. Summary of dataset statistics, utility accuracy, and privacy-leakage indicators for the raw and safe subsets. The table reports training/validation sample counts, top-1 accuracy, and the number of detected text, faces, and race attributes.

	#Samples		Utility	Privacy Instances		
Split	**Train**	**Val**	**Acc@1**	**Text**	**Faces**	**Races**
Caltech101 [13]						
Original	2000	980	94.27	251	82	71
Safe subset	1552	756	83.49	111	24	0
MIT Indoor67 [37]						
Original	3991	1317	83.88	551	2184	1332
Safe subset	1605	512	51.12	84	78	1

tively privacy-sensitive attributes have been removed while preserving scene semantics.

S2.4.3. Utility Score

Classification. For classification experiments, we adopt Masked Autoencoders (ImageMAE) [19] as the backbone and fine-tune a randomly initialized linear classification head. All models are initialized from ImageNet-pretrained weights. Training uses batch size 64, learning rate 5×10^{-4}, gradient accumulation of 4 (effective batch size 256), and 100 epochs. Data augmentation follows the ImageMAE setup using RandAugment with parameters $(n = 2, m = 9, mstd = 0.5)$.

Image Captioning. To evaluate semantic preservation in generative tasks, we fine-tune BLIP-2 [28] on the filtered MS-COCO dataset using the human-annotated captions. Caption quality is evaluated using BLEU-4 and CIDEr.

Visual Question Answering. For VQA, we fine-tune a Qwen3VL-2B [52] model on question–answer pairs from the OK-VQA dataset [32] and report answer accuracy.

In all tasks, model selection is performed on the anonymized validation set constructed in the same manner as the anonymized training set. Final performance is reported on the *original* test sets to measure preservation of task-relevant semantics under the original data distribution. All experiments are conducted on NVIDIA A100 GPUs.

S3. Evaluation of Stage 1: Privacy Inspection

S3.1. Is Image Privacy Anonymization Necessary?

Given the images categorized into safe and unsafe partitions, we need to perform a robust anonymization pipeline on the unsafe images while keep the safe images intact. We evaluate the extreme setting in which *only* images flagged as safe by the VLM are used for training. This subset inevitably contains some false positives (images incorrectly flagged as safe) which allows us to assess how well the model performs without any anonymization applied to unsafe images.

Table S1 summarizes the key statistics. As expected, compared to the model trained on the full (original) dataset, training solely on the safe subset leads to a substantial drop in downstream utility due to the significant reduction in data volume and diversity. The remaining samples are highly sanitized and lack many of the visual cues needed for effective classification, resulting in noticeably weaker accuracy.

However, the privacy-leakage signals are correspondingly minimal: the VLM detects very few readable texts, faces, or identifiable racial attributes in this safe-only set. This confirms that the privacy detector is conservative and effective as most privacy-sensitive images are successfully excluded. At the same time, the sharp utility degradation highlights the necessity of anonymizing unsafe images rather than discarding them, which helps the model to recover both dataset scale and semantic richness while maintaining strong privacy guarantees.

Table S2. **Alignment to MS-COCO annotations of different text priors produced by Stage 1 under the FLEUR [27] metric** (higher is better). The public caption c^{pub} maintains FLEUR scores close to the private caption c^{priv}, indicating strong preservation of global scene semantics despite removing sensitive details. The LLM prior , while lower due to the introduction of additional synthesized attributes, still retains meaningful semantic alignment.

Text Prior	c^{priv}	c^{pub}	LLM
FLEUR (↑)	80.45	78.93	58.27

S3.2. Are Captions Generated by Our Pipeline Helpful for Utility and Quality Preservation?

We provide an evaluation of the captions produced in Stage 1. Although modern VLMs are generally strong captioners, it is important to quantify how well the generated captions align with the underlying visual content, especially when private details are removed or rewritten. To assess caption fidelity, we use the FLEUR benchmark [27], which measures consistency between the generated caption, the image, and the five human-annotated reference captions from MS-COCO [29]. We use the val2014 split, which is also our test set, to obtain the quality score for c^{priv}, c^{pub}, and LLM.

As shown in Table S2, the public caption c^{pub} exhibits only a modest decrease in FLEUR compared to the private caption c^{priv}. This comparable score indicates that the public captions still capture the essential scene semantics that humans perceive while successfully omitting privacy-sensitive content. Notably, the LLM-composed captions also maintain reasonably high alignment despite introducing synthetic, privacy-safe attributes, demonstrating that our

Table S3. **Extend utility and privacy comparison for different generative backbones.** Our Unsafe2Safe, which uses [24] as the underlying image editor, achieves comparable performance on downstream tasks while successfully anonymizing privacy-sensitive information, unlike other anonymization models. The **best** value (yellow) and the **second-best** value (blue) are highlighted per column. † denotes the model was finetuned with our dataset.

Model	Text Prior	Utility Score		Privacy Score					
		Accuracy (↑)		FaceSim (↓)		TextSim (↓)		Race Entropy (↑)	
		Cal101	Indoor	Cal101	Indoor	Cal101	Indoor	Cal101	Indoor
Raw Images	–	94.277	83.881	1.0000	1.0000	1.0000	1.0000	0.4384	0.7443
Unsafe2Safe (FlowEdit [24])	c_{private}	94.334	79.925	0.4378	0.2666	0.6611	0.4210	0.5552	0.7399
	class	93.857	80.448	0.4965	0.3743	**0.4856**	**0.2077**	0.4051	0.6508
	c_{public}	94.487	80.746	0.4881	0.2883	0.5395	0.2896	0.6409	0.7208
	c_{edit}	94.792	77.090	0.3658	**0.2077**	0.5238	0.2393	0.7646	0.7589
	LLM (c_{edit}, c_{public})	92.884	80.746	**0.3428**	0.2294	**0.4881**	**0.2119**	**0.8751**	0.7643
Unsafe2Safe (FreePrompt [30])	c_{private}	**94.926**	81.567	0.5026	0.2764	0.6382	0.2757	0.5421	0.6963
	class	94.506	79.254	0.5474	0.2966	0.5651	0.2314	0.4991	0.5960
	c_{public}	**94.849**	82.537	0.5085	0.2811	0.5949	0.2423	0.5790	0.6899
	c_{edit}	93.857	78.881	0.3693	0.2165	0.5672	0.2361	0.7516	**0.8276**
	LLM (c_{edit}, c_{public})	94.105	80.522	0.4539	0.2159	0.5580	0.2217	0.7967	0.7806
Unsafe2Safe (OminiControl [44])†	c_{edit}	94.506	80.000	**0.3585**	**0.1925**	0.6132	0.3350	**0.8425**	**0.8058**
DeepPrivacy2 [22]	–	94.601	**84.030**	0.3921	0.3547	0.9569	0.8653	0.7315	0.7547
FaceAnonSimple [25]	–	**94.849**	**84.030**	0.4586	0.5045	0.9355	0.7701	0.6091	0.7407

edit-instruction generation preserves global scene meaning even when enriching the caption with additional identity-neutral details.

S4. Additional Results of Stage 2: Safe Image Generation

S4.1. Quantitative resuls

In Table S3, we report the performance of our pipeline when leveraging FreePrompt [30] and OminiControl [44] as the editing diffusion models. These results strengthen our conclusion that the **Unsafe2Safe** provides a robust framework not only for obtaining privacy-preserving edit instructions, but also for curating an effective dataset for teaching a diffusion model to perform anonymization.

S4.2. Analysis of attention maps from SafeAttention

We further analyze the resulting attention maps to visualize how **SafeAttention** affects the editing behavior. Following Liu et al. [30], we compute the averaged attention maps over all diffusion steps for each of the 16 transformer layers in the UNet. For Safe Cross Attention, we additionally separate the attention maps into the components corresponding to c^{pub} and c^{edit}, allowing us to examine how each text source influences the model.

Figure S2 shows the attention maps for the Cross Attention and Safe Cross Attention modules at the 13$^{\text{th}}$ transformer layer (counting from 1). As illustrated, in vanilla InstructPix2Pix [7], attention is spread diffusely across the entire image, causing edits to leak into regions that should remain unchanged. In contrast, Safe Cross Attention produces a clear separation when we inspect the maps per token group: public-caption tokens attend primarily to stable background regions and task-relevant objects, while edit-instruction tokens concentrate their attention on the sensitive areas identified for anonymization. This structured attention pattern demonstrates that Safe Cross Attention provides a more controlled and interpretable mechanism for privacy-preserving editing, facilitating denoising that is both more selective and more faithful to the intended anonymization behavior.

S4.3. Does Our Method's Controllability Promote Fairness?

Throughout the aforementioned experiments, the LLM was free to propose any identity attributes. To assess demographic controllability, we introduced a simple intervention: we constructed a list of racial groups, and for each image, we uniformly sampled one race and asked the LLM to integrate it into the edit ideas. The racial groups include `White`, `Black`, `East Asian`, `South Asian`, `Southeast Asian`, `Middle Eastern/North African`, `Indigenous/Pacific Islander`, and `Hispanic/Latino`. We use demographic predictions as a proxy for diversity. Figure S3 presents an example under the sampled label `Indigenous/Pacific Islander`. After applying our pipeline, the anonymized output reconstructs the scene

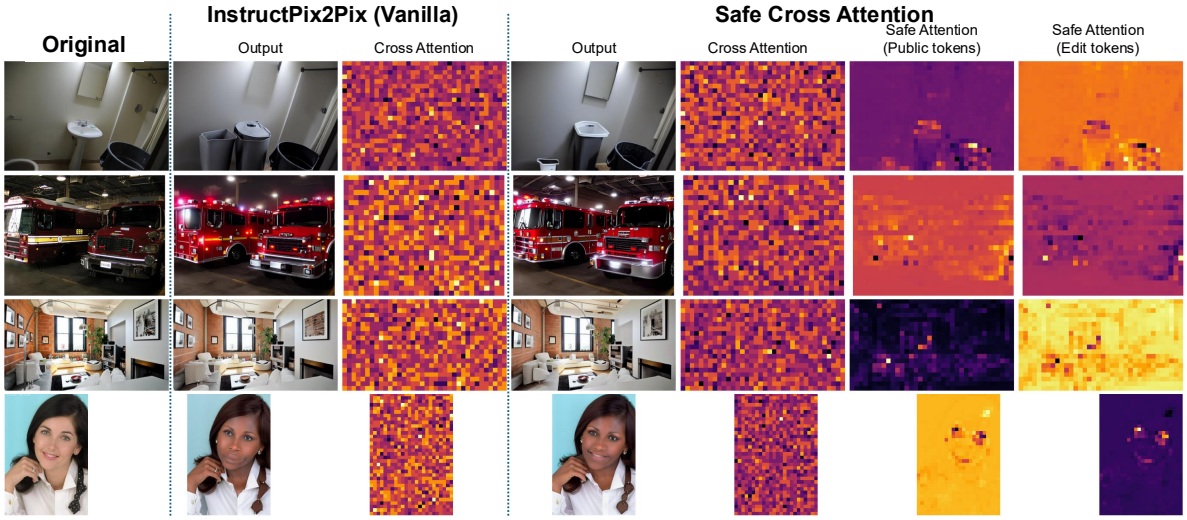

Figure S2. Attention maps at the 13th transformer block in the UNet. From left to right: (1) the original private image; (2) the anonymized output produced by vanilla InstructPix2Pix along with its standard cross-attention map; (3) the anonymized output produced by our Safe Cross Attention model, shown with its vanilla cross-attention map, the Safe Attention map corresponding to public-caption tokens, and the Safe Attention map corresponding to edit-instruction tokens.

Figure S3. Example of demographic-controlled anonymization under the *randomly* sampled "Indigenous / Pacific Islander" condition.

with entirely new identities whose appearance, such as skin tone, hairstyle, and traditional clothing, aligns with the sampled demographic category. At the same time, the global scene geometry and activity are preserved: the individuals are still gathered around the same cake, positioned in the same layout, and engaged in the same collective action. The LLM-generated edit instruction also faithfully reflects the sampled demographic, producing a coherent description that guides the editor toward culturally consistent attributes while removing sensitive cues such as uniforms, text, or identifiable faces. This intervention demonstrates that our framework not only anonymizes identity but also provides fine-grained control over demographic attributes simply via text condition when explicitly requested.

S4.4. Qualitative results

How Are Images Anonymized Differently Across Text Priors? Figure S4 shows the results when using different text priors to the FlowEdit backbone [24]. The result for c^{edit} is omitted because the text prior is not suitable for the model, which expects a description of the target scene rather than edit instructions.

We observe that existing face anonymization frameworks fail to properly address non-human privacy concerns and instead prioritize modifying the entire face or body, which can sometimes be missed. In contrast, our method applies editing to the whole image and preserves only the information necessary for downstream tasks.

It is important to note that the priors c^{priv} and c^{class} represent unprocessed information that is not fully derived from our Stage 1 and is typically available in common image datasets, such as human-annotated captions or class labels. These priors are not recommended in our pipeline and tend to either fail to protect sensitive attributes or discard valuable diversity in the images.

In comparison, safe text prompts including c^{pub}, c^{edit}, and LLM are much more suitable for effective downstream learning. This improvement is due to their alignment with

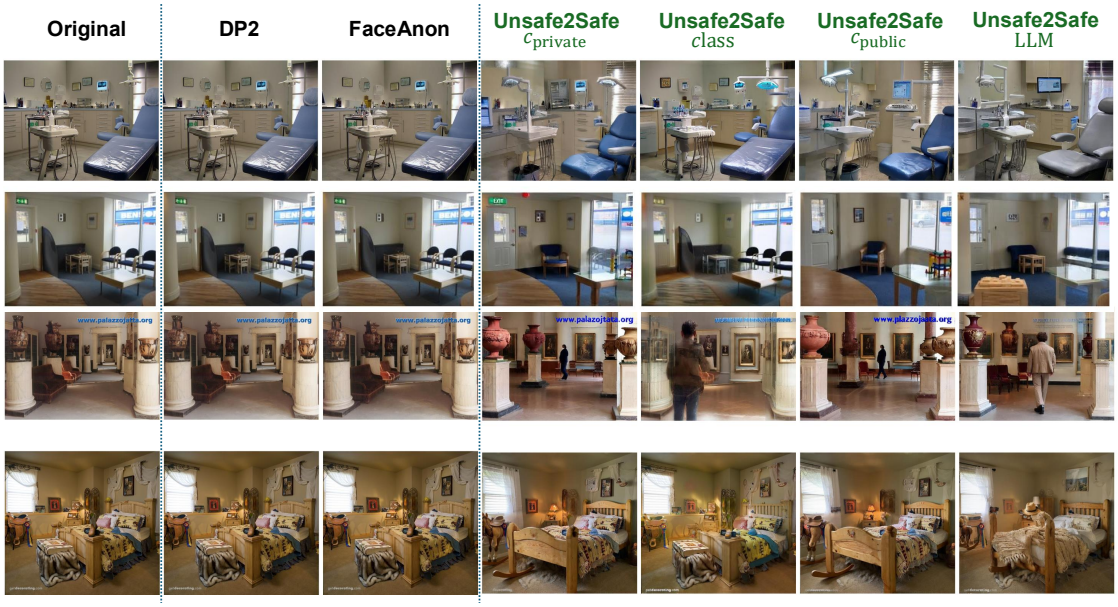

Figure S4. Qualitative results of our method with different text priors and existing face anonymizers. Our results were generated by the FlowEdit [24] model.

Figure S5. Qualitative results of our method with different generative backbones and existing face anonymizers. FlowEdit [24] takes LLM as the target text prior, while other models take c^{edit} as the text prior.

the original content as well as their ability to preserve fine-grained details and structural dynamics.

How Are Images Anonymized Differently Across Generative Backbones? We also show the quality of anonymization across different backbone diffusion models.

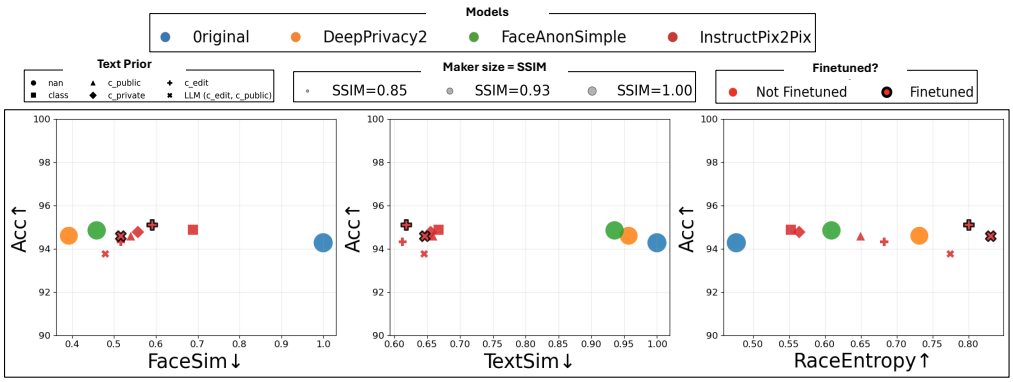

Figure S6. Multi-dimensional evaluation of privacy–utility trade-offs across anonymization models with the Caltech101 dataset [13]. Each subplot visualizes one privacy metric (FaceSim, TextSim, or RaceEntropy) against downstream utility (Acc@1). Colors denote different editing backbones, and marker shapes represent the text priors used during generation. Marker size is proportional to the SSIM *cheating* score, and a black outline indicates models fine-tuned on our anonymization dataset. This visualization highlights how fine-tuning and text conditioning jointly influence privacy protection, realism, and task performance.

Figure S5 shows outputs of all 4 diffusion models with the most compatible text prior as suggested by their authors: FreePrompt with c^{edit}, finetuned InstructPix2Pix with c^{edit}, FLUX-based OminiControl with c^{edit}, and FlowEdit with the target caption LLM. Regardless of the generator choice, our Unsafe2Safe generates images that highly align to the original content while revealing little leakage.

S5. Visualized Trade-Off Among All Evaluation Dimensions

Figure S6 summarizes the quantitative evaluation across multiple dimensions. Each subplot uses a different *privacy* indicator on the x-axis (Face Similarity, Text Similarity, or Race Entropy) and downstream *utility* (top-1 accuracy) on the y-axis. Marker size reflects the *cheating* structural similarity (SSIM), while color, shape, and outline respectively encode the generative backbone, the text prior used during editing, and whether the model was fine-tuned.

This consolidated visualization makes the privacy-utility-fidelity trade-off directly interpretable. Face-only anonymization baselines perform well on face similarity and achieve slightly higher classification accuracy, but they fail to remove low-level cues, textual information, and demographic signals, resulting in substantially weaker overall privacy protection. In contrast, as revealed by the scatterplots, our editing models generally occupy *favorable* regions of the trade-off space: they reduce identifiable content while maintaining competitive utility and without relying on structural similarity.

S6. Prompts to Language Models

S6.1. Prompts used in Stage 1

Figures S7, S8, S9, S10, and S11 illustrate the prompting components that shape how Stage 1 decomposes privacy reasoning into a sequence of explicit and controllable text-generation steps. Figure S7 defines the privacy criteria used throughout Stage 1 by summarizing VISPR [34]'s 67 attributes into nine interpretable categories.

Figure S8 shows the prompt used to obtain the `PRIVACY_FLAG`. This prompt is intentionally conservative, defaulting to `TRUE` under any ambiguity, to minimize false negatives, since any missed detection would allow sensitive content to pass downstream. The expected output is intentionally short and "explanation-free," facilitating fast, large-scale screening across all images.

Figure S9 presents the structured captioning prompt, which serves as the backbone of Stage 1. Although the `PRIVACY_REVIEW` is not directly used in later steps, it forces the VLM to explicitly enumerate privacy-relevant elements, thereby making omissions auditable and ensuring that the subsequent private and public captions are grounded in a clear semantic separation.

Figure S10 provides the prompt used by the LLM to generate edit instructions. It requires attribute-level rewriting (e.g., gender, hair, clothing, body shape, cultural markers) while prohibiting repetition of unchanged details. Importantly, the image itself is not provided to the LLM at this stage; the model must reason solely from the privacy-preserving caption.

Finally, Figure S11 shows the prompt used to merge the public caption and the edit instruction into a single compact description. Despite its simplicity, this step addresses two practical constraints: (i) diffusion editors often impose strict

token budgets, and (ii) the merged caption enables seamless integration of available ground-truth labels for preserving task-relevant semantics during editing.

S6.2. Prompts used in evaluation

We report here the prompts used for all VLM-based evaluation components. These cover four tasks: (1) privacy flagging under a restricted set of criteria, (2) extraction of readable text from the image, (3) detection of demographic attributes, and (4) judement of anonymization quality. The corresponding prompts are shown in Figure S12, Figure S13, Figure S14, and Figure S15, respectively.

S7. Deployment

We provide the implementation for a user-friendly Graphical User Interface (GUI) of our pipeline via Gradio [2]. The interface demonstrates how an uploaded image is automatically processed by our VLM- and LLM-based inspectors to produce safe captions and an anonymized proxy generated by the selected editing model.

To promote accessibility and reproducibility, we plan to publicly release this interface to the research community.

```
- **People**, faces (full or partial), visible skin in private contexts, tattoos, or emotional expressions
- **Appearance cues**: clothing, accessories, or features revealing gender, culture, religion, status, or roles
- **Health indicators**: casts, wheelchairs, bandages, medications, disabilities, or visible illness/injury
- **Texts or symbols**: signs, receipts, documents, screens, handwriting, license plates, usernames, phone numbers
- **Locations or timestamps**: homes, offices, bedrooms, addresses, classrooms, GPS data, clocks, date/time stamps
- **Personal interactions**: small groups, families, friends, teams, protests, celebrations, or close relationships
- **Professional/school settings**: uniforms, tools, name tags, diplomas, workspaces, formal events
- **Cultural/political/religious identifiers**: flags, slogans, attire, ceremonies, posters
- **Private belongings or behaviors**: hobbies, photos, mail, social media, travel documents, receipts
```

Figure S7. Privacy-related visual criteria used in the Stage 1 of our pipeline. These criteria are summarized from the set of 67 criteria introduced by VISPR [34].

```
SECTION: PRIVACY_FLAG
Determine whether the image contains **any content that could disclose private, sensitive, or
personally identifying information**. Do not include explanations, justification, or extra text.
---
Set: **PRIVACY_FLAG: TRUE**
if the image contains *even partially visible*, *blurred*, or *contextually implied* content from
the list below. If uncertain, ambiguous, low-resolution, or partially occluded|**default to TRUE**.
---
Flag the image if you observe or reasonably infer the presence of:
[CRITERIA]
---
**Treat these ambiguous cases as DEFINITE PRIVACY_FLAG: TRUE**:
- Partial faces in mirrors or reflections
- Blurred or occluded name tags, signs, or screens
- Visible groups of people even without faces
- Partially legible receipts or handwritten notes
- Niche personal objects (e.g. medals, family photos, memorabilia)
```

Figure S8. Assistant prompt used for the privacy filtering step in **Stage 1**. When a private issue in any image is detected, the VLM sets PRIVACY_FLAG to TRUE, even when the issues were partially visible, contextually implied, ambiguous, or low-resolution cases.

```
Your task is to analyze the image and generate **structured, privacy-aware captions** in 3 sections:
- `PRIVACY_REVIEW`: Identify sensitive or private elements
- `PRIVATE_CAPTION`: Full factual visual description
- `PUBLIC_CAPTION`: Safe, anonymized caption suitable for generative training
---
SECTION: PRIVACY_REVIEW
List up to 10 elements in the image that pose a privacy risk. For each, describe what is visible
and why it is sensitive. Use this format:
- ITEM: [Short factual phrase of the visible element]
- REASON: [Why it is privacy-sensitive, based on the list below]
Consider the following privacy risk categories:
[CRITERIA]
Example:
- ITEM: A woman hugging a child in a decorated living room
- REASON: Relationship, identity, and private setting
---
SECTION: PRIVATE_CAPTION
Provide a richly detailed narrative ( 50 tokens) of everything visible in the image, even if
privacy-sensitive. Include:
- People, animals, objects, or vehicles and their appearances
- Visible logos, brand names, or text
- The full environment and any identifying or contextual features
---
SECTION: PUBLIC_CAPTION
Write a **safe and fluent caption** ( 50 tokens) that:
- Rephrases or removes all sensitive elements listed in PRIVACY_REVIEW.
- Do NOT include identity, brand names, text, or logos
- Focuses on scenery, layout, lighting, objects, and anonymous presence (e.g., \a person" is allowed
if not identifiable)
- Use fluent, image-grounded language. Do **not** censor awkwardly or produce incomplete fragments.
The caption should stand alone.
```

Figure S9. User prompts provided to the VLM for generating structured, privacy-aware captions of private images. The prompt consists of three sections: PRIVACY_REVIEW, which identifies potentially sensitive elements; PRIVATE_CAPTION, which offers a detailed caption of the entire scene; and PUBLIC_CAPTION, which produces a concise, privacy-compliant caption suitable for use as guidance for diffusion models.

```
You are given a caption that describes a real-world photo:
Caption: public_caption
Your task is to propose realistic identity modifications while the content of the scene (layout, objects,
actions) must remain exactly the same.
These descriptions will be applied as editing instructions for a Stable Diffusion model.
- Make the modifications significant, specific, and easy to generate by a Stable Diffusion.
- Change every person and every object with privacy risks.
- For each person, include at least 3 specific changes regarding gender, hair, body shape, skin
tone, clothing, or culture.
- Neutralize objects or background cues with privacy risks with plain or generic alternatives.
The changes may include:
[CRITERIA]
Rules:
- Preserve the current object appearence, layout, or actions.
- Output only the modifications, not rewrite of the caption.
- Do NOT repeat the unchanged details from the original caption.
- Keep it realistic, specific, and  45 words continuous, natural phrase fragment
```

Figure S10. Prompt to the LLM for generating edit ideas as instructions conditioned on the public caption. The model proposes identity-neutral but realistic attribute modifications, with a focus on faces, while preserving layout and scene semantics.

```
You are given a caption describing a real-world
photo and an edit instruction.
Original caption: public_caption
Edit instruction: edit_caption

Combine them into a single, natural-sounding
↪  caption
that reflects ALL edits while preserving the
unchanged parts. Keep the result concise (under
50 words). After combining, the caption has to
describe an image of class_name.
```

Figure S11. Prompt for merging the public caption and edit instruction into a single compact description compatible with token-limited diffusion editors. When a class name or ground-truth annotation is available, users can simply include it into the prompt for a task-relevant caption.

```
SECTION: PRIVACY_FLAG
Determine whether the image contains any instance
of **criteria**. Respond with only one line:
**PRIVACY_FLAG: TRUE** or **PRIVACY_FLAG: FALSE**
Do not include explanations or additional text.
---
**Set PRIVACY_FLAG: TRUE** if the image clearly,
partially, or contextually depicts any CRITERIA_NAME.
If the presence of criteria is uncertain, ambiguous,
low-resolution, or partially occluded, **default to
TRUE**.
**Set PRIVACY_FLAG: FALSE** if the image does not
contain criteria, or if any privacy concern is
unrelated to criteria.
---
Flag the image as TRUE if you can observe or
reasonably infer any of the following:
CRITERIA_DESCRIPTION
```

Figure S12. Prompt used to ablate the flagging step with custom, less-inclusive privacy criteria.

```
You are an image analysis assistant. Analyze the
provided image and extract all visible text that
appears anywhere in the scene (including logos,
signs,  menus, or labels). If no readable text is
visible, output exactly: NO_TEXT

Otherwise, respond strictly in the following format:
### TEXT
<verbatim text as it appears, preserving
capitalization, punctuation, and line breaks>

Rules:
- Include only the text visible in the image - no
descriptions or interpretations.
- Preserve exact wording, capitalization,
and line breaks.
- If part of the text is unclear, replace
it with [unclear].
- Output exactly one block following the format
above.
- If no text is visible, output only "NO_TEXT".
```

Figure S13. User prompt used for detecting text in the image.

```
You are an image analysis assistant. Analyze the
provided image and identify all apparent genders
and races  visible in the scene. If no clearly
visible person (no recognizable face or body)
is present, output exactly: NO_HUMAN
Otherwise, respond strictly in the following format:
### GENDER
<male and/or female>
### RACE
<list of races: White, Black, Asian, Hispanic,
Middle Eastern, Indian, Other>
Rules:
- Use only the categories provided.
- Include all that clearly appear, separated by commas.
- Do NOT include "unknown" or any explanation.
- Output exactly one block following the format above.
- If no human is visible, output only "NO_HUMAN".
```

Figure S14. User prompt used for detecting demographical attributes (gender and race) in the image.

```
System Prompt: You are a vision-language evaluator for privacy-preserving image anonymization. You must
↪  compare the left and right images and judge anonymization strictly based on visual differences.
↪  Assign a score from 0 to 100 based on how much that difference reduces the ability to identify or
↪  link specific people, objects, or attributesfrom the LEFT image to the RIGHT image. Do not infer or
↪  assume sensitive information beyond what is visually identifiable.

User Prompt:
You are given two images shown side by side:
- Left image: the original image
- Right image: the edited / anonymized image
Your task is to assign an anonymization score that measures how much the RIGHT image reduces the
↪  ability to link the LEFT image to the same specific identities or personal specifics.
Privacy-sensitive information includes (but is not limited to):
[CRITERIA]
You MUST follow this decision process:
STEP 1 | Instance identity check:
Decide whether the LEFT and RIGHT images depict the same characterstics of privacy-senstive
↪  information.
If all of them they appear to be the same instance with no meaningful visual change, the score will be
↪  0.
- If people, faces, bodies, clothing, or group composition differ, treat the images as DIFFERENT
↪  instances.
- The same layout, camera angle, or furniture arrangement does NOT imply the same instance.
STEP 2 | Unlinkability strength:
If the RIGHT image shows a different instance, identity, or attributes, score how strongly this change
↪  prevents linking the LEFT image to the same specific details. The general scene, activity, or
↪  layout may remain similar.
Preserving scene structure, pose, or activity does NOT reduce the score as long as identity or instance
↪  linkability is broken.
---
### Scoring (0{100)
- **0**: Same identities or same instance (no anonymization).
- **20{40**: Weak unlinkability (small but real instance or attribute change).
- **40{70**: Moderate unlinkability (multiple identity-related cues changed).
- **70{90**: Strong unlinkability (identities or specifics are hard to link).
- **90{100**: Near-complete unlinkability of identities or personal specifics.
High scores are allowed even when scene/layout/activity is preserved, if identities or spersonal
↪  specifics are replaced or neutralized.
Reward evidence of:
- people/face identity replacement or strong masking (blur/occlusion/distortion)
- removal/replacement of tattoos, distinctive clothing/accessories, name tags
- removal/replacement of readable text (documents, screens, plates, usernames)
- replacement of distinctive personal objects or environment instance details
---
Rules:
- Give score 0 for identical pair.
- Give score 100 for totally different image pair.
- Do NOT penalize preservation of scene structure or activity.
- Ignore image quality, realism, or aesthetics.
- Do not infer sensitive information that is not clearly visible.
Output:
ANONYMIZATION_SCORE: <integer 0-100>
```

Figure S15. Prompt to the VLM-based judge for evaluating anonymization results. If no meaningful anonymization is observed, the judge assigns a score of 0; otherwise, it follows the defined rubics to assess the extent to which instance-specific identity attributes have been modified.