TrustCLIP: Learning Private Visual Features via Adversarial Reconstruction

Nikos Athanasiou[1,2⋆], Ilya A. Petrov[1,3⋆], Angela Yao[1,4], Shugao Ma[1], Eric Sauser[1], Edoardo Remelli[1], Shreyas Hampali[1], Johannes Schönberger[1], Fadime Sener[1], and Bugra Tekin[1]

[1] Meta, Zürich, Switzerland
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3] University of Tübingen, Germany
[4] National University of Singapore, Singapore

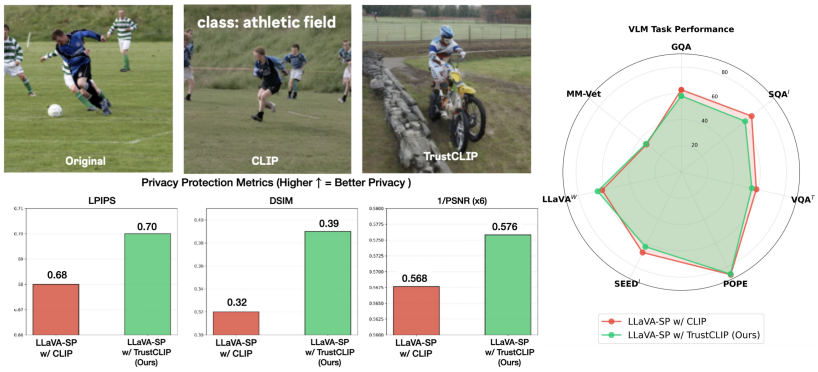

Fig. 1: TrustCLIP ensures privacy for visual understanding tasks while preserving task utility. (top left) Reconstructions from CLIP features reveal detailed content, whereas TrustCLIP features prevent meaningful recovery of the original image (while still maintaining class semantics *i.e.* 'semantic field'). (right) Despite the privacy enhancement, LLaVA-SP using TrustCLIP maintains competitive performance compared to the unprotected baseline across standard VLM benchmarks. (bottom left) Privacy metrics further confirm that TrustCLIP produces substantially less reconstructible features.

Abstract. Vision and vision–language models rely on high-level visual representations that are increasingly used across recognition, retrieval, and multimodal reasoning pipelines. However, recent advances in generative modeling have shown that such features can often be inverted, enabling realistic reconstructions of the underlying image and raising significant privacy risks. We revisit this problem through the lens of reconstruction and propose *TrustCLIP*, a reconstruction-driven framework that treats a feature-conditioned generator as an explicit privacy adversary. TrustCLIP learns a projection between encoder features and

⋆ Work done during Meta internship.

downstream modules that is explicitly optimized to degrade the reconstructions produced by generative attackers while retaining the necessary signals for downstream tasks. Unlike prior defenses that rely on discriminative privacy metrics, TrustCLIP directly optimizes against a generative reconstruction attacker, targeting a threat not captured by standard evaluation protocols. We demonstrate its effectiveness in both conventional classification and multimodal large language model pipelines. Across these settings, TrustCLIP consistently reduces the fidelity of generative inversions while maintaining downstream task performance. Project page: `atnikos.github.io/trustclip`.

1 Introduction

Vision-language models such as CLIP [34] have become central components in modern computer vision. The models are robust and widely used, enabling zero-shot recognition, retrieval, and other open-world reasoning tasks. However, this power comes with a significant privacy concern. Recent works have shown that CLIP features can be inverted to produce high-fidelity reconstructions of the original image [6, 9, 40, 46]. Given how widely these features are deployed in modern systems, this threat is far from theoretical. If feature vectors collected on-device or shared with remote services are intercepted or misused, adversaries can reconstruct faces [36, 42], private home environments, medical imagery, or other sensitive content [45, 49]. Such reconstructions can enable re-identification, profiling, or even blackmail, with consequences ranging from personal harm to breaching data-protection regulations (e.g., GDPR, HIPAA, the EU AI Act). As vision-language features are increasingly deployed in healthcare [37], autonomous driving [12], and on personal devices, the ability to recover raw images from stored or transmitted embeddings poses a tangible risk to user privacy and trust.

Mitigating this risk, however, is not straightforward. Downstream tasks such as classification, visual question answering, and multimodal reasoning depend on the rich semantic content encoded in CLIP features. Naively suppressing information—for example by adding noise or quantizing the representation—degrades these features indiscriminately, harming both the reconstruction-relevant and the task-relevant components alike. The core difficulty is that the visual information exploited by a reconstruction attacker is deeply entangled with the information needed to solve downstream tasks: both rely on the same high-dimensional feature space produced by visual encoders. Moreover, most existing privacy defenses are evaluated against *discriminative* metrics—such as the accuracy of an attribute classifier on the protected features—which fail to capture *generative* leakage, where a diffusion-based attacker reconstructs visually coherent images that reveal sensitive content even when re-identification accuracy is low. Any privacy mechanism must therefore be selective, suppressing the components that enable faithful image recovery while preserving those that carry task-critical semantics. This raises a fundamental question: *how can we preserve the utility of CLIP features for downstream tasks, while mitigating the privacy risks that arise from their misuse?*

Prior work has explored privacy-preserving representations through adversarial projection [7], video obfuscation [20], and secure aggregation [41], yet existing defenses either sacrifice task performance or only partially reduce reconstruction fidelity, leaving an unsatisfactory privacy–utility trade-off. In this work, we propose *TrustCLIP*, a framework for learning privacy-preserving visual representations against generative inversion attacks without sacrificing task-specific performance. We consider a realistic threat model in which an adversary has access to intermediate vision features[5] produced by a frozen encoder and to a powerful image-conditioned generative model that can synthesize images directly from these features (see Fig. 2). Rather than reasoning about privacy only through proxy metrics on latent space, we explicitly model this reconstruction pathway and treat *invertibility under a generative attacker* as the primary privacy risk. Our approach inserts a lightweight, adversarially-trained projection between the frozen CLIP encoder and different downstream networks using its features. The projection is optimized under two competing objectives: a *task loss* that preserves downstream accuracy, and a *reconstruction loss* that penalizes the fidelity of images recovered by a generative attacker conditioned on the projected features. This adversarial formulation forces the projection to discover which components of the representation are critical for reconstruction versus those that are critical for the task, and to selectively suppress the former. The premise underlying this approach is that downstream tasks and reconstruction attacks do not depend on the *same* information within CLIP features to the same extent. Tasks such as classification and visual question answering primarily depend on high-level semantic structure—object categories, spatial layout, and coarse scene understanding—whereas faithful image reconstruction additionally requires instance-specific detail: fine textures, skin tones, facial geometry, and identity-revealing cues. Because these informational demands are not fully overlapping, a learned projection can degrade reconstruction fidelity without proportionally harming task performance. Our experiments confirm this separation (Fig. 1): TrustCLIP features preserve scene-level semantics (Top-1 accuracy within 0.5% of the unprotected baseline) while obliterating the fine-grained detail that enables identity recovery (DSIM improves 2.5×). The projection itself is a lightweight MLP that integrates naturally into existing pipelines—we demonstrate its effectiveness in both image classification and vision–language models, showing that the adversarial training framework generalizes across diverse downstream architectures.

We instantiate our framework using IP-Adapter [46], a diffusion-based generator conditioned on CLIP features, which represents one of the strongest publicly available reconstruction attacks for vision–language embeddings. Although we focus on CLIP and IP-Adapter, the adversarial training formulation is not specific to this combination and can be applied to other encoders or generative attackers. Our contributions are as follows: **(i)** We identify generative leakage—the ability of a diffusion-based attacker to reconstruct privacy-revealing images

[5] For example, in client-server inference where a device sends CLIP features to a cloud VLM or in retrieval systems that cache embeddings.

Fig. 2: CLIP features reveal privacy information. Top: original images. Bottom: IP-Adapter reconstructions from frozen CLIP Vision features. Outputs closely match the originals, preserving layout, pose, and fine-grained semantics, including sensitive cues such as perceived gender, skin tone, age (children vs. adults), facial affect (e.g., desperation vs. smile), and detailed indoor/outdoor context.

from encoder features—as a distinct and underexplored threat not captured by standard discriminative privacy metrics, and introduce an evaluation framework that directly measures it. **(ii)** We propose a lightweight, adversarially-trained projection that explicitly reduces reconstruction fidelity under the attacker while maintaining task utility, yielding improved privacy–utility trade-offs. **(iii)** We construct privacy-preserving variants of both conventional image classification models and multimodal large language models (MLLMs), demonstrating that the projection layer integrates naturally into these architectures and supports privacy across diverse downstream tasks. **(iv)** We perform extensive evaluations across privacy metrics (LPIPS, DreamSim) and downstream benchmarks in classification and vision–language modeling, demonstrating that the proposed method significantly improves privacy preservation while retaining competitive performance.

2 Related Work

Feature inversion & privacy leakage. Deep representations retain enough information to reconstruct input images, whether by optimizing pixels to match activations [27], learning feature-to-image decoders [9], or exploiting models to reveal training-data prototypes [4,13]. Powerful generative priors encoded in GANs and diffusion models dramatically improve reconstruction fidelity [5, 8, 16, 28]. CLIP [19, 34] is especially vulnerable: its patch tokens encode rich semantics while staying closely aligned to natural-image statistics through web-scale contrastive pretraining. This property makes CLIP embeddings more reconstructible than CNN or self-supervised descriptors [6, 21].

Diffusion adapters & CLIP-conditioned generation. Diffusion models [17, 35,38] are the dominant backbone for high-fidelity synthesis. Lightweight adapters such as ControlNet [47], T2I-Adapter [31], IP-Adapter [46] condition the denoising process on spatial or feature-level guidance while keeping the backbone frozen. IP-Adapter maps CLIP features into the text-conditioning space, letting

a frozen diffusion U-Net reconstruct rich semantic and perceptual detail from visual tokens alone. Because both CLIP and the diffusion backbone are trained on natural image–text alignments, their latent spaces are inherently compatible—making this a strong yet practical adversary, as all components are publicly released. We exploit this property as our white-box attack.

Privacy-preserving representations. Approaches to suppress sensitive information while preserving task utility include adversarial disentanglement [11,22], information bottlenecks [1]. In video, SPAct [7] and STPrivacy [23] adversarially train encoders that maintain action recognition while concealing identity; hardware solutions such as coded-aperture recognition [43] offer non-invertible capture; and Ilic et al. [20] use optical flow with DINO features for motion-consistent obfuscation. DP-CLIP [18] applies differential privacy at the batch level during contrastive training, offering formal guarantees at the cost of accuracy. A key limitation of most defenses is their reliance on *proxy* privacy metrics—attribute-classifier accuracy or mutual-information estimates—which capture *discriminative* leakage but miss *generative* leakage, where a diffusion-based adversary reconstructs privacy-revealing images despite low re-identification accuracy. TrustCLIP addresses this gap: it is, to our knowledge, the first defense that directly optimizes against a generative reconstruction attacker, treating the fidelity of diffusion-based inversion—rather than attribute-classifier accuracy—as the primary privacy objective. Our framework explicitly evaluates privacy under such a generative threat.

Inversion-resistant descriptors & scene-level privacy. In retrieval and localization, NinjaDesc [32] adversarially trains descriptors that resist decoder-based reconstruction, while Adversarial Affine Subspace Embeddings [10] obfuscate features while preserving geometric matching. At the scene level, geometric representations also leak appearance: SfM point clouds can be inverted to recover images [33], and subsequent work has proposed mitigations through geometric substitutions [39], uncalibrated localization [15], and ray clouds [30]. These methods target descriptor-level or geometric reconstruction. In contrast, we address a complementary and increasingly critical issue: the inversion of language-aligned vision features—specifically CLIP tokens—through diffusion-based generation, a threat that grows in relevance as such features become standard between vision encoders and downstream models in modern VLMs.

3 Method

Figure 3 illustrates the TrustCLIP framework. A frozen vision encoder f_v (e.g. CLIP ViT-L/14) maps an input image x to a sequence of feature tokens $z = f_v(x) \in \mathbb{R}^{T \times D}$. A lightweight privacy projection P_θ transforms these tokens into $\tilde{z} = P_\theta(z)$ before they reach any downstream consumer. Two modules compete during training: a *task head* h_{task} that produces predictions $\hat{y} = h_{\text{task}}(\tilde{z})$ under a task-specific loss (§3.3), and a generative *attacker* G_ϕ that attempts to reconstruct the original image $\tilde{x} = G_\phi(\tilde{z})$ (§3.2). The projection P_θ and task head are jointly optimized to minimize task loss while maximizing reconstruction er-

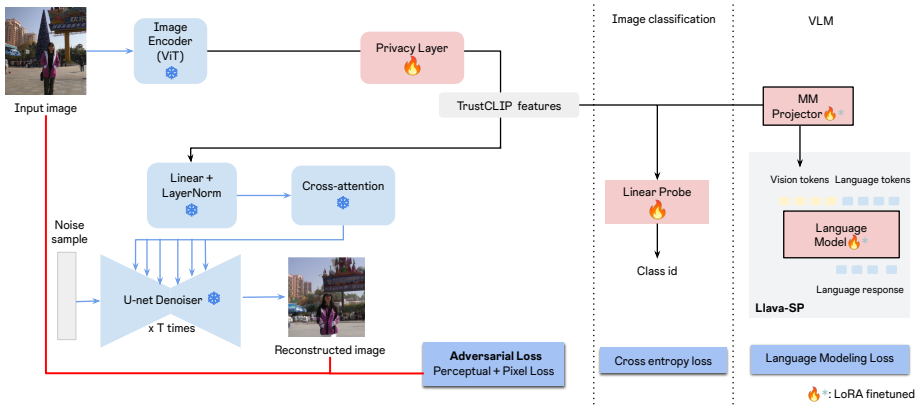

Fig. 3: Overview of TrustCLIP. A frozen vision encoder extracts feature tokens from an input image. The privacy projection P_θ transforms these features before they are consumed by any downstream module. During training (red path), a frozen generative attacker G_ϕ attempts to reconstruct the original image from the projected features; the reconstruction loss gradient is backpropagated through the attacker to update P_θ. The task heads (image classification/VLM) simultaneously optimize task performance. At deployment, only the non-red path is active: the attacker is discarded, and P_θ functions as a lightweight, drop-in privacy layer. ❄ frozen; 🔥 trainable.

ror, so that P_θ learns to suppress information that enables image recovery while preserving information that supports the downstream task. At deployment, only the green path in Figure 3 is used: the attacker is discarded, and P_θ serves as a drop-in privacy layer that adds negligible overhead and requires no architectural modification to the encoder or downstream model. We emphasize that, while P_θ itself is lightweight, the downstream task head is fine-tuned jointly with it—a linear probe for classification and LoRA adapters for the VLM (§3.3)—so the contribution is a privacy layer that requires no change to the network *architecture*, rather than one that leaves the downstream model entirely untouched.

Threat Model. We consider a setting in which an adversary gains access to intermediate vision features—for example, CLIP embeddings transmitted from a client device to a cloud-hosted VLM, cached in a retrieval database, or exposed through a multi-tenant storage failure. Given these features, the adversary attempts to reconstruct the original image using a strong generative model: concretely, an IP-Adapter [46] conditioned on Stable Diffusion [35]. To ensure a rigorous evaluation, each attacker is an IP-Adapter *finetuned on the feature distribution it attacks*: the attacker for unprotected CLIP features is trained on CLIP features, and the attacker for TrustCLIP features is trained on TrustCLIP features. This gives every method its own strongest possible adversary. The goal of TrustCLIP is to learn a projection P_θ that degrades the fidelity of images reconstructed from the projected features, while preserving the semantic con-

tent needed for downstream tasks. §3.1 argues why this selective suppression is feasible.

3.1 Privacy-Preserving Projection

Why selective suppression is feasible. The success of TrustCLIP rests on a structural property of CLIP representations: downstream tasks primarily require high-level semantic content—object categories, spatial layout, coarse scene structure—whereas pixel-accurate reconstruction additionally requires instance-specific detail such as textures, skin tones, and facial geometry. Because these informational demands do not fully overlap, a learned projection can suppress the latter while preserving the former. Our experiments confirm this empirically: TrustCLIP features retain scene-level accuracy within 0.5% of the baseline while degrading reconstruction fidelity by 2.5× in DSIM (§4.2). The naive baselines in Tab. 2 further validate this premise: Gaussian noise cannot achieve this separation because it perturbs all feature dimensions indiscriminately, whereas TrustCLIP's adversarial training learns to target only the reconstruction-critical components. We require $P_\theta : \mathbb{R}^{T\times D} \to \mathbb{R}^{T\times D}$ transforms encoder features into $\tilde{z} = P_\theta(z)$ such that reconstructible information is suppressed while task-relevant structure is maintained. We require P_θ to satisfy three constraints: (i) *Model-agnostic*: it operates purely on the feature sequence, independent of the specific encoder or task head. (ii) *Lightweight*: it adds negligible runtime overhead compared to the backbone encoder. (iii) *Non-destructive at initialization*: at the start of training, it preserves the original representation to avoid cold-start degradation of utility. We implement P_θ as a shallow, token-wise network with a residual connection, $P_\theta(z) = z + f_\theta(z)$, where f_θ is a small MLP applied independently to each token (optionally with layer normalization). The last linear layer of f_θ is initialized to zero, so P_θ starts as the identity mapping. This yields a smooth *privacy–utility knob*: as λ_{rec} increases (§3.3), P_θ gradually deviates from identity to suppress the reconstructible components while the task loss protects the semantic subspace. Although our experiments instantiate f_v with CLIP, the projection itself is agnostic to the backbone and can be placed in front of any consumer of tokenized vision features, including VLMs and non-contrastive encoders.

3.2 Generative Attacker

The attacker $G_\phi : \mathbb{R}^{T\times D} \to \mathcal{X}$ is a generative model that takes feature tokens as conditioning and produces a reconstruction $\tilde{x} = G_\phi(\tilde{z})$. In principle, G_ϕ can be any feature-conditioned image decoder: a diffusion model with cross-attention adapters, a GAN with feature conditioning, or a latent diffusion model.

Training-time attacker (fixed). In our experiments, G_ϕ is an IP-Adapter module built on a Stable Diffusion backbone [35, 46]. The projected features $\tilde{z}$ are converted into conditioning tokens for the diffusion U-Net via a lightweight adapter. This model is pre-trained on unprotected CLIP features and remains *frozen*

while P_θ is optimized. Freezing the attacker during projection training provides a stable adversarial signal and avoids the instabilities of alternating min-max optimization [29].

Evaluation-time attacker (adaptive). To evaluate robustness under a worst-case adversary, we conduct a second stage after P_θ converges: a new generative model $G_{\phi'}$ is retrained *directly* on the protected features $\tilde{z}$ produced by the converged projection $P_\theta^\star$:

$$\phi' = \arg\min_{\phi} \; \mathbb{E}_{x\sim\mathcal{D}}\big[\mathcal{L}_{\text{rec}}(x,\; G_\phi(P_{\theta^\star}(f_v(x))))\big]. \tag{1}$$

This attacker has full knowledge of the defense—access to the projected feature distribution and freedom to optimize its decoder—representing the adaptive adversary of §3.

IP-Adapter combined with Stable Diffusion represents the current frontier for feature-conditioned image generation: the diffusion backbone is trained on billions of images, and the adapter is optimized on millions of image–feature pairs, making it a near-optimal decoder for CLIP features. All components are publicly released, so any motivated adversary can deploy this attack. By evaluating against both the fixed attacker and a stronger adaptive variant retrained on the protected feature distribution (Eq. 1), we test TrustCLIP under a realistic worst-case scenario. Although we focus on this specific attacker, the projection reduces the information available in $\tilde{z}$ for *any* reconstruction method: by the data processing inequality, no downstream function of $\tilde{z}$ can recover more about x than $\tilde{z}$ itself contains.

3.3 Training Objective

Given an input x with task label y, encoder features $z = f_v(x)$, and projected features $\tilde{z} = P_\theta(z)$, the task head produces predictions $\hat{y} = h_{\text{task}}(\tilde{z})$ and the frozen attacker produces a reconstruction $\tilde{x} = G_\phi(\tilde{z})$. The reconstruction loss combines pixel-level and perceptual distances: $\mathcal{L}_{\text{rec}}(x, \tilde{x}) = \alpha \, \|x - \tilde{x}\|_p + (1 - \alpha)\, \text{LPIPS}(x, \tilde{x})$, where $\|\cdot\|_p$ is an ℓ_p pixel distance ($p \in \{1, 2\}$) and LPIPS [48] provides perceptual similarity. The full objective jointly optimizes the projection P_θ (parameters θ) and the task head h_{task} (parameters ψ):

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{x,y}[\mathcal{L}_{\text{task}}(h_\psi(\tilde{z}),\, y) \;-\; \lambda_{\text{rec}}\, \mathcal{L}_{\text{rec}}(x,\, G_\phi(\tilde{z}))]\,. \tag{2}$$

The first term drives task performance; the second, with a negative sign, encourages P_θ to *increase* reconstruction error under the fixed attacker. The scalar $\lambda_{\text{rec}} \geq 0$ controls the privacy–utility trade-off. In practice, we differentiate through the frozen attacker and backpropagate the negative reconstruction gradient to θ.

Task heads. The framework supports arbitrary downstream tasks. For *image classification*, h_{task} is a linear probe on pooled features trained with cross-entropy. For *vision–language modeling*, P_θ is placed before the VLM visual adapter (e.g., LLaVA-SP's multimodal projector [26]) and $\mathcal{L}_{\text{task}}$ is the standard language-modeling loss. In both settings, f_v is frozen; only P_θ and h_{task} are updated. Full task-head details and hyperparameters are provided in the supplementary material.

Why joint optimization? We train end-to-end by design. A two-stage baseline—optimize P_θ for privacy, then train h_{task} on frozen projected features—would corrupt features indiscriminately, since the privacy objective alone provides no signal about what to preserve. Joint optimization balances gradients from $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{rec}}$, suppressing reconstructible information while retaining task-relevant structure. Identity initialization (§3.1) further stabilizes this process: starting from the original CLIP representation lets both objectives co-adapt gradually, rather than requiring the projection to relearn useful structure after an unconstrained privacy-only phase.

Identity initialization and warm-up. The projection starts as the identity (f_θ initialized to zero), so training begins from the original CLIP features. For classification, λ_{rec} is constant from the start. For VLM training, we adopt a gradual warm-up: λ_{rec} is set to zero for the first N steps while the task head adapts to the feature space, and is then linearly increased to its target value. More details in the supplementary material.

4 Experiments

We evaluate TrustCLIP on image classification (SUN397) and multimodal reasoning (LLaVA-SP), in classification and VLM benchmarks. All qualitative results use the adaptive attacker unless noted otherwise.

4.1 Setup

Backbones. For classification, we attach a linear probe to the frozen CLIP ViT-L/14 encoder. For multimodal reasoning, we integrate TrustCLIP into LLaVA-SP [26], yielding *TrustLLaVA*. The VLM uses CLIP-ViT-L/14@336 as the vision encoder and Vicuna-1.5-7B as the LLM, trained with 558K image–text pairs and 665K instruction-following examples in a two-stage pre-train + LoRA regime. For a fair comparison, TrustLLaVA follows the *exact* LLaVA-SP pipeline (identical training data and LoRA configuration), differing only by P_θ and the adversarial loss; the LLaVA-SP row in Tab. 3 is thus its matched, in-distribution-finetuned counterpart, not an off-the-shelf VLM.
Datasets. For classification: SUN397 [44] with standard train/val/test splits; images resized to 512×512 for reconstruction and 224×224 for feature extraction. For VLMs: the standard benchmark suite used in LLaVA-SP [26].

Attacker. Our generative attacker is an IP-Adapter [46] module built on Stable Diffusion v1.5 [35]. Projected features are mapped to conditioning tokens for the diffusion U-Net, which reconstructs images using DDIM with 20 inference steps. Random seeds are fixed so that pre-/post-projection comparisons share identical noise. Attacker training details and capacity analysis are provided in the supplementary material.

Metrics. *Privacy:* PSNR, SSIM (pixel fidelity), LPIPS [48], DreamSim (DSIM) [14] (perceptual/semantic distance). Lower PSNR/SSIM and higher LPIPS/DSIM indicate stronger privacy. *Utility:* Top-1, Top-5, and mean-class accuracy for classification; official benchmark metrics for VLMs.

Baselines. We compare TrustCLIP to the unprotected baseline in which CLIP features are passed directly to both the attacker and downstream heads (*CLIP+IP-Adapter*), illustrating the privacy risk in current systems. For VLMs, we report alongside Qwen-VL [2], Qwen-VL-Chat [2], LLaVA-1.5 [25], and LLaVA-SP [26]. To our knowledge, TrustCLIP is the first method to adversarially modify language-aligned vision features using a generative attacker. Prior methods such as [32] and DP-CLIP [18] target fundamentally different settings—sparse local descriptors for geometric matching, and training-time differential privacy for the CLIP pretraining corpus, respectively—and are therefore not directly comparable. These differences are definitional: DP-CLIP [18] bounds *training-set membership* rather than inference-time reconstruction, while NinjaDesc [32] and SPAct [7] defend a single task against *discriminative* attackers—whereas we defend foundation-model features serving many tasks against a *generative* inverter. Porting them to dense CLIP tokens under a generative attacker would change each method beyond recognition, so the matched-ℓ_2 Gaussian control (§4.2) instead gives the apples-to-apples comparison they cannot.

Implementation details. During training, the CLIP encoder and all attacker components are frozen; only P_θ and the task head are updated. At evaluation time, we additionally test against an adaptive attacker retrained on the protected features (§4.2). P_θ is initialized to identity (§3.1), ensuring training begins from the original CLIP representation. We train with AdamW using fixed random seeds. Full hyperparameters are provided in the supplementary material.

4.2 Main Results

Classification. Tab. 1 reports both task accuracy and reconstruction metrics on SUN397. Across all TrustCLIP configurations, Top-1 accuracy remains within 0.5% of the unprotected CLIP baseline (83.4%), with the best setting (L2+LPIPS, λ_{rec}=0.25) achieving 82.9%. Simultaneously, PSNR drops from 13.58 to 10.47–10.69 (a 21–23% reduction), SSIM from 0.288 to 0.187–0.204, and DSIM improves 2.5×. The unprotected baseline confirms the severity of the privacy risk: despite strong accuracy, its reconstructions remain highly faithful.

Comparison with naive baselines. Tab. 2 compares TrustCLIP against Gaussian noise at varying intensities, evaluated on 500 SUN397 test samples under the

Table 1: Privacy-utility trade-offs on image classification. Evaluated using classification and reconstruction metrics for TrustCLIP variants and the CLIP using IP-Adapter for adversarial attack. Higher values indicate better classification performance; lower PSNR/SSIM and higher LPIPS/DSIM reflect stronger privacy preservation.

				Classification metrics			**Reconstruction errors (mean ± std)**			
Setting	Pixel	Perc.	λ_{rec}	Top-1	Top-5	Mean Cl.	PSNR ↓	SSIM ↓	LPIPS ↑	DSIM ↑
CLIP	—	—	—	83.36	97.76	80.46	13.581 ± 2.203	0.288 ± 0.147	0.522 ± 0.066	0.215 ± 0.055
TrustCLIP	L1	—	0.25	82.662	97.361	79.871	10.555 ± 1.608	0.191 ± 0.102	0.704 ± 0.051	0.522 ± 0.098
TrustCLIP	L1	LPIPS	0.25	78.547	95.899	72.376	10.639 ± 1.659	0.204 ± 0.109	0.702 ± 0.054	0.535 ± 0.098
TrustCLIP	L2	—	0.25	82.487	97.411	79.621	10.527 ± 1.576	$\mathbf{0.187 \pm 0.097}$	0.702 ± 0.051	0.514 ± 0.097
TrustCLIP	L2	LPIPS	0.25	**82.915**	97.582	79.270	10.687 ± 1.655	0.203 ± 0.109	0.699 ± 0.053	0.514 ± 0.097
TrustCLIP	L2	—	0.5	82.653	97.434	79.892	10.681 ± 1.622	0.195 ± 0.101	0.701 ± 0.049	0.523 ± 0.099
TrustCLIP	L2	LPIPS	0.5	81.411	97.140	76.771	10.685 ± 1.661	0.200 ± 0.108	0.697 ± 0.052	0.510 ± 0.095
TrustCLIP	L2	—	1	82.869	97.517	**80.068**	$\mathbf{10.465 \pm 1.616}$	0.188 ± 0.100	0.714 ± 0.051	0.556 ± 0.110
TrustCLIP	L2	LPIPS	1	82.763	**97.669**	79.070	10.617 ± 1.590	0.203 ± 0.093	$\mathbf{0.776 \pm 0.061}$	$\mathbf{0.799 \pm 0.068}$

Table 2: Comparison with naive baselines on SUN397. All methods evaluated under the same fixed attacker (trained on unprotected CLIP features) on test samples. TrustCLIP achieves substantially stronger privacy at comparable or higher accuracy than any noise level.

Method	Top-1	LPIPS ↑	DSIM ↑
CLIP (no defense)	83.9	0.58	0.34
Noise σ=0.01	84.1	0.59	0.34
Noise σ=0.05	83.8	0.58	0.33
Noise σ=0.1	82.7	0.58	0.33
Noise σ=0.25	75.2	0.60	0.36
Noise σ=0.5	54.4	0.64	0.44
Noise σ=1.0	17.4	0.71	0.68
TrustCLIP (ours)	**85.1**	**0.90**	**0.88**

same fixed attacker. The results expose a fundamental limitation of indiscriminate perturbation: at low noise ($\sigma \leq 0.1$), accuracy is preserved but privacy barely improves over the undefended baseline (DSIM: 0.33 vs. 0.34). At high noise ($\sigma = 1.0$), accuracy collapses to 17.4% yet DSIM reaches only 0.68—still well below TrustCLIP. In contrast, TrustCLIP maintains accuracy above the baseline (85.1%) while achieving DSIM of 0.88, confirming that learned, selective suppression is fundamentally superior to indiscriminate perturbation. To rule out that this gain is merely an artefact of perturbation magnitude, we add a control that matches the noise to TrustCLIP at equal feature displacement: isotropic Gaussian noise calibrated to the same ℓ_2 perturbation $\|\tilde{z} - z\|_2$ as our projection. At this matched budget the Gaussian control reaches only DSIM 0.64 at 12.5% Top-1, whereas TrustCLIP attains DSIM 0.80 at 82.9% (undefended CLIP: 0.21 / 83.4%)—strictly worse on both privacy and utility. The advantage of TrustCLIP is thus a property of *where* information is removed, not *how much*.

Vision-language modeling. Tab. 3 (top) reports TrustLLaVA alongside strong VLMs on the standard benchmark suite, with reconstruction metrics in the lower

Table 3: VLM utility and privacy evaluation. *Top:* Comparison with SoTA methods. *Bottom:* Reconstruction quality when each method is attacked by a dedicated IP-Adapter finetuned on its own feature distribution (i.e., CLIP features for LLaVA-SP, TrustCLIP features for TrustLLaVA). Lower PSNR/SSIM and higher LPIPS/DSIM indicate stronger privacy.

Method	LLM	Res.	VQA^{v2}	GQA	SQA^{I}	VQA^{T}	POPE	MME^{P}	SEED^{I}	LLaVA^{W}	MM-Vet
Qwen-VL [3]	Qwen-7B	448	78.8	59.3	67.1	**63.8**	–	–	56.3	–	–
Qwen-VL-Chat [3]	Qwen-7B	448	78.2	57.5	68.2	61.5	–	1487.5	58.2	–	–
LLaVA-1.5 [24]	Vicuna-7B	336	78.5	62.0	66.8	58.2	85.9	**1510.7**	66.2	63.4	30.5
LLaVA-1.5 [24]	Vicuna-7B	336	78.4	61.9	67.6	56.2	85.8	1477.4	67.0	64.2	32.1
LLaVA-SP [26]	Vicuna-7B	336	**79.2**	**62.7**	**68.4**	58.5	**86.6**	1470.7	**67.9**	61.7	33.8
TrustLLaVA	Vicuna-7B	336	76.3	58.0	62.1	55.0	86.1	1390.8	63.0	**65.3**	**34.5**
TrustLLaVA (MLP)	Vicuna-7B	336	72.8	55.4	60.5	51.1	85.3	1296.8	58.2	51.5	27.8

	Reconstruction metrics (mean ± std)			
	PSNR ↓	SSIM ↓	LPIPS ↑	DSIM ↑
LLaVA-SP [26]	10.57 ± 1.83	0.26 ± 0.13	0.68 ± 0.06	0.32 ± 0.08
TrustLLaVA	10.42 ± 1.72	0.24 ± 0.12	0.70 ± 0.06	0.39 ± 0.09
TrustLLaVA (MLP)	**7.13** ± 1.22	**0.12** ± 0.08	**0.81** ± 0.06	**0.62** ± 0.10

panel. TrustLLaVA preserves competitive performance on coarse-grained tasks: POPE (86.1 vs. 86.6 for LLaVA-SP) and VQAv2 (76.3 vs. 79.2), and notably *improves* on LLaVA-Bench (+3.6) and MM-Vet (+0.7), suggesting that adversarially shaping the representation can reduce hallucination-style errors. On the privacy side, DSIM improves from 0.32 (LLaVA-SP) to 0.39, with corresponding gains in LPIPS and reductions in PSNR and SSIM. Tab. 4 breaks down MME-Perception and SEED-Bench by category type. Tasks relying on high-level semantics—existence detection, artwork identification, action recognition—are fully preserved (avg. Δ: –1.9% on MME, –0.2% on SEED). Tasks depending on fine-grained visual detail—counting, OCR, localization—show larger degradation (avg. Δ: –11.1% on MME, –6.4% on SEED). This pattern directly confirms the premise of §3.1: the projection selectively suppresses instance-specific cues while retaining the semantic structure that drives most VLM capabilities. The privacy gains justify these modest costs on fine-grained tasks, while the categories most relevant to typical VLM usage remain intact.

Robustness under adaptive and unseen attackers. To test the strongest threat model, we retrain a new IP-Adapter directly on TrustCLIP features (Eq. 1), giving the adversary full knowledge of the defense. On SUN397 this adaptive attacker recovers only marginally more detail than the fixed variant (PSNR: 11.04 vs. 10.62) and remains far less effective than the vanilla CLIP attacker on unprotected features (PSNR: 13.58); for VLMs, the supplementary material confirms consistent gains (DSIM: 0.42–0.43 for identity-init, 0.59–0.62 for MLP, vs. 0.32 baseline). The defense also holds against attackers it never co-trained with: a higher-capacity *IP-Adapter Plus* transfer attacker recovers *less* detail than

Table 4: Per-category analysis of TrustLLaVA vs. LLaVA-SP. Semantic tasks are preserved while fine-grained tasks degrade, confirming selective suppression of instance-specific detail (§3.1).

	Avg. Δ	Top categories
MME-Perception		
Semantic	**−1.9%**	Exist. (0.0), Art. (0.0), Pos. (–2.5)
Fine-grained	–11.1%	Post. (–18.9), Count (–14.5), OCR (–12.2)
SEED-Bench		
Semantic	**−0.2%**	Act.P (+2.9), Act.R (+0.9), Inter. (0.0)
Fine-grained	–6.4%	Count (–9.0), Loc. (–7.5), ID (–6.6)

Fig. 4: Qualitative comparison on SUN397 [44]: **original** (top), reconstructions from the **vanilla CLIP** attacker (middle), and from the **TrustCLIP** attacker (bottom) under the adaptive threat model (§3). Vanilla CLIP reveals faces, pets, textures, and distinctive color patterns; TrustCLIP obfuscates these while preserving scene semantics and class-level structure.

our adaptive attacker (DSIM 0.88 vs. 0.80), and a non-diffusion *CNN decoder* fails likewise (PSNR 13.77 / DSIM 0.45 vs. 14.40 / 0.36 on unprotected CLIP). Consistent failure across families—diffusion and feed-forward alike—indicates TrustCLIP reduces the recoverable information itself rather than overfitting one inverter (§3.2); see the supplementary material.

4.3 Ablation Studies

Identity initialization vs. standard MLP. We compare two projection architectures on VLM benchmarks: (i) identity-initialized projection (default) and (ii) standard MLP without identity initialization. The identity-initialized variant maintains near-baseline performance (MM-Vet: 32.3–34.2 vs. 34.0 for LLaVA-SP; POPE: 86.2–86.6 vs. 86.6) with meaningful privacy gains (DSIM: 0.42–0.43 vs. 0.32). The MLP variant achieves stronger privacy (DSIM: 0.59–0.62) at significant utility cost (MM-Vet: 26.5–27.8). These configurations span a controllable

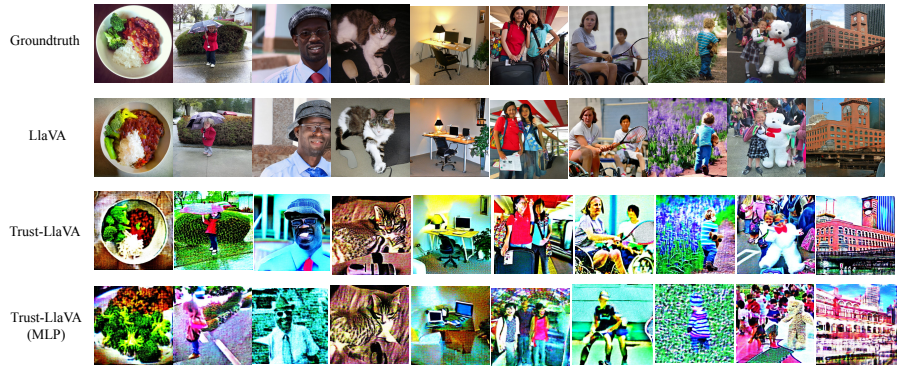

Fig. 5: Privacy spectrum on VLMs. Each column shows the same COCO validation image across four settings. **Row 1:** Ground truth. **Row 2 (LLaVA):** Reconstructions from unprotected CLIP features—faces, identities, objects, and scene details are recovered with high fidelity. **Row 3 (Trust-LLaVA):** Identity-initialized projection—facial features become unrecognizable and fine detail is suppressed, while scene-level semantics are preserved. **Row 4 (Trust-LLaVA MLP):** MLP projection without identity initialization—all personally identifiable information is destroyed at the cost of reduced VLM utility (§4.3). All reconstructions use the adaptive attacker (§3).

privacy–utility spectrum, and the comparison highlights the importance of starting from the CLIP manifold: the identity-initialized variant allows task and privacy gradients to co-adapt gradually, whereas the MLP variant departs early from the original representation and struggles to recover task-relevant structure.

Loss formulation and λ_{rec}. Tab. 1 ablates pixel loss (ℓ_1 vs. ℓ_2), perceptual loss (with/without LPIPS), and $\lambda_{\text{rec}} \in \{0.25, 0.5, 1.0\}$ on SUN397. L2+LPIPS at λ_{rec}=0.25 yields the best trade-off: Top-1 drops 0.5% while DSIM improves 2.5×. For VLMs, λ_{rec}=0.001 suffices; performance remains stable across three orders of magnitude (supplementary material).

Hyperparameters. For the identity-initialized projection, residual weight r and initialization scale ε have minimal impact: DSIM varies by only ∼0.01 across configurations, with r=1.0, ε=0.1 achieving the best trade-off (supplementary material tables). For the MLP variant, freeze duration and warmup length minimally affect privacy (DSIM: 0.59–0.62) but modestly affect utility (Supplementary Tables).

4.4 Qualitative Results

Fig. 4 compares reconstructions across diverse SUN397 scene categories under the adaptive attacker. Vanilla CLIP reveals sensitive details—recognizable faces and clothing in *indoor kennel* and *cockpit*, vivid textures and car shapes in *building facade*, and distinctive color patterns in *kitchen*. TrustCLIP systematically

obfuscates these attributes: humans are reduced to vague silhouettes, facade textures are smoothed into coarse blobs, and fine-grained details are washed out. Critically, scene semantics remain recognizable—the corridor geometry in *medina*, furniture layout in *dinette*, and room structure in *kitchen*—confirming that the projection preserves the task-relevant information identified in §3.1.

VLM reconstructions and privacy spectrum. Fig. 5 compares reconstructions across the full privacy spectrum. LLaVA-SP's unprotected CLIP features (Row 2) enable high-fidelity recovery: the man's face and clothing are clearly recognizable, the cat's fur texture is preserved, group compositions and individual identities in the outdoor scenes remain distinguishable, and indoor layouts are faithfully reproduced. The identity-initialized TrustLLaVA projection (Row 3) suppresses these sensitive attributes while coarse scene structure is preserved: the food plate, garden scene, office layout, and group configurations remain identifiable at the category level (DSIM: 0.42–0.43, MM-Vet: 32.3–34.2). The MLP variant without identity initialization (Row 4) pushes privacy to the extreme: faces are obliterated, colors become arbitrary, and personal items dissolve into abstract patterns, leaving only coarse spatial layout intact (DSIM: 0.59–0.62, MM-Vet: 26.5–27.8). Together, these configurations illustrate the controllable privacy–utility trade-off that TrustCLIP enables.

5 Conclusion

We presented TrustCLIP, a framework that strengthens the privacy of vision–language representations by adversarially training a lightweight projection to suppress the visual detail that enables generative inversion while preserving the semantic content needed for downstream tasks. A central observation is that existing privacy defenses focus on discriminative leakage, overlooking the threat posed by diffusion-based attackers that can reconstruct visually coherent images from encoder features. TrustCLIP directly addresses this gap by optimizing against a generative reconstruction attacker as the primary privacy objective. Experiments across image classification and VLM pipelines confirm that the approach consistently reduces inversion fidelity while maintaining competitive downstream performance, offering a practical privacy–utility trade-off.

Limitations and future work. Our evaluation focuses on CLIP encoders; while we evaluate against transfer and non-diffusion attackers (§4.2), the projection is *trained* against a single attacker family (IP-Adapter + Stable Diffusion). Extending the framework to other vision backbones (e.g., SigLIP, DINOv2) and generative architectures is a natural next step. Privacy suppression has a greater impact on tasks requiring fine-grained visual detail (e.g., OCR, counting) than on coarse-grained semantic tasks; adaptive mechanisms that modulate suppression strength per-token or per-task could mitigate this. Finally, applying TrustCLIP to video and multi-view settings, where temporal consistency introduces additional privacy surfaces, is a promising direction.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv preprint arXiv:1612.00410 (2016)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security '23). pp. 5253–5270 (2023)
5. Chen, C., Liu, D., Shah, M., Xu, C.: Enhancing privacy-utility trade-offs to mitigate memorization in diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8182–8191 (2025)
6. Chen, Y., Wang, X., Wang, S., Ma, X.: Leakyclip: Extracting training data from CLIP. arXiv preprint arXiv:2508.00756 (2025), https://arxiv.org/abs/2508.00756
7. Dave, I.R., Chen, C., Shah, M.: SPAct: Self-supervised privacy preservation for action recognition. In: CVPR. pp. 20164–20173 (2022)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
9. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: CVPR. pp. 4829–4837 (2016)
10. Dusmanu, M., Schonberger, J.L., Sinha, S.N., Pollefeys, M.: Privacy-preserving image features via adversarial affine subspace embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14267–14277 (2021)
11. Edwards, H., Storkey, A.: Censoring representations with an adversary. arXiv preprint arXiv:1511.05897 (2015)
12. Elhenawy, M., Ashqar, H.I., Rakotonirainy, A., Alhadidi, T.I., Jaber, A., Tami, M.A.: Vision-language models for autonomous driving: CLIP-based dynamic scene understanding. Electronics **14**(7), 1282 (2025). https://doi.org/10.3390/electronics14071282
13. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 1322–1333 (2015)
14. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: Advances in Neural Information Processing Systems. vol. 36, pp. 50742–50768 (2023)
15. Geppert, M., Larsson, V., Speciale, P., Schönberger, J.L., Pollefeys, M.: Privacy preserving localization and mapping from uncalibrated cameras. In: CVPR. pp. 3316–3326 (2021)
16. Hintersdorf, D., Struppek, L., Brack, M., Friedrich, F., Schramowski, P., Kersting, K.: Does clip know my face? Journal of Artificial Intelligence Research **80**, 1033–1062 (2024)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)

18. Huang, A., Liu, P., Nakada, R., Zhang, L., Zhang, W.: Safeguarding data in multimodal ai: A differentially private approach to clip training. arXiv preprint arXiv:2306.08173 (2023)
19. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., et al.: OpenCLIP. https://github.com/mlfoundations/open_clip (2021)
20. Ilic, S., Meier, L., Pollefeys, M., Oswald, M.R.: Selective, interpretable and motion consistent privacy attribute obfuscation for action recognition. In: CVPR (2024)
21. Kazemi, H., Chegini, A., Geiping, J., Feizi, S., Goldstein, T.: What do we learn from inverting clip models? arXiv preprint arXiv:2403.02580 (2024)
22. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9012–9020 (2019)
23. Li, M., Xu, X., Fan, H., Zhou, P., Liu, J., Liu, J.W., Li, J., Keppo, J., Shou, M.Z., Yan, S.: Stprivacy: Spatio-temporal privacy-preserving action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5106–5115 (2023)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv:2310.03744 (2023)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
26. Lou, H., Fan, C., Liu, Z., Wu, Y., Wang, X.: Llava-sp: Enhancing visual representation with visual spatial tokens for mllms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22014–22024 (October 2025)
27. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR. pp. 5188–5196 (2015)
28. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
29. Mescheder, L.M., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International Conference on Machine Learning (2018), https://api.semanticscholar.org/CorpusID:3345317
30. Moon, H., Lee, C., Hong, J.H.: Efficient privacy-preserving visual localization using 3d ray clouds. In: CVPR. pp. 9773–9783 (2024)
31. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1965–1973 (2024)
32. Ng, T., Kim, H.J., Lee, V.T., DeTone, D., Yang, T.Y., Shen, T., Ilg, E., Balntas, V., Mikolajczyk, K., Sweeney, C.: Ninjadesc: Content-concealing visual descriptors via adversarial learning. In: CVPR. pp. 12787–12797 (2022)
33. Pittaluga, F., Koppal, S.J., Kang, S.B., Sinha, S.N.: Revealing scenes by inverting structure from motion reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 145–154 (2019)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference of Machine Learning. pp. 8748–8763. PMLR (2021)

35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
36. Shamshad, F., Naseer, M., Nandakumar, K.: CLIP2Protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20595–20605 (June 2023). https://doi.org/10.1109/CVPR52729.2023.01973
37. Sharshar, A., Khan, L.U., Ullah, W., Guizani, M.: Vision-language models for edge networks: A comprehensive survey. IEEE Internet of Things Journal (2025). https://doi.org/10.1109/JIOT.2025.3579032
38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
39. Speciale, P., Schönberger, J.L., Kang, S.B., Sinha, S.N., Pollefeys, M.: Privacy preserving image-based localization. In: CVPR. pp. 5493–5503 (2019)
40. Struppek, L., Hintersdorf, D., Correia, A., Adler, A., Kersting, K.: Plug & play attacks: Towards robust and flexible model inversion attacks. In: International Conference of Machine Learning. pp. 20522–20545. PMLR (2022)
41. Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., Zhou, Y.: A hybrid approach to privacy-preserving federated learning. In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. pp. 1–11 (2019)
42. Wang, Z., Wang, H., Jin, S., Zhang, W., Hu, J., Wang, Y., Sun, P., Yuan, W., Liu, K., Ren, K.: Privacy-preserving adversarial facial features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8212–8221 (June 2023). https://doi.org/10.1109/CVPR52729.2023.00794
43. Wang, Z.W., Vineet, V., Pittaluga, F., Sinha, S.N., Cossairt, O., Bing Kang, S.: Privacy-preserving action recognition using coded aperture videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
44. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 3485–3492 (2010), https://api.semanticscholar.org/CorpusID:1309931
45. Xiu, K., Zhang, S.Q.: CapRecover: A cross-modality feature inversion attack framework on vision language models. In: Proceedings of the 33rd ACM International Conference on Multimedia (MM '25). ACM, Dublin, Ireland (2025). https://doi.org/10.1145/3746027.3755203
46. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
49. Zhou, Z., Zhu, J., Yu, F., Li, X., Peng, X., Liu, T., Han, B.: Model inversion attacks: A survey of approaches and countermeasures. arXiv preprint arXiv:2411.10023 (2024)