3D FaceShell: Attribute Transfer in 3D Face Avatars as a VLM Defense Mechanism

Weston Bondurant, Srijan Das†, Hieu Le†, and Stephanie Schuckers†

University of North Carolina at Charlotte
wbondura@charlotte.edu
https://westonbond.github.io/3DFS/

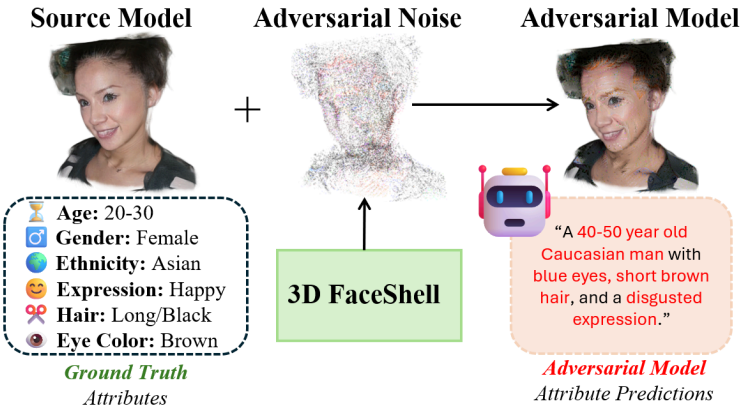

Fig. 1: Illustration of our 3D FaceShell method. By optimizing a minimally perceptible 3D Gaussian shell, our method successfully subverts VLM attribute predictions across rendered views while preserving the original facial identity and photorealistic appearance.

Abstract. Photorealistic 3D face avatars are increasingly deployed as reusable digital assets across applications such as telepresence, animation, and personalized media. At the same time, vision–language models (VLMs) can infer sensitive attributes from rendered images with open-ended semantic reasoning without any fine-tuning. This creates a new privacy challenge: once a 3D face avatar is shared, any of its renderings can be analyzed to extract high-level facial attributes. Existing defenses largely operate in 2D image space and do not address identity-preserving semantic manipulation of 3D facial representations.

We propose **3D FaceShell**, a framework for steering VLM interpretations of faces rendered from 3D models while preserving geometric fidelity and facial identity. 3D FaceShell augments the original 3D representation with a learnable Gaussian shell that produces subtle, spatially distributed perturbations optimized through multi-view embedding alignment. The perturbations are designed to be visually inconspicuous yet sufficient to redirect VLM-based attribute inference in a view-consistent manner.

† Equal advising.

Extensive experiments on reconstructed celebrity face avatars and multiple black-box VLMs demonstrate that 3D FaceShell significantly increases attribute injection and mismatch rates while maintaining high perceptual similarity and identity consistency. Our results show that it is possible to manipulate VLM-level semantic interpretation of 3D faces without compromising their human-recognizable appearance.

Keywords: 3D · Gaussian Splatting · Vision–Language Models · Face Attribute Inference

1 Introduction

Photorealistic 3D face avatars are rapidly becoming reusable digital assets. Modern pipelines reconstruct detailed 3D heads from short videos or even single images, enabling talking-head synthesis, virtual telepresence, personalized media, and immersive applications [18, 25, 42]. Unlike a static photograph, a 3D face model is persistent: it can be rendered under arbitrary viewpoints and lighting conditions long after its creation. At the same time, Vision–Language Models (VLMs), including both commercial foundation models such as GPT-4 or Gemini [23, 29] and open source models like VideoLLaMA3 or LLaVa-NeXT [12, 41], have evolved into open-ended semantic reasoners capable of inferring attributes such as *age*, *ethnicity*, *gender*, and *expression* from any rendered image without task-specific training [7, 16]. Together, these developments create a structural privacy risk: once a 3D face avatar is released, every possible rendering becomes analyzable by powerful semantic models, exposing sensitive attributes beyond simple identity recognition [3].

Existing defenses are fundamentally mismatched to this setting. Most face anonymization and adversarial cloaking methods operate in 2D image space, protecting only a specific rendering rather than the underlying 3D asset. A new viewpoint instantly bypasses such defenses. While recent adversarial work has explored perturbations in 3D representations, these approaches primarily target identity detectors and do not address identity-preserving semantic manipulation in VLM embedding space [14, 31, 36, 46].

To this end, we propose **3D FaceShell**, a framework that steers VLM interpretations of 3D face avatars while preserving the underlying facial identity and appearance as highlighted in Fig. 1. Our objective is to *alter specific attributes* inferred by VLMs from a 3D face avatar while preserving the underlying facial identity and geometry. Unlike generic 3D objects, facial avatars are designed to maintain a stable and recognizable digital likeness for legitimate uses such as telepresence, animation, and personalized media. Any intervention must therefore balance two competing requirements: it must be strong enough to redirect high-level semantic interpretation in the VLM embedding space, yet subtle enough to preserve geometric structure and identity-defining appearance across viewpoints. Naive perturbations that visibly distort shape or degrade texture undermine the asset's utility. The challenge is thus not merely to attack a model, but to manipulate semantic inference in a way that remains perceptually faithful and multi-view consistent.

To realize this objective, 3D FaceShell operates directly in the parameter space of 3D Gaussian Splatting (3DGS) face avatars. Starting from a pre-optimized facial representation, we introduce a set of auxiliary Gaussians positioned near the facial surface, forming a learnable Gaussian shell. The baseline Gaussians remain fixed, separating adversarial perturbations from underlying geometry. This approach enables semantic manipulation while, by design, never deforming facial structure. The auxiliary shell is optimized to produce subtle, spatially distributed perturbations that are minimally perceptible in rendered views yet sufficient to shift VLM embeddings toward a targeted semantic interpretation. During training, multiple rendered views are passed through a pre-trained VLM, and the shell parameters are updated via multi-view embedding alignment against a pose-consistent, attribute-altered target. This design ensures that the semantic manipulation is both view-consistent and visually inconspicuous.

Through extensive evaluation on reconstructed celebrity face avatars and multiple opaque-box VLMs, we show that 3D FaceShell consistently increases attribute injection and mismatch rates while maintaining high perceptual similarity and identity consistency. Compared to state-of-the-art 2D semantic attack baselines, 3D FaceShell achieves stronger cross-view robustness and substantially better preservation of facial structure, demonstrating that semantic steering and visual fidelity need not be mutually exclusive.

We summarize our contributions as follows:

- We introduce and formalize **the problem of manipulating VLM perception of semantic face attributes for 3D face avatars** under strict identity and geometric preservation constraints. Defending against VLM inference for 3D avatars is entirely new, differing fundamentally from prior 3D adversarial attacks that target face detectors and more general adversarial attacks on VLMs that do not preserve face quality or identity.
- We propose **3D FaceShell, a novel additive Gaussian shell framework** that injects spatially distributed, minimally perceptible perturbations into 3DGS face representations and optimizes them through multi-view embedding alignment in the vision encoder space to manipulate VLM perception. This design enables view-consistent semantic steering while leaving the baseline geometry untouched.
- We demonstrate **strong cross-view robustness and improved identity preservation** compared to state-of-the-art 2D semantic attack baselines across multiple opaque-box VLMs, showing that semantic manipulation and perceptual fidelity can be achieved simultaneously in 3D.

2 Related Work

2.1 3D Gaussian Splatting

The introduction of 3DGS has established a highly efficient paradigm for real-time radiance field rendering [10]. By utilizing millions of 3D Gaussians characterized by learnable opacity, scale, rotation, and spherical harmonics, 3DGS

achieves state-of-the-art view synthesis without the massive computational bottlenecks of earlier methods [6, 33, 34, 38]. Given this explicit editability and rendering speed, 3DGS has been widely adopted for modeling dynamic and highly detailed human face and body avatars. Recent frameworks [18, 25, 42] enable the generation of full body or face avatars from 2D video or images. However, as these 3D head avatars become increasingly common and easily shareable digital assets, the underlying geometry and identity textures are left inherently exposed to malicious downstream processing and automated semantic parsing by VLMs [1, 20, 30, 32]. This danger necessitates privacy defenses for 3D face avatars.

2.2 Adversarial Attacks on 3D Models

As the use of 3D representations has become more common, the vulnerabilities of 3D representations to adversarial manipulation have become a critical security domain [17, 19]. Discrete 3D structures, such as point clouds, are highly susceptible to targeted perturbations either through point shifting or malicious point generation [35]. However, adversarial attacks in the 3D domain are not limited to explicit point sets. Implicit representations like Neural Radiance Fields (NeRFs) are vulnerable to targeted perturbations that hallucinate or obscure objects during view synthesis [9]. Recently, this threat has been explicitly extended to 3DGS. Recent works inject adversarial noise in the 3DGS space to fool object detectors or vision encoders [17, 39]. Unlike these methods which shift overall semantic representations in object detectors or vision encoders for objects, we aim instead to shift VLM understanding of a more complex domain, *faces*, targeting specific attributes for transfer rather than broader semantics of the 3D representation.

Attacks on 3D Face Avatars. Biometric authentication systems are very frequently targeted by adversarial attacks leading to the development of specialized physical and 3D face attacks that bypass conventional defenses [14,31,36,46]. These methods, however, focus on attacks utilizing physical 3D masks or traditional methods of 3D representation like point clouds. In contrast, our 3D FaceShell proposes an attack on 3DGS face avatars, utilizing exclusively additive noise which leaves the underlying face representation undisturbed.

2.3 Adversarial Attacks on VLMs

Due to their extensive cross-modal training, foundation VLMs possess powerful reasoning capabilities but remain highly susceptible to visual adversarial attacks [24]. Images with adversarial noise applied to them can bypass safety guardrails and execute multimodal jailbreaks, coercing models into generating harmful or unintended text [22,37]. The threat landscape includes both targeted attacks on specific commercial systems [5] along with targeted and untargeted transferable frameworks designed to disrupt generalized vision-language alignment across a variety of VLMs [8, 15, 21, 43, 45]. 3D FaceShell differs from these by moving the problem of attacking VLMs with adversarial noise into the 3D space and focusing on shifting the attributes of a face representation while still maintaining a high level of detail and realism to retain the perceptual and identity similarity relative to the source face.

3 Problem Definition

We represent a 3D face using 3DGS as a set of N anisotropic Gaussians $\mathcal{G} = \{g_i\}_{i=1}^{N}$, where each Gaussian g_i is parameterized by its spatial mean $\mathbf{x}_i \in \mathbb{R}^3$, covariance $\Sigma_i \in \mathbb{R}^{3\times3}$, and opacity $\alpha_i \in \mathbb{R}$. Let $R(\mathcal{G};\pi)$ denote the differentiable renderer that produces a 2D image from $\mathcal{G}$ under camera parameters π. Simultaneously, we consider a pretrained VLM $f_\theta(\cdot)$, composed of a visual encoder $\mathrm{VE}(\cdot)$ followed by a large language model $\mathrm{LLM}(\cdot)$: $f_\theta(x) = \mathrm{LLM}(\mathrm{VE}(x))$, where θ denotes fixed model parameters.

Then, given a source 3D face $\mathcal{G}_s$ with ground-truth semantic description c_{gt}, our objective is to learn an additive perturbation δ in the Gaussian parameter space, yielding a modified 3D representation $\mathcal{G}_t = \mathcal{G}_s + \delta$, such that rendered views $R(\mathcal{G}_t;\pi)$ are semantically aligned with a predefined target description $\mathcal{T}_{tar}$ (targeted attack), while preserving the original facial identity and geometric structure. Under a white-box setting with access to f_θ, the optimization problem is formulated as:

$$\min_{\delta} \; \mathcal{L}_{adv}\big(f_\theta(R(\mathcal{G}_s+\delta;\pi)), \mathcal{T}_{tar}\big) + \lambda\, \mathcal{L}_{reg}(\mathcal{G}_s, \mathcal{G}_s+\delta), \tag{1}$$

where $\mathcal{L}_{adv}$ drives semantic alignment with the target description in the VLM output space, and $\mathcal{L}_{reg}$ constrains the perturbation to preserve geometric fidelity, identity consistency, and photorealistic rendering quality.

4 Method: 3D FaceShell

We propose **3D FaceShell**, an identity-preserving framework for learning additive adversarial perturbations directly in the parameter space of 3DGS facial representations to induce targeted semantic misalignment in pretrained VLMs.

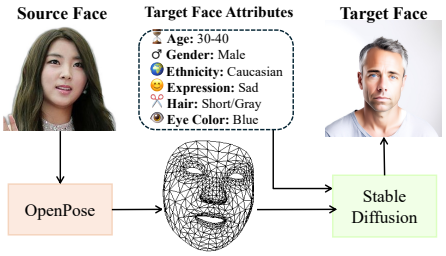

Fig. 2: Multi-Attributed Target Image Generation. (1) OpenPose extracts the structural pose from a source face. (2) A Stable Diffusion model, conditioned on this pose and a target attribute text prompt, synthesizes a pose-aligned, semantically different target image.

The proposed framework consists of four key components: (a) *Multi-attributed target image generation*, which synthesizes pose-aligned target faces with semantically opposing attributes to guide embedding manipulation; (b) *Gaussian shell initialization*, where a secondary set of auxiliary Gaussians is introduced to parameterize controlled additive perturbations without modifying the baseline geometry; (c) *Multi-view adversarial training*, which aligns rendered views of the perturbed 3D model with target embeddings under stochastic augmentations for robustness; and (d) *Face-aware regularization mechanisms*, which constrain perturbations in identity-sensitive regions and suppress unnatural luminance artifacts. Together, these components enable semantically effective yet

structurally consistent adversarial manipulation of 3D faces in VLM embedding space.

4.1 Multi-Attributed Target Image Generation

To construct supervision for the adversarial objective defined in Sec. 3, we synthesize target images that reflect the desired semantic manipulation c_{tar}. Given an input rendering $R(\mathcal{G}_s, \pi)$, we first extract its pose representation using OpenPose [2], producing a structural pose map $\mathbf{P}$ that preserves geometric configuration while disentangling appearance. We then condition a pretrained text-to-image diffusion model on $\mathbf{P}$ together with a text prompt $\mathcal{T}'_{tar}$ encoding a set of desired facial attributes. This process is illustrated in Fig. 2. The target prompt is constructed using a multi-attribute composition strategy, where selected attributes are deliberately chosen to be semantically different to those of the original face. While the user may explicitly specify a subset of attributes to modify, the remaining facial attributes are automatically initialized with values opposite to the source description. This multi-attribute formulation encourages a stronger shift in the VLM embedding space, as aligning multiple semantic factors simultaneously produces a more pronounced deviation toward $\mathcal{T}_{tar}$ compared to single-attribute manipulation, analogous to multi-task supervision. Finally, the synthesized image I_{tar} serves as a semantic anchor for guiding the optimization of the Gaussian shell parameters.

Thus, we reformulate the VLM text-injection objective into a target image injection objective, replacing purely text-driven alignment with embedding-space supervision via a synthesized target image. If $x = R(\mathcal{G}_s; \pi)$ denote the original rendering and $x_\delta = R(\mathcal{G}_s + \delta; \pi)$ the perturbed rendering, then we re-formulate Eq. 1 as:

$$\min_{\delta} \ \mathcal{L}\big(f_\theta(x_\delta), f_\theta(I_{\text{tar}})\big) \quad \text{s.t.} \quad I_{\text{tar}} \neq x, \tag{2}$$

where $\mathcal{L}$ denotes a similarity measure in the VLM embedding space. The perturbation δ is optimized such that the embedding of the rendered adversarial image aligns with that of the target image, while remaining visually and geometrically consistent with the original face. This formulation of 3D FaceShell explicitly drives the VLM representation of the perturbed face toward a pose-consistent yet semantically divergent target embedding.

4.2 Gaussian Shell Initialization

For the adversarial noise training, we initialize from a pre-optimized 3DGS representation $\mathcal{G}_s$ consisting of N_{base} Gaussians while keeping all baseline parameters fixed throughout optimization to preserve geometry and identity. We instantiate a secondary set of N_{shell} auxiliary Gaussians, referred to as a *Gaussian shell* to parameterize the additive perturbation δ. Each Gaussian shell is initialized by randomly sampling the spatial mean of a baseline Gaussian and applying a small uniform perturbation $\epsilon \sim \mathcal{U}([-0.01, 0.01]^3)$, ensuring close spatial proximity to the original surface. The covariance, rotation, and spatial parameters of the shell Gaussians are frozen after initialization. Also, optimization is restricted to their appearance attributes, specifically color and opacity, which together define the

additive perturbation δ in the rendered space. This constrained parameterization enforces that adversarial manipulation operates purely through localized appearance and density modulation, yielding semantically effective yet geometrically non-destructive perturbations of the underlying 3D face.

4.3 Multi-View Adversarial Noise Training

At each optimization step of 3D FaceShell, the perturbed 3D representation $\mathcal{G}_s+\delta$ is rendered into K differentiable views $\{x_k\}_{k=1}^{K}$ using a predefined camera set $\mathit{\Pi} = \{\pi_k\}_{k=1}^{K}$, where each camera π_k is parameterized by azimuth θ_k, elevation ϕ_k, radius r, and field of view ψ. Formally,

$$x_k = R(\mathcal{G}_s + \delta; \pi_k) \tag{3}$$

The adversarial objective of 3D FaceShell is computed over a fixed training subset of $\mathit{\Pi}$ to enforce cross-view consistency of the perturbation in 3D space. We apply stochastic augmentations to each rendered view prior to encoding to prevent overfitting to specific viewpoints. Specifically, we apply a set of augmentations, denoted by $\mathcal{A}(\cdot)$, simulate 8-bit image discretization via quantization noise and apply random resized crops and color jittering. The augmented rendered views x_k and the synthetic multi-attribute target image I_{tar} are processed by the visual encoder $\text{VE}(\cdot)$ of the pretrained VLM f_θ to obtain their corresponding feature embeddings: $z_k = \text{VE}(\mathcal{A}(x_k))$ and $z_{\text{tar}} = \text{VE}(\mathcal{A}(I_{\text{tar}}))$. Then, we optimize δ to align these feature embeddings to guide the attribute transfer through image injection. In practice, the perturbation δ is optimized using features extracted from E different visual encoders, corresponding to the visual backbones of multiple VLMs. This multi-encoder supervision encourages consistent feature alignment across diverse representations, improving robustness and enabling the learned perturbations to generalize in a model-agnostic manner. Thus, the feature alignment loss aggregates over all training views as

$$\mathcal{L}_{\text{feature}} = \frac{1}{EK} \sum_{e=1}^{E} \sum_{k=1}^{K} \Big(1 - \text{sim}\big(z_k, z_{\text{tar}}\big)\Big), \tag{4}$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}$ denotes cosine similarity. The overall optimization jointly minimizes $\mathcal{L}_{\text{feature}}$ and regularization terms on the shell parameters, yielding semantically targeted yet geometrically consistent adversarial perturbations across viewpoints.

4.4 Face-Aware Regularization Mechanisms

Finally, we introduce face-aware regularization terms to preserve identity and photorealism of the underlying 3D face. While $\mathcal{L}_{\text{feature}}$ drives semantic embedding shifts, the following constraints suppress unnecessary perturbations and ensure that the learned noise remains localized and structurally consistent.

Landmark Consistency. In order to protect identity-critical regions, we extract fixed binary masks M corresponding to sensitive facial areas (e.g., *eyes*, *nose*, and *lips*) from the baseline renders. Let I_k and I_k^{base} denote the current and baseline renders at view π_k, respectively. We penalize pixel-level deviations within the masked regions across all training views:

$$\mathcal{L}_{\text{landmarks}} = \frac{1}{K} \sum_{k=1}^{K} \left\| M \odot (I_k - I_k^{\text{base}}) \right\|_2^2, \tag{5}$$

where $\odot$ denotes element-wise multiplication. This constraint discourages distortions in identity-sensitive areas while allowing controlled modifications elsewhere.

Chroma-Constrained Texture Regularization. Texture-based adversarial perturbations often manifest as high-frequency luminance artifacts. We mitigate this effect and encourage smooth semantic transfer by mapping the additive texture parameters of the auxiliary shell Gaussians directly into the YUV color space, emphasizing our regularization in the luminance channel. Let (Y_i, U_i, V_i) denote the YUV decomposition of the explicit texture noise applied to the i-th shell Gaussian, out of N_{shell} total auxiliary Gaussians. We define:

$$\mathcal{L}_{\text{chroma}} = \frac{1}{N_{\text{shell}}} \sum_{i=1}^{N_{\text{shell}}} \left(\lambda_Y Y_i^2 + U_i^2 + V_i^2 \right), \tag{6}$$

4.5 Overall Optimization of 3D FaceShell

The optimization of 3D FaceShell jointly balances semantic embedding manipulation and face-aware regularization in the 3D Gaussian shell parameter space as illustrated in Fig. 3. At each iteration, gradients from the multi-view feature alignment loss are backpropagated through the differentiable renderer to update only the shell parameters (color and opacity), while the baseline Gaussians remain frozen. The overall objective combines latent feature matching with structural and photometric constraints:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{feature}} + \lambda_{\text{chroma}} \mathcal{L}_{\text{chroma}} + \lambda_{\text{landmarks}} \mathcal{L}_{\text{landmarks}}. \tag{7}$$

where λ_{chroma} and $\lambda_{\text{landmarks}}$ are the strengths of landmark consistency and luminance suppression, respectively. Together, these losses constrain the adversarial optimization to produce semantically effective yet visually coherent perturbations, preserving facial identity and structural realism across viewpoints.

5 Experiments

5.1 Implementation Details

For the adversarial feature alignment in 3D FaceShell, we use a multi-encoder training with $E = 2$ distinct vision encoders: SigLIP (ViT-SO400M-Patch14) [40] and CLIP (ViT-Large-Patch14-336) [26]. For stable optimization, the weight

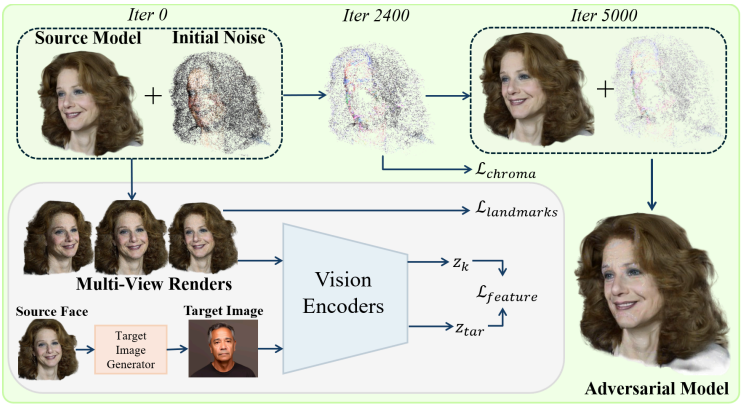

Fig. 3: The 3D FaceShell Optimization Pipeline. **3D Operations (Green):** A learnable Gaussian shell is initialized over a frozen 3DGS baseline and shell parameters are iteratively optimized and Gaussian color is regulated ($\mathcal{L}_{\text{chroma}}$). **2D Operations (Gray):** Rendered views are generated to guide feature alignment in shell Gaussians ($\mathcal{L}_{\text{feature}}$) and enforce identity preservation ($\mathcal{L}_{\text{landmarks}}$).

of the chroma-constrained texture regularization is dynamically annealed using a cosine progression schedule. If $p = \frac{i}{T}$ denote the normalized training progress at iteration i out of T total iterations, and define the cosine progression factor $C = \frac{1}{2}(1 - \cos(\pi p))$. The iteration-dependent weight is then given by $\lambda_{\text{chroma}} = \lambda_0 C$, where λ_0 is a constant scaling factor. Additionally, we further regularize the optimization and prevent opacity-driven geometric artifacts by imposing a progressively tightening upper bound on the shell Gaussian opacity. The maximum allowable opacity is dynamically clamped as $O_{\max} = -0.5 - 2.5C$. As training progresses, this schedule gradually pushes Gaussian shells toward increased transparency, suppressing perturbations that do not meaningfully contribute to the semantic objective. This progressive clamping acts as an additional geometric regularizer, ensuring that the adversarial signal remains minimally intrusive while preserving structural fidelity of the underlying 3D face.

Hyperparameters. Unless otherwise specified, rendered images are generated at a resolution of 384×384 with a field of view $\psi = 50°$ and a fixed camera radius $r = 2.7$. We use $K = 5$ fixed camera poses for optimization: frontal $(270°, 0°)$, slight left $(255°, 0°)$, slight right $(285°, 0°)$, slight up $(270°, 15°)$, and slight down $(270°, -15°)$. To evaluate cross-view generalization, we introduce a novel validation view at $(285°, 15°)$, which is excluded from optimization and used exclusively for evaluation. We optimize the color and opacity parameters of $N_{shell} = 50,000$ shell Gaussians using the Adam optimizer for 10,000 total iterations, setting the learning rate to 0.1 for both parameter groups. For our face-aware regularization mechanisms, the landmark consistency loss weight is fixed at $\lambda_{\text{landmarks}} = 500$. In the chroma-constrained texture regularization, the luminance penalty is enforced with $\lambda_Y = 10.0$ while we scale our overall chroma-constrained texture loss with $\lambda_{\text{chroma}} = 0.4$. Finally, during the multi-

view adversarial noise training, the stochastic augmentations $\mathcal{A}(\cdot)$ applied to the rendered views consist of simulated 8-bit quantization noise (uniformly sampled between $\pm$ 2/255), a random resized crop with a scale factor between 0.8 and 1.0, and random color jittering adjusting brightness and contrast by up to 10%.

5.2 Evaluation Settings

Dataset. In order to evaluate our method and the competing baselines, we construct a benchmark consisting of 100 celebrity faces curated from the IMDB-WIKI dataset [27]. For each subject, the facial attributes *eye color*, *hair color*, *hair type*, and *expression* are manually annotated, while *age*, *gender*, and *ethnicity* are obtained from publicly available metadata corresponding to the time the source photograph was captured. Each 2D face image is then converted into a 3D Gaussian Splatting (3DGS) representation using FaceLift [18]. The resulting benchmark, comprising 3D faces paired with curated attribute annotations, will be publicly released to facilitate further research in this direction.

Metrics. We assess the capabilities of each attack method to successfully transfer target attributes onto rendered views of each source 3D face avatar. We provide the rate at which a given VLM outputs target attributes when provided with a rendered view of the source face (Injection Rate) and the rate at which a given VLM outputs an attribute different from the ground truth attribute of a given source face when provided with a rendered view of that source face (Mismatch Rate). Importantly, we assess the quality of the outputs for each method using LPIPS [44] to measure the perceptual distance between the synthesized images and the ground truth references. Finally, we utilize ArcFace [4] to determine the identity similarity between the generated face views produced by each method and the faces in the ground truth images.

Selection Criteria. To determine the optimal stopping point for our adversarial optimization and ensure consistent perceptual fidelity, we evaluate the training progression on a validation set consisting of 20 separate faces. As illustrated in Fig. 4, the optimization process reveals a clear trade-off between the VLM injection accuracy and the perceptual quality (LPIPS) over the course of 10,000 training iterations. To ensure our method balances high visual fidelity with maximum semantic manipulation, we apply a strict threshold-based selection criterion for our final evaluations. For each source face, we extract the rendering from the iteration that yields the highest injection accuracy, strictly bottlenecked by the requirement that the LPIPS score must remain below 0.165. In edge cases where the optimization fails to produce any samples under this 0.165 threshold, we default to selecting the iteration that achieved the absolute lowest LPIPS score overall. This guarantees that the final perturbed 3D avatars remain perceptually grounded and structurally consistent, preventing

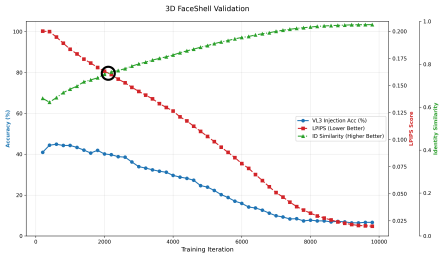

Fig. 4: Validation results on VideoLLaMA3 used to select the best-performing samples during optimization.

unbounded visual degradation while still capturing the peak adversarial transfer rate.

Models. We evaluate all methods on four black-box VLMs: VideoLLaMA3 [41], LLaVA-NeXT [12], LLaVA-OV [11], and BLIP-2 [13]. These models are selected due to their strong multimodal reasoning capabilities and the diversity of their visual encoders. In particular, BLIP-2 employs a CLIP-based visual encoder, whereas the others utilize SigLIP-based visual encoders.

Baselines. As there are no other methods for adversarial semantic editing in the 3DGS space, we benchmark 3D FaceShell against six state-of-the-art 2D adversarial attacks targeting VLMs: AnyAttack [43], V-Attack Ensemble [21], AdvDiffVLM [8], SSA-CWA [5], M-Attack [15], and AttackVLM [45]. We selected these specific baselines because they represent a comprehensive cross-section of the most effective modern 2D attack paradigms, including diffusion-based, self-supervised, and highly transferable embedding-space perturbations.

5.3 Quantitative Results

State of the Art Comparison. As shown in Table 1, 3D FaceShell achieves competitive performance in terms of both Injection Rate (IR) and Mismatch Rate (MR) across all evaluated vision-language models. While aggressive 2D baselines such as M-Attack and V-Attack obtain slightly higher mismatch rates overall and higher injection rates on certain models (e.g., BLIP-2 and LLaVA-NeXT), the semantic performance gap remains small. For instance, 3D FaceShell trails the strongest baseline (M-Attack) by only a 3.9% relative difference in LLaVA-OV MR and by 10.3% relative difference compared to V-Attack on VideoLLaMA3 MR. Moreover, 3D FaceShell surpasses both M-Attack and V-Attack in absolute Injection Rate on VideoLLaMA3 and LLaVA-OV. Importantly, the marginal semantic gains of these aggressive 2D approaches come at the cost of significant degradation in *visual quality* and *identity preservation*. In contrast, 3D FaceShell maintains substantially higher perceptual and structural fidelity, achieving a $+71.5\%$ improvement in LPIPS compared to V-Attack and a $+60.7\%$ improvement over M-Attack. Similarly, 3D FaceShell improves Identity Similarity by $+88.9\%$ and $+56.2\%$ relative to V-Attack and M-Attack, respectively. This distinction is critical, as preserving facial identity and geometric structure is essential for realistic and minimally intrusive adversarial manipulation.

Multi-Attribute vs Single-Attribute Transfer. Table 2 presents a quantitative comparison of Transfer Rate and Mismatch Rate on VideoLLaMA3 when modifying all targeted semantic attributes simultaneously versus isolating a single attribute and keeping the rest constant. Our results indicate that shifting all attributes at once yields superior performance for most individual attributes. For instance, transferring all attributes simultaneously results in a $+50.7\%$ relative increase in the transfer rate for hair color and a $+96.7\%$ relative increase in the transfer rate for expression compared to isolated single-attribute attacks. Additionally, Gender and Ethnicity transfer rates see relative increases of $+3.2\%$ and $+9.7\%$, respectively. This result confirms that inducing a compounded semantic shift via multi-attribute target image injection provides a stronger gradient signal, thereby improving the attack success rate across individual attributes.

Table 1: Quantitative comparison of Injection Rate (IR) and Mismatch Rate (MR) across multiple vision-language models, alongside perceptual (LPIPS) and structural (Identity Similarity) preservation metrics across frontal face views.

Method	VideoLLaMA3 (%)		LLaVA-NeXT (%)		LLaVA-OV (%)		BLIP-2 (%)		LPIPS ↓	ID Sim ↑
	IR	MR	IR	MR	IR	MR	IR	MR		
Nothing	5.9	19.3	4.6	17.1	4.4	15.7	13.1	39.7	0.0000	1.0000
Random Noise	8.1	22.5	6.9	21.6	6.9	20.6	12.5	39.2	0.0862	0.9535
AnyAttack [43]	6.6	22.9	5.7	21.1	6.4	19.3	13.6	41.1	0.4452	0.7795
AttackVLM [45]	8.9	24.7	7.1	21.0	7.6	21.3	11.4	38.3	0.2428	**0.8506**
AdvDiffVLM [8]	6.9	23.6	6.7	22.1	5.9	21.9	12.1	40.6	0.2638	0.6526
SSA-CWA [5]	33.9	57.0	21.7	41.6	32.0	50.4	37.7	68.3	0.4626	0.5513
V-Attack [21]	44.4	**74.0**	35.7	63.4	51.6	75.1	33.9	66.6	0.4039	0.5255
M-Attack [15]	47.4	69.6	**46.3**	**70.3**	56.4	**77.4**	**37.9**	**70.3**	0.3814	0.4884
3D FaceShell	**48.7**	66.4	31.3	53.0	**56.9**	74.4	26.7	53.7	**0.1499**	0.7629

Table 2: Quantitative comparison of Transfer Rate (TR) and Mismatch Rate (MR) across training and validation views for each targeted semantic attribute on VideoLLaMA3.

Method	Age (%)		Gender (%)		Ethnicity (%)		Expression (%)		Hair (%)		Hair Color (%)		Eye Color (%)	
	IR	MR	IR	MR	IR	MR	IR	MR	IR	MR	IR	MR	IR	MR
3D FaceShell	40.8	83.3	64.1	64.1	34.0	68.3	17.9	49.7	37.6	60.2	21.1	72.5	50.4	70.3
Individual	35.6	82.0	62.1	62.1	31.0	69.1	9.1	50.0	31.8	63.1	14.0	58.7	57.3	42.7

Table 3: Quantitative comparison between training-view results and novel validation-view results.

Method	VideoLLaMA3 (%)		LLaVA-NeXT (%)		LLaVA-OV (%)		BLIP-2 (%)		LPIPS ↓	ID Sim ↑
	IR	MR	IR	MR	IR	MR	IR	MR		
Ours – Training Views	**46.8**	**65.2**	**30.3**	**52.8**	**56.0**	**73.5**	**26.2**	**53.2**	0.1523	0.7550
Ours – Validation View	36.4	56.3	23.6	45.0	41.1	59.1	23.3	51.1	**0.1465**	0.7289
Front View Training – Validation View	23.4	42.0	13.7	33.0	26.6	44.1	17.1	45.3	0.1474	**0.7596**

Commercial Model Evaluation. Table 4 extends our primary evaluation to the commercial Gemini-2.5-Flash [29] VLM, comparing our method against our 2D adversarial attack baselines. Our 3D FaceShell successfully generalizes to this closed-source model, achieving a competitive injection rate and mismatch rates. Crucially, while our method observes a relatively low difference in semantic manipulation metrics, falling short of V-Attack's injection rate by a relative 13.4% and M-Attack's by a relative 28.3%, it yields a disproportionately massive improvement in visual fidelity. Specifically, 3D FaceShell achieves an LPIPS score of 0.1499, representing an impressive 62.9% relative improvement over V-Attack and a 60.7% improvement over M-Attack. Similarly, 3D FaceShell achieves an identity similarity score of 0.7629, structurally outperforming V-Attack [21] by 45.2% and M-Attack [15] by 56.2%. This demonstrates that 3D FaceShell maintains semantic competitiveness without relying on the severe, unrealistic face degradation characteristic of 2D baselines. It should be noted that we additionally tried to evaluate each of the methods on GPT-4o [23] but this model largely

refused to answer questions about face attributes despite prompt engineering efforts.

Cross-View Generalization. We assess the necessity of our multi-view adversarial noise training in Table 3. When comparing our multi-view training strategy against a model trained exclusively on a single frontal view, it is evident that multi-view training contributes heavily to successful transfer onto novel views. The model trained only on the front view experiences a severe drop in performance on the validation view, whereas our multi-view approach maintains robust generalization. Specifically, multi-view adversarial noise training achieves a +55.6% higher relative Injection Rate on the unseen validation view for VideoLLaMA3 compared to the single-view baseline.

Table 4: Quantitative comparison of Injection Rate (IR) and Mismatch Rate (MR) on Gemini, along with perceptual (LPIPS) and identity preservation (ID Sim) metrics across frontal face views.

Method	Gemini (%)		LPIPS ↓	ID Sim ↑
	IR	MR		
Nothing	2.9	15.4	0.0000	1.0000
Random Noise	4.4	19.3	0.0862	0.9535
AnyAttack [43]	4.0	20.7	0.4452	0.7795
AttackVLM [45]	5.4	23.6	0.2428	**0.8506**
AdvDiffVLM [8]	3.9	21.7	0.2638	0.6526
SSA-CWA [5]	26.0	48.3	0.4626	0.5513
V-Attack [21]	35.9	**76.0**	0.4039	0.5255
M-Attack [15]	**43.4**	67.3	0.3814	0.4884
3D FaceShell	31.1	50.4	**0.1499**	0.7629

5.4 Qualitative Results

Fig. 5 displays the large improvement in identity preservation and visual clarity of our method compared to state-of-the-art 2D methods of semantic shifting. While baseline 2D attacks introduce noticeable pixelated noise, severe color distortions, or structural warping that immediately compromise the photorealism of the face, 3D FaceShell introduces far less intrusive noise. By confining the optimization to our learnable Gaussian shell and heavily regularizing the perturbations, our approach ensures the rendered outputs look naturally persistent. The qualitative comparisons confirm that 3D FaceShell successfully balances the dual objectives of manipulating VLM semantic interpretations while leaving the underlying human-recognizable geometry and textures perceptually undisturbed.

Table 5: Ablation study of loss functions and augmentations evaluated across all training and validation views.

Method	VideoLLaMA3 (%)		LLaVA-NeXT (%)		LLaVA-OV (%)		BLIP-2 (%)		LPIPS ↓	ID Sim ↑
	IR	MR	IR	MR	IR	MR	IR	MR		
No Augmentation	17.6	34.7	16.1	33.9	28.8	45.3	15.6	42.8	**0.1438**	**0.8397**
No Chroma Constraint	53.4	**72.5**	**43.1**	**66.7**	**61.7**	**80.4**	**34.5**	**61.9**	0.1911	0.5695
No Landmark Consis.	**54.6**	72.2	**43.1**	66.0	59.7	77.5	28.8	56.1	0.1512	0.6184
No Clamp	44.6	65.0	31.5	53.5	57.0	76.3	27.7	55.1	0.1894	0.7098
3D FaceShell	45.0	63.7	29.2	51.5	53.5	71.1	25.7	52.9	0.1513	0.7507

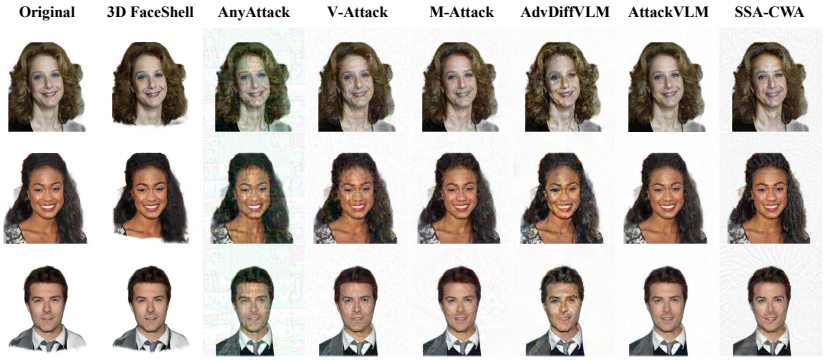

Fig. 5: 3D FaceShell achieves superior perceptual fidelity over baselines. While state-of-the-art 2D attacks introduce severe pixelated noise, color distortion, and structural warping to shift VLM perception, our method produces visually inconspicuous perturbations that successfully preserve original photorealism and facial identity.

5.5 Ablation Studies

Impact of Losses and Augmentation. In Table 5 we conduct an ablation study analyzing the impact of our loss functions and augmentations. Removing the data augmentations leads to a drastic drop in both Injection Rate and Mismatch Rate across all VLMs. For instance, causing a 60.9% relative drop in VideoLLaMA3 Injection Rate and a 46.2% relative drop in LLaVA-OV Injection Rate. This demonstrates that augmentations are strictly necessary to prevent the adversarial shell from overfitting to specific rendered views. Furthermore, removing the chroma loss ($\mathcal{L}_{\text{chroma}}$) or the landmark loss ($\mathcal{L}_{\text{landmarks}}$) significantly degrades the incredibly important perceptual quality and identity similarity of the rendered images compared to the source images. Quantitatively, disabling the chroma loss causes a 24.1% relative drop in identity similarity and a 26.3% decrease in perceptual similarity. Similarly, disabling the landmark loss drops identity similarity by 17.6%. This confirms that the face-aware regularization mechanisms are vital for achieving the optimal balance between a high rate of semantic transfer and identity preservation. Finally, omitting dynamic opacity clamping ("No Clamp") yields marginal semantic gains, such as a relative 7.9% increase in LLaVA-NeXT Injection Rate, but introduces unacceptable visual artifacts, causing a 25.2% relative reduction in perceptual similarity and a 5.4% drop in identity similarity.

Table 6: Quantitative comparison of 3D FaceShell against variants trained using a single visual encoder.

Method	VideoLLaMA3 (%)		LLaVA-NeXT (%)		LLaVA-OV (%)		BLIP-2 (%)		LPIPS ↓	ID Sim ↑
	IR	MR	IR	MR	IR	MR	IR	MR		
3D FaceShell (SigLIP)	42.6	61.4	19.0	39.2	53.0	70.9	20.4	47.9	0.1484	**0.7830**
3D FaceShell (CLIP)	21.6	39.7	25.5	45.7	18.9	35.4	21.1	48.8	**0.1306**	0.7789
3D FaceShell	**45.0**	**63.7**	**29.2**	**51.5**	**53.5**	**71.1**	**25.7**	**52.9**	0.1513	0.7507

Multi-Encoder Supervision. Additionally, Table 6 highlights the benefit of utilizing multiple visual encoders during the feature alignment step. Utilizing multiple encoders in our training architecture produces more effective overall semantic attribute shifting across the black-box VLMs than using either SigLIP or CLIP in isolation. For example, ensembling the models yields a +108.3% relative increase in VideoLLaMA3 Injection Rate compared to using CLIP alone, and a +26.0% relative increase in BLIP-2 Injection Rate compared to using SigLIP alone. This demonstrates that ensembling diverse representation spaces during optimization prevents the Gaussian shell from overfitting to a single model's feature space, ultimately yielding more robust and transferrable adversarial perturbations.

6 Conclusion

In this work, we introduced 3D FaceShell, a novel privacy defense framework designed to protect 3D face avatars from unauthorized semantic inference by VLMs. Our method remains robust through multiple rendered viewpoints by operating directly in the 3D Gaussian space and successfully transfers face attribute semantics without significantly degrading visual clarity or destroying the identity of the source face. By introducing a heavily regularized, learnable auxiliary Gaussian shell, we successfully inject perturbations that manipulate VLM understanding while leaving the underlying baseline geometry and identity-defining appearance completely untouched.

Our extensive evaluations across multiple state-of-the-art black-box VLMs demonstrate that 3D FaceShell achieves highly competitive targeted attribute transfer and mismatch rates. Crucially, it accomplishes this while maintaining significantly higher perceptual quality and identity similarity than current state-of-the-art 2D semantic manipulation baselines. Ablation studies further validate that our multi-attribute synthesis, multi-view adversarial training, and face-aware regularization mechanisms are strictly essential for achieving robust, view-consistent semantic steering.

While our current framework successfully defends static 3D face avatars, a promising direction for future work involves extending this mechanism into the 4D Gaussian Splatting space. Adapting our learnable auxiliary shell to maintain temporal consistency could safeguard fully animatable and dynamic portrait avatars [28] against continuous semantic inference in video formats.

Acknowledgments

This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant No. 2601332. This work is also supported in part by the National Science Foundation (IIS-2245652). We would also like to thank Charlotte Vision Lab members for our valuable discussions.

References

1. AlDahoul, N., Tan, M.J.T., Kasireddy, H.R., Zaki, Y.: Exploring vision language models for facial attribute recognition: Emotion, race, gender, and age (2024), https://arxiv.org/abs/2410.24148
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 172–186 (2021). https://doi.org/10.1109/TPAMI.2019.2929257
3. Chaubey, A., Guan, X., Soleymani, M.: Face-llava: Facial expression and attribute understanding through instruction tuning. IEEE/CVF Winter Conference on Applications of Computer Vision 2026
4. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (Oct 2022). https://doi.org/10.1109/tpami.2021.3087709, http://dx.doi.org/10.1109/TPAMI.2021.3087709
5. Dong, Y., Chen, H., Chen, J., Fang, Z., Yang, X., Zhang, Y., Tian, Y., Su, H., Zhu, J.: How robust is google's bard to adversarial image attacks? (2023), https://arxiv.org/abs/2309.11751
6. Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., Wang, Z.: Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ FPS. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=6AeIDnrTN2
7. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
8. Guo, Q., Pang, S., Jia, X., Liu, Y., Guo, Q.: Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models (2024), https://arxiv.org/abs/2404.10335
9. Horváth, A., Józsa, C.M.: Targeted adversarial attacks on generalizable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3718–3727 (October 2023)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
11. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), https://arxiv.org/abs/2408.03326
12. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models (2024), https://arxiv.org/abs/2407.07895
13. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
14. Li, Y., Li, Y., Dai, X., Guo, S., Xiao, B.: Physical-world optical adversarial attacks on 3d face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24699–24708 (2023)
15. Li, Z., Zhao, X., Wu, D.D., Cui, J., Shen, Z.: A frustratingly simple yet highly effective attack baseline: Over 90

16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
17. Lou, T., Jia, X., Liang, S., Liang, J., Zhang, M., Xiao, Y., Cao, X.: 3d gaussian splatting driven multi-view robust physical adversarial camouflage generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28752–28762 (2025)
18. Lyu, W., Zhou, Y., Yang, M.H., Shu, Z.: Facelift: Learning generalizable single image 3d face reconstruction from synthetic heads. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12691–12701 (October 2025)
19. Naderi, H., Bajić, I.V.: Adversarial attacks and defenses on 3d point cloud classification: A survey. IEEE Access **11**, 144274–144295 (2023)
20. Narayan, K., Vibashan, V., Patel, V.M.: Facexbench: Evaluating multimodal llms on face understanding. IEEE Transactions on Biometrics, Behavior, and Identity Science (2026)
21. Nie, S., Zhang, J., Yan, J., Shan, S., Chen, X.: V-attack: Targeting disentangled value features for controllable adversarial attacks on lvlms (2025), https://arxiv.org/abs/2511.20223
22. Niu, Z., Ren, H., Gao, X., Hua, G., Jin, R.: Jailbreaking attack against multimodal large language model (2024), https://arxiv.org/abs/2402.02309
23. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
24. Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., Mittal, P.: Visual adversarial examples jailbreak aligned large language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 21527–21536 (2024)
25. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024)
26. 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 on machine learning. pp. 8748–8763. PmLR (2021)
27. Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision **126**(2-4), 144–157 (2018)
28. Taubner, F., Zhang, R., Tuli, M., Lindell, D.B.: Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5318–5330 (June 2025)
29. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: A family of highly capable multimodal models. arxiv 2023. arXiv preprint arXiv:2312.11805 (2024)
30. Tömekçe, B., Vero, M., Staab, R., Vechev, M.: Private attribute inference from images with vision-language models. Advances in Neural Information Processing Systems **37**, 103619–103651 (2024)
31. Wang, M., Zhou, J., Li, T., Meng, G., Chen, K.: A survey on physical adversarial attacks against face recognition systems. Neurocomputing p. 132485 (2025)
32. Wilson, E., Bindschaedler, V., Jörg, S., Sheikholeslam, S., Butler, K., Jain, E.: Towards privacy-preserving photorealistic self-avatars in mixed reality (2025), https://arxiv.org/abs/2507.22153

33. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
34. Wu, T., Yuan, Y.J., Zhang, L.X., Yang, J., Cao, Y.P., Yan, L.Q., Gao, L.: Recent advances in 3d gaussian splatting (2024), https://arxiv.org/abs/2403.11134
35. Xiang, C., Qi, C.R., Li, B.: Generating 3d adversarial point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
36. Yang, X., Liu, C., Xu, L., Wang, Y., Dong, Y., Chen, N., Su, H., Zhu, J.: Towards effective adversarial textured 3d meshes on physical face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4119–4128 (2023)
37. Yin, Z., Cao, Y., Liu, H., Wang, T., Chen, J., Ma, F.: Towards robust multimodal large language models against jailbreak attacks (2025), https://arxiv.org/abs/2502.00653
38. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
39. Zeybey, A., Ergezer, M., Nguyen, T.: Gaussian splatting under attack: Investigating adversarial noise in 3d objects (2024), https://arxiv.org/abs/2412.02803
40. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision (2023)
41. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., Zhao, D.: Videollama 3: Frontier multimodal foundation models for image and video understanding (2025), https://arxiv.org/abs/2501.13106
42. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Chen, K., Qin, M., Li, Y., Wang, H.: Hravatar: High-quality and relightable gaussian head avatar. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
43. Zhang, J., Ye, J., Ma, X., Li, Y., Yang, Y., Chen, Y., Sang, J., Yeung, D.Y.: Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
45. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems **36**, 54111–54138 (2023)
46. Zhou, F., Zhou, Q., Ling, H., Lu, X.: Adversarial attacks on both face recognition and face anti-spoofing models (2025), https://arxiv.org/abs/2405.16940