DRAGOMETRY: 3D GEOMETRIC CONTROL FOR DRAG-BASED FACIAL IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Point-based drag editing provides an intuitive interface for image manipulation, but existing diffusion-based methods typically operate on learned intermediate representations in which geometric information is encoded only implicitly. Geometry-conditioned diffusion models offer a different setting: explicit structured geometry is already exposed as part of the generative condition. We therefore propose DRAGOMETRY, a point-based editing framework that performs dragging directly in the geometry condition space rather than through learned feature correspondence. We instantiate the framework on MMDM with FLAME for facial editing. User specified 2D handles are associated with persistent 3D surface points, allowing drag constraints to be resolved through different geometric variables. For local deformation, we optimize FLAME expression and jaw parameters and introduce Jacobian-based parameter weighting to favor parameters relevant to the requested motion while suppressing unwanted deformation elsewhere. For viewpoint edits, we instead optimize camera parameters while preserving the underlying facial geometry. We further introduce DragFace, a benchmark containing both facial deformation and rotation edits, together with Geometry Mean Distance, which evaluates drag accuracy using explicit geometric correspondence. Experiments show that our method achieves more accurate geometric point control, more effective facial deformation, and substantially larger coherent viewpoint changes than existing diffusion-based drag editing methods.

1 INTRODUCTION

Drag-based image editing provides an intuitive interface for controllable generation, where a user specifies one or more handle points and corresponding target locations and expects the selected content to move accordingly while the remaining image is preserved. Existing methods typically realize this control through motion supervision and iterative point tracking in learned representations of generative models (Pan et al., 2023; Shi et al., 2024; Mou et al., 2024; Liu et al., 2024; Zhang et al., 2025; Xia et al., 2025). Although these 2D methods optimize different variables, they largely rely on learned representations in which semantic and spatial information is encoded implicitly. Appropriate intermediate features must therefore be selected from particular network layers or diffusion stages for motion supervision, feature matching, or point tracking (Pan et al., 2023; Shi et al., 2024; Mou et al., 2024; Liu et al., 2024).

The reliance on learned internal representations introduces several practical challenges, including sensitivity to feature selection, inaccurate point tracking, and unstable or incomplete motion supervision, which subsequent methods have attempted to alleviate (Ling et al., 2024; Cui et al., 2024; Zhang et al., 2025). To reduce this dependence, recent approaches introduce external 3D geometry, such as reconstructed meshes, to guide generic diffusion models (Koo et al., 2025). Geometry-conditioned diffusion models, however, provide a fundamentally different setting: image generation is already controlled by explicit geometric conditions, such as FLAME/NPHM (Li et al., 2017; Giebenhain et al., 2023) for faces, or SMPL/SMPL-X (Loper et al., 2015; Pavlakos et al., 2019) for human bodies. Unlike geometry implicitly encoded in learned features or introduced as an auxiliary representation, these structured conditions explicitly describe the generated geometry and provide known correspondence across geometric configurations. This motivates us to perform drag editing directly in the geometry condition space rather than recovering and manipulating geometric information indirectly through diffusion representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

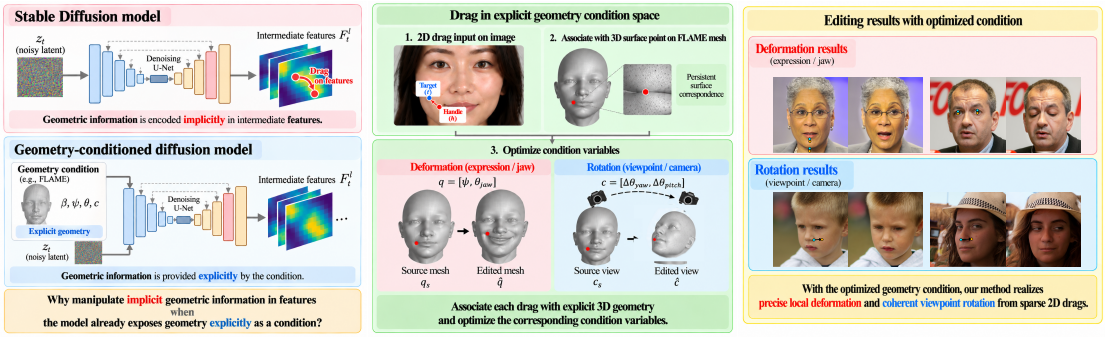

Figure 1: **Motivation and overview of DRAGOMETRY.** Left: geometric information is implicit in diffusion features but explicit in the geometry condition. Middle: we associate each 2D drag with a persistent 3D surface point and optimize deformation or viewpoint variables. Right: the optimized condition produces local facial deformation and coherent viewpoint rotation.

More importantly, directly applying conventional feature-based drag to geometry-conditioned diffusion models is problematic. Intermediate diffusion features are not guaranteed to preserve the pointwise correspondence defined by the underlying geometry, especially when the geometric condition changes. Moreover, manipulating an intermediate feature while keeping the geometry condition fixed creates an inconsistency: subsequent denoising steps remain guided by the original geometry and can counteract the intended edit. We illustrate the correspondence issue using MMDM (Taubner et al., 2025b), a FLAME-conditioned facial diffusion model. As shown in Fig. 2, despite the explicit geometric correspondence from H to T, the nearest feature match F remains closer to H than to T. This motivates performing drag editing directly in the geometry condition space, where correspondence is explicit and the subsequent denoising process is conditioned consistently on the edited geometry.

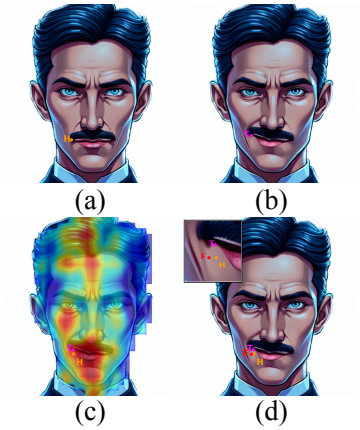

Figure 2: **Pointwise feature correspondence.** (a,b) Source and target under different geometry conditions, with H and T denoting corresponding facial points. (c) Feature similarity to the source feature at H. (d) The nearest feature match F remains closer to H than the geometrically corresponding point T, showing unreliable pointwise correspondence.

A sparse 2D drag may correspond to different underlying transformations, such as local deformation or 3D viewpoint rotation. Conventional 2D drag methods apply the same image-space manipulation regardless of this ambiguity and therefore do not explicitly control which transformation is realized. Geometry-conditioned diffusion models, in contrast, are built on 3D geometric representations whose parameters separately describe deformation and viewpoint, providing a natural way to resolve this ambiguity directly in the condition space.

To address these challenges, we perform point-based editing directly in the explicit geometry condition space, and instantiate our framework on MMDM (Taubner et al., 2025b) as a representative geometry-conditioned diffusion model. For local facial deformation, we optimize the FLAME expression and jaw parameters to move the projected geometric points toward the user specified targets. Since FLAME parameters are entangled, parameters that effectively move the handles may also introduce unwanted deformation elsewhere. We therefore introduce Jacobian-based parameter weighting to identify parameters that support the requested motion while suppressing those that strongly affect unrelated facial regions. For rotation edits, we instead optimize the viewpoint condition while keeping the underlying facial geometry fixed.

We further introduce DragFace, a benchmark containing both deformation and rotation edits. Since conventional Mean Distance (MD) relies on feature matching and may not reflect the motion of the intended geometric point, we propose Geometry MD, which evaluates drag accuracy using explicit geometric correspondence while discounting global face translation. Experiments show that

our method more accurately realizes the requested geometric edits, producing stronger expression changes for deformation and substantially larger, coherent viewpoint changes for rotation than existing diffusion-based drag editing methods.

Our contributions are summarized as follows:

- We introduce a point-based editing framework for geometry-conditioned diffusion models that performs drag editing directly in the explicit geometry condition space, rather than relying on learned feature correspondence. We instantiate the framework on MMDM for facial editing.
- We introduce a Jacobian-based parameter weighting strategy that favors deformation parameters relevant to the requested motion while suppressing parameters that induce excessive changes in unrelated regions.
- We introduce DragFace, a benchmark covering facial deformation and rotation edits, together with Geometry Mean Distance, which evaluates drag accuracy through explicit geometric correspondence while discounting global translation.

2 Related work

Point-based drag editing. DragGAN (Pan et al., 2023) introduced point-based image manipulation through feature-based motion supervision and point tracking. Diffusion-based methods follow this paradigm by exploiting intermediate representations for motion supervision and correspondence (Shi et al., 2024; Mou et al., 2024; Liu et al., 2024). Since these learned representations may provide ambiguous or unstable correspondence, subsequent methods improve feature selection, tracking, and optimization stability (Ling et al., 2024; Cui et al., 2024; Zhang et al., 2025), while some additionally use masks to localize the edit (Shi et al., 2024; Hou et al., 2024). Drag interaction has also been extended directly to 3D representations (Guang et al., 2025; Dong & Wang, 2025; Qu et al., 2025), while FlowDrag (Koo et al., 2025) introduces reconstructed 3D geometry to guide dragging in a generic image diffusion model. However, existing diffusion-based approaches operate on models where geometry is not an explicit generative condition. In contrast, we study point-based editing for geometry-conditioned diffusion models, where structured 3D geometry is directly provided as the conditioning signal.

Geometry-conditioned diffusion models. Explicit 3D representations are increasingly used to provide structured control in diffusion models. For faces and heads, 3DMM/FLAME or neural parametric head models control expression, pose, viewpoint, and appearance (Ding et al., 2023; Lan et al., 2024; Kirschstein et al., 2024; Taubner et al., 2025b;a), while SMPL or SMPL-X conditions support controllable human generation and reconstruction (Buchheim et al., 2025; Ho et al., 2024; Li et al., 2025). These works mainly use geometry as a prescribed condition for synthesis, animation, or reconstruction. In contrast, we treat the geometry condition itself as the editable space and optimize it from sparse handle-target constraints for point-based editing.

3 DRAGOMETRY

3.1 Preliminary

Morphable Multi-view Diffusion Model. We build our method on MMDM (Taubner et al., 2025b), which generates facial images conditioned on reference appearance and spatial maps derived from FLAME geometry and camera view. These conditions are injected throughout denoising to control the generated facial geometry and viewpoint. We denote the generation process as $\mathbf{I} = \mathcal{G}(\mathbf{z}, \mathcal{C}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{c}))$, where $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, $\boldsymbol{\theta}$, and $\mathbf{c}$ denote shape, expression, pose, and camera parameters, respectively, $\mathbf{z}$ is the diffusion latent, and $\mathcal{C}(\cdot)$ denotes the geometry-derived MMDM condition. Keeping the pretrained diffusion model fixed, our goal is to infer edits to this structured condition from sparse handle-target constraints rather than manipulating the diffusion representation directly.

3.2 DRAGOMETRY

Given a source image I_s and N handle-target pairs $\{(\mathbf{h}_i, \mathbf{t}_i)\}_{i=1}^{N}$, DRAGOMETRY performs point-based editing by directly optimizing the geometry condition of MMDM. We first recover the source FLAME and camera parameters $(\boldsymbol{\beta}_s, \boldsymbol{\psi}_s, \boldsymbol{\theta}_s, \mathbf{c}_s)$. FLAME additionally defines a fixed-topology

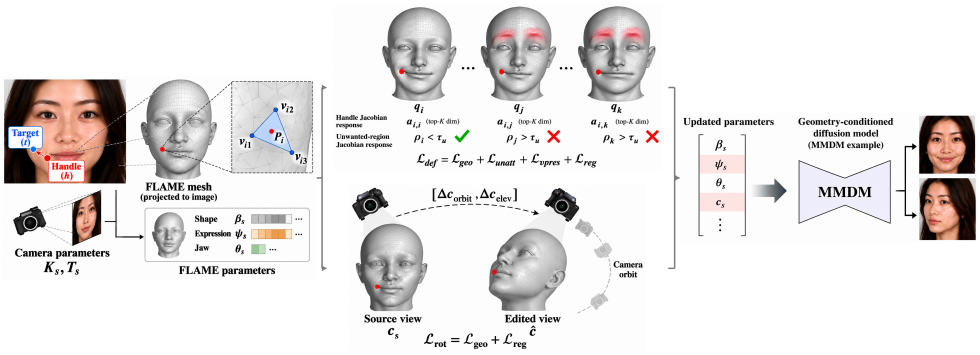

Figure 3: Overview of DRAGOMETRY. Given a source image and sparse handle-target pairs, we recover the FLAME geometry and camera parameters and associate each handle with a persistent 3D surface point. For deformation, Jacobian-based weighting selects deformation parameters that move the handle toward its target while suppressing unwanted changes in distant regions. For rotation edits, the facial geometry is fixed and only the camera viewpoint is optimized. The resulting geometry condition is then passed to the geometry-conditioned diffusion model to synthesize the edited image.

mesh $\mathbf{V} = \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta})$, which provides consistent surface correspondence across parameter configurations. We use this correspondence to associate each 2D handle with a persistent facial surface point. Depending on the requested edit, we then optimize either the deformation or viewpoint parameters so that the projected surface points approach their targets. The optimized condition is finally passed to MMDM to synthesize the edited image.

Geometry initialization. We follow the CAP4D preprocessing pipeline (Taubner et al., 2025b), using Pixel3DMM (Giebenhain et al., 2026) and FlowFace (Taubner et al., 2024) to recover the source FLAME parameters and camera configuration $(\boldsymbol{\beta}_s, \boldsymbol{\psi}_s, \mathbf{K}_s, \mathbf{T}_s)$. During editing, the shape $\boldsymbol{\beta}_s$ and camera intrinsics $\mathbf{K}_s$ remain fixed. We optimize the expression and jaw parameters for deformation edits and the viewpoint derived from $\mathbf{T}_s$ for rotation edits.

Handle-surface association. For each handle $\mathbf{h}_i$, we project the fitted FLAME mesh using $(\mathbf{K}_s, \mathbf{T}_s)$ and locate the projected triangle τ_i containing it. Let k_{ij} and $\mathbf{p}_{ij}$, $j \in \{1, 2, 3\}$, denote its vertex indices and projected vertices. We compute barycentric weights λ_{ij} such that $\mathbf{h}_i = \sum_{j=1}^{3} \lambda_{ij} \mathbf{p}_{ij}$, $\sum_{j=1}^{3} \lambda_{ij} = 1$. Keeping the triangle and barycentric weights fixed defines a persistent surface point $\mathbf{P}_i = \sum_{j=1}^{3} \lambda_{ij} \mathbf{V}_{k_{ij}}$, which provides geometric correspondence throughout both deformation and rotation editing.

3.2.1 Local deformation

For local deformation, we optimize $\mathbf{q} = [\boldsymbol{\psi}, \theta_{\text{jaw}}] \in \mathbb{R}^{66}$, consisting of the 65 FLAME expression coefficients and jaw opening parameter. For each persistent surface point $\mathbf{P}_i(\mathbf{q})$, we denote its projection under the fixed source camera as $\mathbf{p}_i(\mathbf{q}) = \Pi(\mathbf{P}_i(\mathbf{q}); \mathbf{K}_s, \mathbf{T}_s)$, with $\mathbf{p}_i(\mathbf{q}_s) = \mathbf{h}_i$ at the source state. We optimize $\mathbf{q}$ so that the projected surface points approach their targets while limiting unintended deformation elsewhere.

Jacobian-based weighting. FLAME deformation parameters are spatially entangled: a parameter that moves the handle toward its target may also deform unrelated facial regions. We therefore compute two Jacobians at the source state to identify parameters that are both effective for the requested motion and spatially localized.

First, the handle Jacobian

$$\mathbf{J}_i^h = \left. \frac{\partial \mathbf{p}_i(\mathbf{q})}{\partial \mathbf{q}} \right|_{\mathbf{q}=\mathbf{q}_s} \in \mathbb{R}^{2 \times 66} \tag{1}$$

measures how each deformation parameter moves the projected handle. For parameter q_k, we measure its response along the desired drag direction $\mathbf{d}_i$ by $a_{i,k} = |\mathbf{d}_i^\top \mathbf{J}_{i,:,k}^h|$. Therefore, $a_{i,k}$ measures how strongly parameter q_k can control motion along the requested drag direction, independent of the required update sign. Responses are aggregated across handles, and the K strongest dimensions form the candidate set.

A strong handle response alone, however, does not guarantee a desirable deformation, since the same parameter may also cause substantial changes in unrelated facial regions. We therefore compute a second, unwanted-region Jacobian over the full FLAME mesh.

$$\mathbf{J}^u = \left.\frac{\partial \,\mathrm{vec}(\mathbf{V}(\mathbf{q}))}{\partial \mathbf{q}}\right|_{\mathbf{q}=\mathbf{q}_s} \in \mathbb{R}^{3N_v \times 66}, \tag{2}$$

where $\mathbf{V}(\mathbf{q}) \in \mathbb{R}^{N_v \times 3}$ denotes the FLAME vertices. For a deformation parameter q_k, the corresponding Jacobian column is reshaped into N_v three-dimensional displacement vectors, whose norms measure how strongly each mesh vertex responds to that parameter. We define a local region $\mathcal{V}^{\mathrm{local}}$ using a mesh-hop neighborhood around the handle-associated triangles, and treat the remaining vertices as the far region $\mathcal{V}^{\mathrm{far}} = \mathcal{V} \setminus \mathcal{V}^{\mathrm{local}}$. For each parameter q_k, we summarize its influence on unrelated regions by the 95th percentile of the vertex-response magnitudes over $\mathcal{V}^{\mathrm{far}}$, denoted R_k^{far}. We normalize this response as

$\rho_k = \frac{R_k^{\mathrm{far}}}{R_{\mathrm{base}}+\epsilon}$, where R_{base} is the median far-region response among the strongest handle-responsive dimensions. Parameters with $\rho_k > \tau_u$ are rejected. The remaining handle-response scores are normalized to obtain soft weights $w_k \in [0, 1]$, which modulate the corresponding deformation parameters during optimization.

Deformation optimization. Given Jacobian-derived parameter weights, we optimize deformation parameters using three objectives. Geometric drag loss directly enforces the user-specified motion by minimizing the distance between the projected handle-associated surface points and targets, $\mathcal{L}_{\mathrm{geo}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{p}_i(\mathbf{q}) - \mathbf{t}_i\|_2$. Vertex preservation loss discourages changes away from the dragged region while allowing nearby vertices to move. For each source FLAME vertex $\mathbf{V}_i$, we compute its 3D distance d_i to the nearest handle-associated surface point and assign $w_i^{\mathrm{pres}} = 1 - \exp\left[-\left(\frac{d_i}{\sigma}\right)^p\right]$, where σ controls the spatial extent of editable region and p controls the sharpness of the transition. The loss can be written as $\mathcal{L}_{\mathrm{vpres}} = \frac{\sum_i w_i^{\mathrm{pres}} \|\Pi(\mathbf{V}_i(\mathbf{q})) - \Pi(\mathbf{V}_i(\mathbf{q}_s))\|_2^2}{\sum_i w_i^{\mathrm{pres}}}$. Regularization loss limits large deviations from the fitted source deformation parameters, $\mathcal{L}_{\mathrm{reg}} = \|\mathbf{q} - \mathbf{q}_s\|_2^2$, which helps maintain plausible facial geometry and prevents extreme expression. The final deformation objective is therefore

$$\mathcal{L}_{\mathrm{def}} = \lambda_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{vpres}}\mathcal{L}_{\mathrm{vpres}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}. \tag{3}$$

The Jacobian weights modulate the deformation parameter updates during optimization, and the optimized condition is then used by MMDM to synthesize the edited image.

3.2.2 ROTATION EDIT

For rotation edits, we keep the FLAME geometry fixed and optimize only the viewpoint offsets $\Delta\mathbf{c} = [\Delta c_{\mathrm{yaw}}, \Delta c_{\mathrm{pitch}}] \in \mathbb{R}^2$. Using an orbit-based camera update around a fixed 3D pivot, the updated extrinsic $\mathbf{T}(\Delta\mathbf{c})$ reprojects each persistent surface point as $\mathbf{p}_i(\Delta\mathbf{c}) = \Pi(\mathbf{P}_i; \mathbf{K}_s, \mathbf{T}(\Delta\mathbf{c}))$.

Rotation optimization. For each handle-associated surface point $\mathbf{P}_i$, let $\mathbf{p}_i(\Delta\mathbf{c}) = \Pi(\mathbf{P}_i; \mathbf{K}_s, \mathbf{T}(\Delta\mathbf{c}))$ denote its projected location under the updated viewpoint. Since the facial geometry remains fixed, we define the geometric drag loss as $\mathcal{L}_{\mathrm{geo}}^{\mathrm{rot}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{p}_i(\Delta\mathbf{c}) - \mathbf{t}_i\|_2^2$. To discourage unnecessarily large viewpoint changes, we regularize the rotation offsets by $\mathcal{L}_{\mathrm{reg}}^{\mathrm{rot}} = \left\|\frac{\Delta\mathbf{c}}{\Delta\mathbf{c}_{\max}}\right\|_2^2$, where $\Delta\mathbf{c}_{\max}$ specifies the allowed yaw and pitch ranges. The final rotation objective is

$$\mathcal{L}_{\mathrm{rot}} = \lambda_{\mathrm{geo}}^{\mathrm{rot}}\mathcal{L}_{\mathrm{geo}}^{\mathrm{rot}} + \lambda_{\mathrm{reg}}^{\mathrm{rot}}\mathcal{L}_{\mathrm{reg}}^{\mathrm{rot}}. \tag{4}$$

During optimization, the yaw and pitch offsets are constrained to their prescribed ranges. The optimized viewpoint is used to rebuild the MMDM condition and synthesize the rotated image.

4 BENCHMARK

4.1 DRAGFACE BENCHMARK

Existing drag-editing benchmarks are designed for general image manipulation rather than facial editing, and do not explicitly separate different underlying transformations such as deformation and

viewpoint rotation. We therefore introduce DragFace, a benchmark for point-based facial editing covering both transformation types. We randomly select 150 FFHQ images (Karras et al., 2019) and resize them to 512×512, using 130 for deformation and 20 for rotation. Each deformation image contains three editing cases with one to four handle-target pairs, yielding 390 edits, while each rotation image contains one single-pair case, yielding 20 edits. In total, DragFace contains 410 cases, each consisting only of a source image and handle-target pairs, without editing masks.

4.2 Geometry Mean Distance

Existing drag editing methods commonly evaluate point accuracy using DIFT-based Mean Distance (MD). However, DIFT correspondence can become unreliable under substantial semantic, appearance, or viewpoint changes, which can in turn make DIFT-based MD misrepresent editing quality (Tang et al., 2023; Choi et al., 2025). In our setting, MD can additionally be reduced by global face translation rather than the intended local deformation or viewpoint change, as illustrated in Fig. 4. We therefore introduce Geometry Mean Distance (Geometry MD), which tracks corresponding facial surface points using explicit FLAME geometry while discounting global face translation.

We use the handle-to-surface association to identify the corresponding FLAME surface point on the source and edited meshes, denoted as $\mathbf{P}_i^s$ and $\mathbf{P}_i^g$, respectively. For deformation edits, both points are projected using the source camera, $\mathbf{x}_i^s = \Pi_s(\mathbf{P}_i^s), \mathbf{x}_i^g = \Pi_s(\mathbf{P}_i^g)$, so that the measured displacement reflects facial deformation rather than viewpoint variation. For rotation edits, the source point is projected using the source camera while the edited point is projected using the camera recovered from the edited image, $\mathbf{x}_i^s = \Pi_s(\mathbf{P}_i^s), \mathbf{x}_i^g = \Pi_g(\mathbf{P}_i^g)$, thereby retaining the motion caused by the viewpoint change.



Figure 4: Example showing the limitation of conventional MD: global translation can yield a low score, while DIFT feature matching may miss the correct correspondence after facial deformation.

The geometric displacement of the corresponding surface point is $\mathbf{x}_i^g - \mathbf{x}_i^s$. To discount simple global face translation, we similarly compute the projected displacement of a predefined set of stable FLAME head vertices $\mathcal{V}_{\text{stable}}$ and use their component-wise median as the global displacement, $\mathbf{g} = \text{median}_{v \in \mathcal{V}_{\text{stable}}} (\mathbf{x}_v^g - \mathbf{x}_v^s)$, where the same projection convention is used for deformation and rotation as above. Geometry MD is then defined as

$$\text{GeoMD} = \frac{1}{N} \sum_{i=1}^{N} \left\| (\mathbf{x}_i^g - \mathbf{x}_i^s - \mathbf{g}) - (\mathbf{t}_i - \mathbf{h}_i) \right\|_2 . \tag{5}$$

A lower Geometry MD indicates that the corresponding facial surface points follow the requested drag more accurately after discounting global translation. Unlike feature based MD, Geometry MD relies on geometric correspondence rather than searching for a similar feature in the edited image.

5 Experiments

5.1 Experimental setup

We use MMDM with DDIM sampling using 20 denoising steps and a classifier-free guidance scale of 1.5. For deformation editing, we optimize the FLAME deformation parameters for 40 iterations using gradient descent with the proposed Jacobian-based parameter weighting. We retain the $K = 24$ dimensions with the strongest handle responses and set the unwanted-response rejection threshold to $\tau_u = 20$. For rotation editing, we optimize yaw and pitch using Adam for at most 150 iterations, constraining both angles to $[-45°, 45°]$. Additional details are provided in the appendix.

5.2 Baselines and evaluation metrics

We compare our method with four diffusion based point editing methods: DragDiffusion (Shi et al., 2024), DragNoise (Liu et al., 2024), GoodDrag (Zhang et al., 2025), and DragLoRA (Xia et al.,

2025). All methods are evaluated on the same DragFace source images and handle-target annotations. To ensure a common point-only interface, we provide each method with only the source image and handle-target pairs, without additional editing masks, although some baselines optionally support mask guidance. For both deformation and rotation, we report Geometry MD to measure how accurately the corresponding facial surface points follow the requested drag while discounting global face translation. For deformation, we additionally report the distance between the recovered FLAME expression parameters of the source and edited images as an indicator of expression change (Bounareli et al., 2024; Ding et al., 2023; Jang et al., 2025), together with Image Fidelity (IF), to measure source-image appearance preservation. For rotation, we report the recovered head pose change to indicate the magnitude of the induced viewpoint rotation (Bounareli et al., 2024; Wei et al., 2026); since no target rotation angle is specified, this is not treated as an accuracy metric. We further report canonical vertex-to-vertex (V2V) distance to measure preservation of the underlying 3D facial geometry during rotation (Grimmer et al., 2023; Otto et al., 2025; Li et al., 2021).

5.3 RESULTS

Explicit geometric control produces larger expression changes with lower drag error. DRAGOMETRY achieves the lowest mean GeoMD of 12.68 pixels, compared with 13.73 pixels for the strongest baseline DragNoise, with a corresponding improvement in the median (Table 1). Separately, its recovered expression-change magnitude reaches 0.934, compared with 0.463–0.567 for the baselines. This measures the extent of deformation rather than its correctness; the evidence for improved target alignment comes from GeoMD. Source-image similarity is mixed: DRAGOMETRY obtains an IF of 0.942, exceeding DragDiffusion and GoodDrag but falling below DragNoise and DragLoRA. Since IF responds to both intended deformation and unwanted appearance changes, this difference cannot be attributed solely to poorer preservation or to the larger expression changes.

Table 1: Quantitative comparison on deformation edits. GeoMD and expression change are reported as mean / median.

Method	GeoMD ↓	Expr. change	IF ↑
DragDiffusion	15.46 / 15.23	0.463 / 0.439	0.906
DragNoise	13.73 / 13.78	0.567 / 0.537	0.962
GoodDrag	15.10 / 14.67	0.501 / 0.489	0.897
DragLoRA	14.19 / 14.35	0.481 / 0.459	**0.969**
Ours	**12.68 / 12.84**	0.934 / 0.904	0.942

Improved rotation alignment comes with weaker preservation of recovered facial geometry. Table 2 shows DRAGOMETRY achieves the lowest mean GeoMD of 42.85 pixels, decreasing from the next best 47.12 pixels for GoodDrag, while the corresponding median decreases from 38.40 to 31.44 pixels. Its mean pose change is also larger, at $22.28°$ compared with $9.92°$–$13.79°$ for the baselines. Pose change measures how much the head orientation changes, rather than how accurately the dragged point reaches its target. Although DRAGOMETRY yields a higher canonical V2V distance than the baselines (1.117×10^{-3} versus 0.742×10^{-3}–0.853×10^{-3}), the absolute deviation remains small despite the substantially large viewpoint change, suggesting that the underlying 3D facial geometry is largely preserved during rotation.

Table 2: Quantitative comparison on rotation edits. Geometry MD and V2V are reported as mean / median, while pose change is reported as the mean rotation magnitude.

Method	GeoMD ↓	Pose change (°)	V2V ($\times 10^{-3}$) ↓
DragDiffusion	47.50 / 40.48	13.79	0.783 / 0.713
DragNoise	51.75 / 48.79	9.92	0.853 / 0.828
GoodDrag	47.12 / 38.40	12.66	**0.742** / 0.728
DragLoRA	51.26 / 43.79	11.69	0.768 / **0.707**
Ours	**42.85 / 31.44**	22.28	1.117 / 1.029

Explicit geometric control enables facial expression changes to more closely follow the requested drag. Figure 5 compares representative deformation results. Across mouth and eye edits, DRAGOMETRY produces changes in facial expression that more closely follow the specified handle–target directions, while several baselines retain more of the source expression or mainly shift and rescale local appearance. This distinction is particularly visible in the multi-handle mouth examples, where changes in mouth opening and lip configuration jointly accommodate the requested motion. These observations are consistent with the lower GeoMD and larger recovered expression changes reported in Table 1.



Figure 5: Qualitative comparison on deformation edits. Cyan and orange points represent handles and targets in drag edit. Drag directions are shown by an arrow between handles and targets.

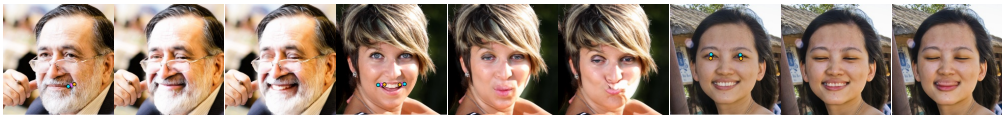

Figure 6: Qualitative ablation of the Jacobian-based parameter weighting for deformation edits. Each triplet compares the source drag specification with the full method and the variant without Jacobian weighting. The proposed weighting better preserves unrelated facial structure while realizing the requested local deformation.

Modeling camera rotation in addition to geometric control better aligns viewpoint changes with the requested drag. As seen in Fig. 7, several baselines remain close to the source head orientation despite the specified point displacement. In contrast, DRAGOMETRY produces visibly larger viewpoint changes that more closely follow the requested drag direction. In the illustrated examples, these changes also exhibit coherent structural details such as plausible neck appearance under large rotations (Fig. 7 middle row), while retaining the overall facial structure. This behavior is consistent with controlling the camera viewpoint while keeping the underlying facial mesh fixed. The qualitative observations complement the lower GeoMD reported in Table 2.

Jacobian weighting improves identity preservation with little change in drag accuracy. As an ablation, we remove the proposed weighting and directly optimize all deformation parameters using the same geometry formulation and MMDM backbone. As shown in Table 3, removing the weighting reduces mean GeoMD from 12.68 to 11.82 pixels, but also reduces mean identity similarity from 0.864 to 0.824. In a paired comparison, the full method achieves higher identity similarity in 308 of 390 edits (79.0%), with an average improvement of 0.040. This suggests that some parameters useful for moving the handle also induce substantial changes in unrelated facial regions. The Jacobian weighting suppresses such spatially entangled parameters, improving deformation locality and identity preservation at a modest cost in point accuracy. Figure 6 further visualizes this behavior, showing that dimensions rejected by the Jacobian can produce clear deformation outside the requested region.

Fig. 8 examines the mechanism underlying these unwanted changes. In both examples, a FLAME expression dimension that contributes to the requested mouth motion also affects the eyes. Without Jacobian weighting, such dimensions can introduce unwanted changes in distant facial regions. Our Jacobian criterion suppresses such entangled dimensions, producing localized deformations that support the requested motion. Fig. 8 illustrates the effect of entangled

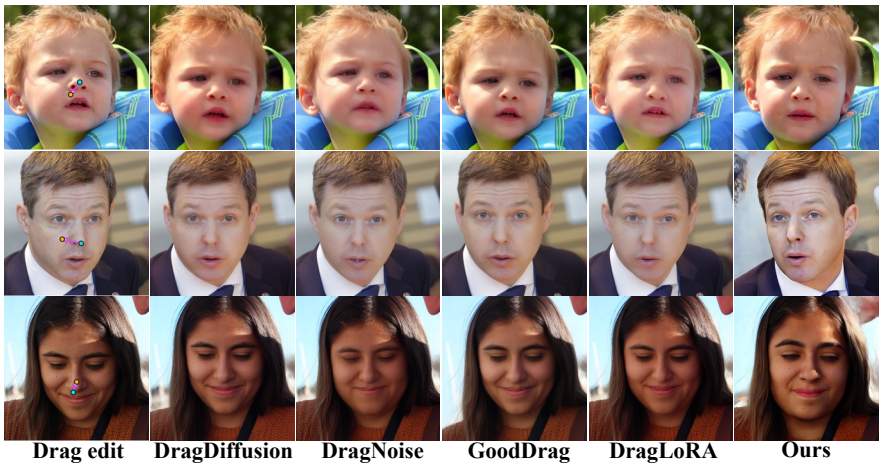

Figure 7: Qualitative comparison on rotation edits. Cyan and orange points represent handles and targets in drag edit. Drag directions are shown by an arrow between handles and targets.



Figure 8: **Effect of Jacobian-based weighting on entangled FLAME dimensions.** The illustrated expression dimensions couple mouth opening/closing with eye opening/closing. The last column in each example activates only the expression dimension rejected by our Jacobian criterion. These dimensions reproduce the unwanted region changes seen in the ablation, showing that they are a main source of the unintended distant-region deformation.

FLAME expression dimensions. In both examples, a dimension that contributes to the requested mouth motion also affects the eyes, coupling mouth opening/closing with eye opening/closing. Without Jacobian weighting, such dimensions can introduce unwanted changes in distant facial regions. Our Jacobian criterion suppresses these entangled dimensions and produces more localized deformation.

Table 3: Ablation of the Jacobian-based parameter weighting on deformation edits. Results are reported as mean / median.

Method	GeoMD ↓	Identity sim. ↑
w/o Jacobian	**11.82 / 11.54**	0.824 / 0.837
Ours	12.68 / 12.84	**0.864 / 0.872**

6 Conclusion

We introduced a point-based editing framework for geometry-conditioned diffusion models that performs dragging directly in the explicit geometry condition space. Instantiated on MMDM with FLAME, our method realizes local deformation through expression parameters with Jacobian-based parameter weighting and viewpoint rotation through camera variables. We further introduced DragFace and Geometry MD for evaluating facial deformation and rotation using explicit geometric correspondence. Experiments demonstrate more accurate geometric point control, stronger facial deformation, and substantially larger coherent viewpoint changes than existing diffusion-based drag editing methods.

Although evaluated on MMDM with FLAME, the formulation does not rely on MMDM-specific internal representations and could be instantiated with other geometry-conditioned models that expose structured geometric parameters and correspondence, such as SMPL/SMPL-X based human generation. Current limitations mainly arise from the expressiveness and fitting accuracy of the underlying geometry representation. Extending the framework to other geometric domains and automatically inferring the intended transformation from sparse drag inputs are promising directions for future work.

AI disclosure. In this work, we used generative AI tools, including ChatGPT, to provide feedback on the research methodology and experimental review, assist with debugging, refine aspects of the conceptual framing and hypotheses, and support interpretation of experimental results. We did not use generative AI to generate synthetic datasets, formulate or prove theoretical claims, translate research material, or clean and reformat the dataset. Additionally, we used generative AI tools to assist with scientific figure creation and modification, suggest experimental parameters, draft and edit parts of the manuscript, improve readability and structure, summarize and search relevant literature, identify related work, and brainstorm presentation and terminology. All AI-assisted code was inspected and tested on our experimental pipeline, experimental results and interpretations were independently checked by the authors, literature-related suggestions were verified against the original publications, and AI-assisted figures and manuscript text were manually reviewed and revised. We take responsibility for the final content of this work, including text, claims, code, experiments, and artifacts produced with the aid of generative AI.

REFERENCES

Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. One-shot neural face reenactment via finding directions in gan's latent space. *International Journal of Computer Vision*, 132(8):3324–3354, 2024.

Benito Buchheim, Max Reimann, and Jürgen Döllner. Controlling human shape and pose in text-to-image diffusion models via domain adaptation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3688–3697. IEEE, 2025.

Yuhao Cheng, Zhuo Chen, Xingyu Ren, Wenhan Zhu, Zhengqin Xu, Di Xu, Changpeng Yang, and Yichao Yan. 3d-aware face editing via warping-guided latent direction learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 916–926, 2024.

Gayoon Choi, Taejin Jeong, Sujung Hong, and Seong Jae Hwang. Dragtext: Rethinking text embedding in point-based image editing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 441–450. IEEE, 2025.

Yutao Cui, Xiaotong Zhao, Guozhen Zhang, Shengming Cao, Kai Ma, and Limin Wang. Stabledrag: Stable dragging for point-based image editing. In *European conference on computer vision*, pp. 340–356. Springer, 2024.

Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12736–12746, 2023.

Jiahua Dong and Yu-Xiong Wang. 3dgs-drag: Dragging gaussians for intuitive point-based 3d editing. In *International Conference on Learning Representations*, volume 2025, pp. 82799–82815, 2025.

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21003–21012, 2023.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. In *International Conference on Learning Representations*, volume 2026, pp. 50353–50376, 2026.

Marcel Grimmer, Christian Rathgeb, Raymond Veldhuis, and Christoph Busch. Neutrex: A 3d quality component measure on facial expression neutrality. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–8. IEEE, 2023.

Weiran Guang, Xiaoguang Gu, Mengqi Huang, and Zhendong Mao. Dragin3d: image editing by dragging in 3d space. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21502–21512. IEEE, 2025.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 538–549, 2024.

Xingzhong Hou, Boxiao Liu, Yi Zhang, Jihao Liu, Yu Liu, and Haihang You. Easydrag: Efficient point-based manipulation on diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8404–8413. IEEE, 2024.

Wooseok Jang, Youngjun Hong, Geonho Cha, and Seungryong Kim. Controlface: Harnessing facial parametric control for face rigging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5614–5624, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

MD Wahiduzzaman Khan, Mingshan Jia, Xiaolin Zhang, En Yu, Caifeng Shan, and Kaska Musial-Gabrys. Instaface: Identity-preserving facial editing with single image inference. In *2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–10. IEEE, 2026.

Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5481–5492, 2024.

Gwanhyeong Koo, Sunjae Yoon, Younghwan Lee, Ji Woo Hong, and Chang D. Yoo. FlowDrag: 3D-aware drag-based image editing with mesh-guided deformation vector flow fields. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 31467–31485, 2025.

Yushi Lan, Feitong Tan, Qiangeng Xu, Di Qiu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas Funkhouser, Chen Change Loy, et al. Loc3diff: Local diffusion for 3d human head synthesis and editing. In *European Conference on Computer Vision*, pp. 49–66. Springer, 2024.

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *Proceedings of the computer vision and pattern recognition conference*, pp. 16008–16018, 2025.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL `https://doi.org/10.1145/3130800.3130813`.

Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. Topologically consistent multi-view face inference using volumetric sampling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3804–3814. IEEE, 2021.

Haonan Lin. Dreamsalon: A staged diffusion framework for preserving identity-context in editable face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8589–8598, 2024.

Pengyang Ling, Lin Chen, Pan Zhang, Huaian Chen, Yi Jin, and Jinjin Zheng. Freedrag: Feature dragging for reliable point-based image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6860–6870, 2024.

Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6743–6752, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 31620–31631, 2024.

Christopher Otto, Prashanth Chandran, Sebastian Weiss, Markus Gross, Gaspard Zoss, and Derek Bradley. Multimodal conditional 3d face geometry generation. *Computers & Graphics*, pp. 104325, 2025.

Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2085–2094, 2021.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27080–27090, 2024.

Yansong Qu, Dian Chen, Xinyang Li, Xiaofan Li, Shengchuan Zhang, Liujuan Cao, and Rongrong Ji. Drag your gaussian: Effective drag-based editing with score distillation for 3d gaussian splatting. In *Proceedings of the special interest group on computer graphics and interactive techniques conference conference papers*, pp. 1–12, 2025.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1532–1540, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252, 2020.

Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8839–8849, 2024.

Kaede Shiohara and Toshihiko Yamasaki. Face2diffusion for fast and editable face personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6850–6859, 2024.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

Felix Taubner, Prashant Raina, Mathieu Tuli, Eu Wern Teh, Chul Lee, and Jinmiao Huang. 3D face tracking from 2D video through iterative dense UV to image flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1227–1237, June 2024.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B Lindell. Mvp4d: Multi-view portrait video diffusion for animatable 4d avatars. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5318–5330. IEEE Computer Society, 2025b.

Mengting Wei, Tuomas Varanka, Yante Li, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Towards consistent and controllable image synthesis for identity-preserving face editing. *Pattern Recognition*, pp. 114322, 2026.

Siwei Xia, Li Sun, Tiantian Sun, and Qingli Li. Draglora: Online optimization of lora adapters for drag-based image editing in diffusion model. In *The International Conference on Machine Learning (ICML)*, 2025.

Zheng Yu, Yaohua Wang, Siying Cui, Aixi Zhang, Wei-Long Zheng, and Senzhang Wang. Fuseany-part: Diffusion-driven facial parts swapping via multiple reference images. *Advances in Neural Information Processing Systems*, 37:80864–80884, 2024.

Zewei Zhang, Huan Liu, Jun Chen, and Xiangyu Xu. Gooddrag: Towards good practices for drag editing with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 13785–13808, 2025.

A Additional implementation details

For all experiments, DDIM sampling uses $\eta = 0$. Given a fixed initial latent, conditioning input, model weights, and random seed, this setting produces deterministic samples.

For deformation editing, we optimize the deformation parameters $\mathbf{q}$ using gradient descent with a learning rate of 0.03 and a gradient multiplier of 2.0. Each parameter update is clipped to a maximum absolute step of 0.05, and the total deviation of each optimized parameter from its source value is constrained to 2.0. For the Jacobian-based parameter weighting, the local facial region is defined as the four-hop neighborhood on the FLAME mesh around the vertices of the handle-associated triangles, while all remaining vertices constitute the unwanted region. We retain the $K = 24$ dimensions with the strongest handle responses and use the unwanted-response rejection threshold $\tau_u = 20$, as specified in the main paper.

For rotation editing, we optimize the two-dimensional viewpoint offset $\Delta\mathbf{c} = [\Delta c_{\text{yaw}}, \Delta c_{\text{pitch}}]$, using Adam with a learning rate of 0.25. The geometric dragging and angle regularization losses use $\lambda_{\text{geo}} = 1.0$ and $\lambda_{\text{reg}} = 0.001$, respectively. Optimization terminates early when the projected handle is within 4 pixels of its target. We update the viewpoint using an orbit-based camera transformation, rotating around a pivot determined by the fitted FLAME translation rather than rotating the camera in place.

B Additional related work

Fine-grained facial editing. Facial editing has been explored through latent manipulation, semantic attribute control, structured parametric models, and region-based guidance. Latent-space methods discover directions for attributes such as age, expression, pose, and appearance (Shen et al., 2020; Härkönen et al., 2020; Shen & Zhou, 2021; Patashnik et al., 2021), while 3DMM-based approaches provide explicit control over facial geometry, expression, pose, and rendering conditions (Ding et al., 2023; Lan et al., 2024; Jang et al., 2025; Khan et al., 2026). Other methods use text prompts, landmarks, masks, or selected facial regions to achieve semantic or local edits (Shiohara & Yamasaki, 2024; Cheng et al., 2024; Lin, 2024; Peng et al., 2024; Yu et al., 2024). Although many of these methods provide fine-grained control, their editing interfaces are typically defined by predefined attributes, parameters, text descriptions, landmarks, or regions. Point-based dragging is complementary to these controls: it can be applied after a semantic or parametric edit to further refine local facial geometry by directly specifying where an individual point should move.

C Additional qualitative results

We provide additional qualitative comparisons for both deformation and rotation edits to complement the examples in the main paper. Fig. 9 presents more local deformation cases with different handle-target configurations, while Fig. 10 shows additional viewpoint rotation results. These examples further illustrate the behavior of our method across a broader range of drag instructions and allow qualitative comparison with the baseline methods.

D Selective failure examples

Failure cases. Fig. 11 shows representative failure cases in which local blur or rendering artifacts appear around the edited region. These artifacts arise during synthesis with MMDM under the optimized geometry condition. To examine whether they are caused by individual FLAME expression dimensions, Fig. 12 visualizes several highly responsive expression dimensions separately. The corresponding single-dimension results do not exhibit the same artifacts, suggesting that the failure is associated with synthesis under the jointly optimized condition rather than a particular expression dimension. This illustrates that our method remains limited by the generation capability of the underlying geometry-conditioned diffusion model.



Figure 9: More qualitative comparison on deformation edits. Cyan and orange points represent handles and targets in drag edit. Drag directions are shown by an arrow between handles and targets.



Figure 10: More qualitative comparison on rotation edits. Cyan and orange points represent handles and targets in drag edit. Drag directions are shown by an arrow between handles and targets.

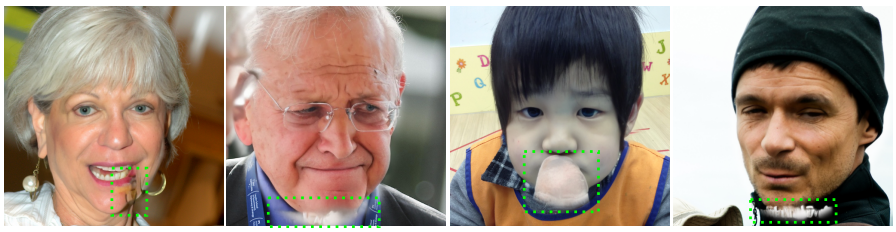

Figure 11: **Representative failure cases.** Local blur or rendering artifacts, circled by green dash boxes, can appear around strongly edited facial regions when MMDM synthesizes the optimized geometry condition, reflecting limitations of the underlying geometry-conditioned diffusion model.



Figure 12: **Analysis of synthesis artifacts.** We compare the drag instruction, our final result, and results obtained by activating individual highly responsive FLAME expression dimensions. The single-dimension results do not reproduce the same local artifacts, suggesting that the failure arises during synthesis of the jointly optimized condition rather than from a particular expression dimension.