GENTRI: GENERATIVE TRIANGULATION FOR OCCLUSION-ROBUST 3D MOUSE POSE ESTIMATION

Anonymous authors
Paper under double-blind review

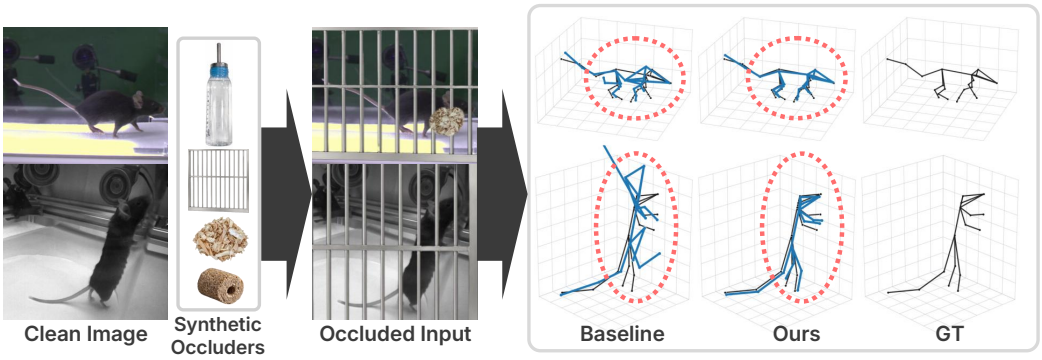


Figure 1: **Under severe occlusion, the baseline fails and GenTri still recovers the 3D pose.** Severe synthetic occlusions are applied to the clean frames (left). On the occluded input, the baseline produces the structural errors marked in red, while GenTri reconstructs anatomically plausible poses that remain close to the ground truth (black). The occluded mouse evaluation set was constructed specifically for this study.

ABSTRACT

Markerless 3D mouse pose estimation enables scalable behavioral analysis in neuroscience, but occlusion remains a major challenge to reliable reconstruction across experimental environments. Limited annotated data make it difficult to handle diverse occlusion patterns through training alone. Applying pose estimation to hours-long behavioral recordings also requires efficient inference. We introduce GenTri, an inference-time framework combining multi-view geometry with a flow-matching pose prior. To reduce the influence of confidently incorrect detections under occlusion, GenTri replaces the L_2 reprojection objective with a redescending Cauchy objective. For efficiency, geometry-aware Gauss–Newton guidance determines each correction's direction and magnitude. This enables GenTri to match its first-order variant with $2\times$ (mouse) to $17\times$ (human) less guidance time. Across multiple 2D detectors and three mouse datasets spanning in-domain and out-of-domain environments, GenTri reduces reconstruction error under synthetic occlusion by 45% on average relative to the reconstruction method paired with each detector. The same guidance formulation extends to human pose benchmarks, demonstrating applicability beyond mouse pose estimation. See our project page: `https://gentri-3dpose.github.io/`.

1 INTRODUCTION

Quantifying mouse behavior is central to studying neural function, motor control, and disease-related phenotypes (Weinreb et al., 2024; Wu et al., 2025). Because marker-based motion capture is difficult to deploy in freely moving animals, markerless pose estimation has become an important tool for behavioral analysis (Mathis et al., 2018; Pereira et al., 2022). Many ethologically relevant

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.

behaviors, such as rearing, grooming, and locomotion, are inherently three-dimensional and poorly resolved from a single view, motivating multi-view systems based on geometric triangulation or learned volumetric fusion (Karashchuk et al., 2021; Dunn et al., 2021; Schneider et al., 2022).

In practical experiments, however, self-occlusion, cage structures, bedding, and experimental apparatus can produce missing or confidently incorrect 2D detections that severely distort multi-view reconstruction (Figure 1). Semi-supervised learning and large-scale animal pretraining improve detector generalization across recording conditions (Li et al., 2023; Biderman et al., 2024; Ye et al., 2024), but cannot eliminate such failures under severe, previously unseen occlusions.

A complementary strategy is to handle such failures at inference time. In human pose estimation, diffusion- and flow-based generative priors (Ho et al., 2020; Lipman et al., 2023; Ci et al., 2023) constrain reconstruction to plausible 3D poses, while test-time guidance enforces consistency with available 2D observations (Stathopoulos et al., 2024). Existing guidance, however, remains sensitive to severely corrupted observations. Under the standard L_2 reprojection objective, the influence of a detection grows with its residual, allowing a few incorrect detections to dominate the reconstruction. Recent robust diffusion solvers mitigate this issue through robust losses or reweighting (Yang et al., 2026; Zheng et al., 2026). This alone does not account for how calibrated multi-view geometry constrains corrections along different 3D directions.

In addition, behavioral experiments can involve hours of recording repeated across multiple days, making inference efficiency important in practice (Marshall et al., 2021; Markowitz et al., 2023). Most observation-based guidance rules inherit from DPS (Chung et al., 2023). They apply repeated first-order corrections controlled by a prescribed step size and back-propagate the observation loss through the denoiser. Each step computes a Jacobian of the generative network and stable inference requires fine integration. Existing guidance therefore falls short of jointly suppressing corrupted observations and exploiting calibrated geometry for effective 3D corrections in few steps.

To address these issues, we introduce **GenTri**, a robust and efficient inference-time framework. A flow-matching prior trained on clean 3D poses supplies plausible pose structure where detections are missing or unreliable, while the remaining detections guide it at inference. At every flow step, the pose predicted by the prior is refined toward the multi-view detections by a Gauss–Newton step (Triggs et al., 2000; Hartley & Zisserman, 2004) under a redescending Cauchy reprojection objective. Large reprojection residuals are therefore downweighted rather than fit quadratically. An identity approximation of the endpoint Jacobian avoids back-propagation through the frozen flow network, while interpolant transport carries the endpoint correction to the current flow state. Under severe synthetic occlusion simulating cage environments, GenTri reduces the absolute mean per-joint position error (MPJPE) over three mouse test sets from 31.6 mm with triangulation to 10.2 mm. These gains hold for every evaluated detector and extend to Human3.6M-Occluded. To assess generalization, the mouse evaluation covers out-of-domain recordings, including our newly collected and manually annotated treadmill test set, without test-time adaptation.

- To address confidently incorrect detections without occlusion-specific training, we propose an inference pipeline that guides a frozen flow-matching pose prior with a scale-adaptive Cauchy reprojection energy.
- Our Gauss–Newton correction at the predicted endpoint requires no back-propagation through the prior and matches first-order guidance with $2\times$ to $17\times$ less guidance time.
- Across three 2D detectors and three mouse domains, in and out of domain, GenTri reduces error by 45% on average and transfers to Human3.6M-Occluded without retuning.

2 Background

2.1 Multi-View 3D Animal Pose Estimation

Markerless systems such as DeepLabCut (Mathis et al., 2018) and SLEAP (Pereira et al., 2022) have enabled 2D landmark detection across diverse species and experimental settings. Multi-view animal systems recover 3D pose through geometric or temporal triangulation, as in Anipose (Karashchuk et al., 2021), or by fusing multi-view image evidence in a common 3D space, as in DANNCE (Dunn et al., 2021) and FreiPose (Schneider et al., 2022). Recent work further improves data efficiency and generalization through unlabeled video and large-scale animal pretraining (Biderman et al., 2024;

Ye et al., 2024). Despite these advances, detector failures caused by novel occlusions can still arise at inference.

Multi-view human pose estimation is comparatively mature, with learnable triangulation and cross-view fusion as standard paradigms (Iskakov et al., 2019; Qiu et al., 2019; He et al., 2020), and its occlusion-robust variants use adaptive fusion, uncertainty-aware reasoning, or geometry–appearance fusion (Zhang et al., 2021; Davoodnia et al., 2024; Liao et al., 2024), with geometry-guided diffusion also explored (Zhang et al., 2024). Despite exploiting cross-view information, these methods, such as learnable triangulation (LT) and AdaFuse, remain vulnerable to confidently incorrect 2D detections under severe occlusion.

2.2 Multi-View Triangulation and Gauss–Newton Refinement

Given M calibrated and fixed cameras with projections $\Pi_m : \mathbb{R}^3 \to \mathbb{R}^2$ and the 2D keypoints $\mathbf{y}_m \in \mathbb{R}^2$, $m = 1, \dots, M$, of one point, its absolute 3D position $\mathbf{p} \in \mathbb{R}^3$ is estimated by minimizing the reprojection error (Hartley & Zisserman, 2004)

$$\min_{\mathbf{p}} \ \frac{1}{2} \sum_{m=1}^{M} \left\| \mathbf{r}_m(\mathbf{p}) \right\|_2^2, \qquad \mathbf{r}_m(\mathbf{p}) = \mathbf{y}_m - \Pi_m(\mathbf{p}). \tag{1}$$

Each view provides a 2D observation of the same 3D point; the residual $\mathbf{r}_m$ measures the discrepancy between this observation and the reprojection of the current 3D estimate in the image plane. Because perspective projection is nonlinear, Gauss–Newton (Triggs et al., 2000; Nocedal & Wright, 2006) linearizes each residual as $\mathbf{r}_m(\mathbf{p} + \Delta\mathbf{p}) \approx \mathbf{r}_m(\mathbf{p}) + \mathbf{J}_m^{\text{rep}} \Delta\mathbf{p}$ with $\mathbf{J}_m^{\text{rep}} = \partial \mathbf{r}_m / \partial \mathbf{p} \in \mathbb{R}^{2\times 3}$, giving the gradient $\mathbf{g}$ and the Gauss–Newton matrix $\mathbf{H}$,

$$\mathbf{g} = \sum_{m=1}^{M} \left(\mathbf{J}_m^{\text{rep}}\right)^\top \mathbf{r}_m, \qquad \mathbf{H} = \sum_{m=1}^{M} \left(\mathbf{J}_m^{\text{rep}}\right)^\top \mathbf{J}_m^{\text{rep}}, \tag{2}$$

where $\mathbf{H} \approx \partial \mathbf{g} / \partial \mathbf{p}$ approximates the Hessian by dropping the second derivatives of Π_m. The Gauss–Newton step $\Delta\mathbf{p} = -\mathbf{H}^{-1}\mathbf{g}$ then updates $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$; when $\mathbf{H}$ is nearly singular along directions the cameras barely observe, the damped Levenberg–Marquardt step $-(\mathbf{H} + \kappa \operatorname{diag} \mathbf{H})^{-1}\mathbf{g}$ is used instead (Levenberg, 1944; Marquardt, 1963). $\mathbf{H}$ captures how strongly the cameras constrain each 3D direction, and our endpoint correction uses it to determine the direction and magnitude of each update (Sec. 3.3).

2.3 Generative Models for 3D Pose

Under occlusion, missing or confidently incorrect keypoints no longer constrain the 3D pose uniquely, and a triangulation-style estimator returns a pose that fits the corrupted detections. Generative models address this by sampling from a learned distribution of plausible poses, either as unconditional diffusion priors that regularize estimation, denoising, and completion (Lu et al., 2023; 2025), or as 2D-conditioned score and diffusion models that generate multiple 3D hypotheses (Ci et al., 2023), which later methods aggregate or rank (Shan et al., 2023; Zhang & Carlone, 2025); flow matching has recently replaced diffusion in the conditional family (Wang et al., 2026; Le et al., 2026) because its linear paths admit few-step sampling and a clean estimate at every step. A rectified flow (Lipman et al., 2023; Liu et al., 2023) connects a Gaussian source $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a clean 3D pose $\mathbf{x}_1$ along the linear interpolant and learns the velocity field by regression,

$$\mathbf{x}_t = (1-t)\,\mathbf{x}_0 + t\,\mathbf{x}_1, \qquad \mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x}_1, \mathbf{x}_0, t}\left[\left\| v_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0) \right\|_2^2 \right], \qquad t \sim \mathcal{U}(0,1). \tag{3}$$

Sampling integrates $\mathrm{d}\mathbf{x}_t / \mathrm{d}t = v_\theta(\mathbf{x}_t, t)$ from $t = 0$ to 1 in steps of size Δt, and at every step the conditional mean $\hat{\mathbf{x}}_1 = \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_t]$ follows from Tweedie's formula (Efron, 2011; Kim & Ye, 2021),

$$\hat{\mathbf{x}}_1 = \mathbf{x}_t + (1-t) v_\theta(\mathbf{x}_t, t), \tag{4}$$

The conditional methods above learn $v_\theta(\mathbf{x}_t, t, \mathbf{c})$ with the 2D observation $\mathbf{c}$ as an input, thereby tying the learned distribution to a particular form of visual evidence. Task-independent priors instead keep $v_\theta(\mathbf{x}_t, t)$ unconditional and introduce observations only during inference, allowing the same prior to accommodate different downstream constraints (Lu et al., 2023; 2025). GenTri takes the latter route: $v_\theta(\mathbf{x}_t, t)$ is trained on clean root-relative poses, and the detections enter only through the guidance of Sec. 2.4, applied at the clean estimate of Eq. 4.

2.4 Observation-Guided Inference with Generative Priors

A generative prior alone produces plausible samples; when an observation $\mathbf{y}$ is available, guidance steers the sampling toward it by adding the gradient of a loss $\mathcal{L}$ that scores the current sample against $\mathbf{y}$ (Chung et al., 2023; He et al., 2024b),

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \Delta t \left[v_\theta(\mathbf{x}_t, t) - \lambda \nabla_{\mathbf{x}_t} \mathcal{L}\big(\hat{\mathbf{x}}_1(\mathbf{x}_t); \mathbf{y}\big) \right]. \tag{5}$$

The loss is evaluated at the clean estimate $\hat{\mathbf{x}}_1$ because no measurement model is defined for the noisy state, and the standard choice is the L_2 loss $\mathcal{L} = \|\mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_1)\|_2^2$ for a forward operator $\mathcal{A}$ (here the projections Π_m of Sec. 2.2). Existing rules differ along two axes (Table 1). The loss: DPS (Chung et al., 2023) and MPGD (He et al., 2024b) keep the L_2 form, whereas RDP (Yang et al., 2026) and Robust-CG (Zheng et al., 2026) bound the influence of large residuals with inverse multi-quadric (IMQ) or Huber weights.

Method	3D pose	Inference	Loss	Step rule	w/o net. grad.
DPS	✗	sampling	L_2	first-order	✗
RDP	✗	sampling	IMQ	first-order	✗
Robust-CG	✗	sampling	Huber	inner CG solve	✓
CLAMP	✗	sampling	L_2	Gauss–Newton	✗
ScoreHMR	✓	sampling	L_2	first-order	✗
NRDF	✓	optim.	Geman–McClure	first-order	✗
DPoser-X	✓	optim.	Geman–McClure	first-order	✓
Ours	✓	sampling	Cauchy	Gauss–Newton	✓

Table 1: **Inference rules by loss and step rule.** optim.: test-time optimization with the prior as a regularizer; w/o net. grad.: the step does not differentiate through the generative network.

The step rule: DPS, RDP, and MPGD take a first-order step scaled by λ, where DPS and RDP back-propagate through the network and MPGD applies the step at the clean estimate without network gradients; Robust-CG and DDS (Chung et al., 2024) instead solve an inner conjugate-gradient system at the clean estimate, and CLAMP (Shin et al., 2026) replaces the first-order step by a damped Gauss–Newton solve but still differentiates through the denoiser. These rules were developed for image inverse problems with a known forward operator. In 3D human pose, ScoreHMR (Stathopoulos et al., 2024) follows the same recipe with a confidence-weighted L_2 reprojection loss, and DPoser-X (Lu et al., 2025) and NRDF (He et al., 2024a) run a first-order test-time optimization with the prior as a regularizer; none uses calibrated multi-view geometry. GenTri changes both axes: like MPGD and DDS, it corrects the clean estimate without network derivatives, but solves it in closed form under a redescending energy (Sec. 3).

3 Main Contribution: GenTri

3.1 Problem Setup and Update Rule

GenTri integrates robust triangulation refinement into generative pose sampling. At each step, the prior predicts a plausible pose, and a robust geometric correction steers it toward the observed keypoints (Figure 2). As the pose prior, we adopt the rectified flow described in Sec. 2.3 (Appendix F). The prior v_θ is trained once on clean root-relative poses and then frozen, while the detector is trained independently on per-view images (Sec. 4.1). The two components are coupled only at inference.

For each joint $j = 1, \ldots, J$, treated as a 3D point as in Sec. 2.2, the detector provides in view m a keypoint $\mathbf{y}_{mj} \in \mathbb{R}^2$ and confidence $\alpha_{mj} \in [0, 1]$. Since the prior is trained on root-centered poses normalized by the training-set scale s, we recover the absolute joint positions at flow step t, $\mathbf{p}_{t,j} \in \mathbb{R}^3$, collectively $\mathbf{p}_t \in \mathbb{R}^{3J}$, from the predicted endpoint as

$$\mathbf{p}_{t,j} = s\, \hat{\mathbf{x}}_{1,j}(\mathbf{x}_t) + \boldsymbol{\tau}_t, \tag{6}$$

where $\hat{\mathbf{x}}_1(\mathbf{x}_t) \in \mathbb{R}^{3J}$ is the normalized pose estimate predicted by the prior at the current step (Eq. 4), $\hat{\mathbf{x}}_{1,j} \in \mathbb{R}^3$ is its joint j, $\boldsymbol{\tau}_t \in \mathbb{R}^3$ is the global translation (estimated at inference, Appendix B.2), and s, the training-set standard deviation of root-relative joints, restores the metric scale lost by normalization.

Rather than applying an observation gradient to the flow state as in Eq. 5, our update corrects the predicted endpoint at every flow step,

$$\mathbf{x}_{t'} = \mathbf{x}_t + \Delta t\, v_\theta(\mathbf{x}_t, t) - \underbrace{\frac{t'}{s}}_{\text{(iii)}} \underbrace{\mathbf{M}_t^{-1}}_{\text{(ii)}} \underbrace{\nabla_{\mathbf{p}_t} \mathcal{E}(\mathbf{p}_t)}_{\text{(i)}}, \qquad t' = t + \Delta t, \tag{7}$$

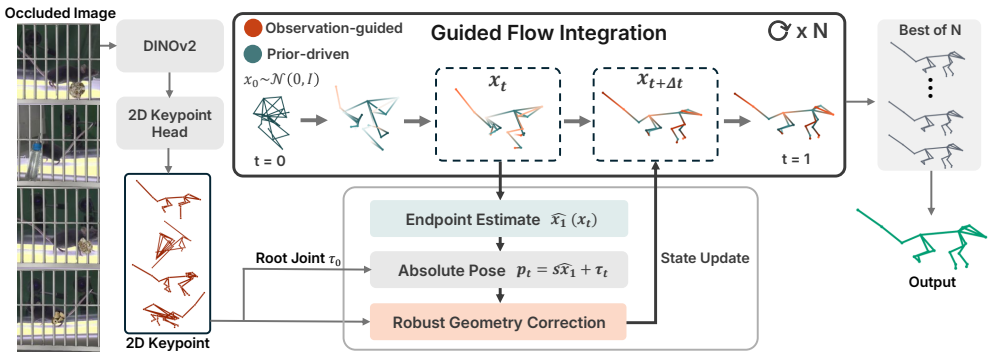


Figure 2: **Overview of GenTri inference.** At every flow step, a Gauss–Newton step on the robust reprojection energy corrects the prior's clean-pose estimate using the calibrated cameras and a second step updates the global translation; the correction is carried back to the flow state, and the lowest-energy hypothesis among the N hypotheses is returned.

where (i) $\mathcal{E}$ is a robust observation energy, differentiated in the endpoint space where the reprojection geometry lives, that replaces the L_2 loss of Eq. 5; (ii) $\mathbf{M}_t^{-1}$, the inverse of a damped Gauss–Newton matrix with one 3×3 block $\mathbf{M}_{t,j}$ per joint, replaces the scalar λ; and (iii) t' carries the endpoint displacement to the current flow time. The remaining sections detail each factor, with Sec. 3.2 defining $\mathcal{E}$ and Sec. 3.3 deriving $\mathbf{M}_t$ and the transport factor t'.

3.2 Robust Multi-View Observation Energy

As discussed in Sec. 3.1, we build $\mathcal{E}$ from a robust penalty rather than the L_2 error of Eq. 1. With the joints of Eq. 6, the residual of Sec. 2.2 is taken per joint and view, $\mathbf{r}_{mj}(\mathbf{p}_t) = \mathbf{y}_{mj} - \Pi_m(\mathbf{p}_{t,j})$ with $r_{mj}(\mathbf{p}_t) = \|\mathbf{r}_{mj}(\mathbf{p}_t)\|_2$, and the energy sums a penalty ρ of these errors. Under L_2, $\rho(r) = r^2$, the influence function (Huber, 1981), $\rho'(r) = 2r$, grows without bound, so the large residuals of a few occluded keypoints dominate the correction. We therefore adopt the Cauchy penalty (Black & Rangarajan, 1996) with scale $C_t > 0$,

$$\rho(r) = C_t^2 \log\left(1 + \frac{r^2}{C_t^2}\right), \qquad \rho'(r) = \frac{2r}{1 + r^2/C_t^2}, \tag{8}$$

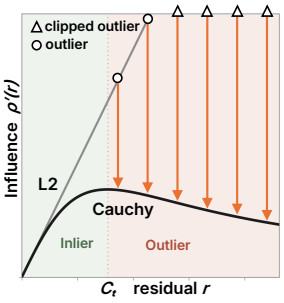


Figure 3: L_2 **and Cauchy influence** $\rho'(r)$**.**

which coincides with r^2 for $r \ll C_t$ and grows only logarithmically beyond, so that $\rho'(r) \to 0$ for large residuals (Figure 3: the Cauchy influence peaks near C_t and decays). The scale C_t is re-estimated at every flow step from the median of the current residuals (Appendix B.2), so no dataset-specific threshold is introduced. The observation energy for our guidance, evaluated at the absolute endpoint $\mathbf{p}_t$ at every step, is

$$\mathcal{E}(\mathbf{p}_t) = \sum_{m=1}^{M} \sum_{j=1}^{J} \alpha_{mj}\, \rho\big(r_{mj}(\mathbf{p}_t)\big). \tag{9}$$

Each view m and joint j enters as one term: α_{mj} down-weights uncertain detections and the Cauchy penalty bounds the contribution of any single view with a large residual, so a joint occluded in one view is still pulled by its remaining views. $\mathcal{E}$ replaces $\mathcal{L}$ of Eq. 5 in the guidance step.

3.3 Efficient Geometry-Aware Guidance

Geometry-aware step size. The scalar step size in Eq. 5 does not account for the direction-dependent curvature of the multi-view observation energy. For example, when two cameras have nearly parallel viewing directions, a joint is weakly constrained along the viewing direction but strongly constrained in the perpendicular directions. A single λ must therefore balance the risk of overshooting in strongly constrained directions against slow progress in weakly constrained

Algorithm 1 GenTri Inference

Require: detections $(\mathbf{y}_{mj}, \alpha_{mj})$, cameras Π_m, prior v_θ, scale s, steps S, damping κ
$\Delta t \leftarrow 1/S$; $\quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = 0, \Delta t, \ldots, 1{-}\Delta t$ **do**
  $\hat{\mathbf{x}}_1 \leftarrow \mathbf{x}_t + (1{-}t)\, v_\theta(\mathbf{x}_t, t)$; $\quad \mathbf{p}_t \leftarrow s\,\hat{\mathbf{x}}_1 + \boldsymbol{\tau}_t$ ▷ predicted endpoint, absolute pose
  $\mathbf{H}_j,\ \mathbf{M}_{t,j} \leftarrow$ Eqs. 10, 11, $j = 1, \ldots, J$ ▷ per-joint Gauss–Newton
  $\mathbf{x}_{t+\Delta t} \leftarrow \mathbf{x}_t + \Delta t\, v_\theta(\mathbf{x}_t, t) - (t{+}\Delta t)/s\, \mathbf{M}_t^{-1} \nabla_{\mathbf{p}_t} \mathcal{E}(\mathbf{p}_t)$ ▷ final update rule, Eq. 7
end for
return $\mathbf{x}_1$ ▷ normalized pose state

ones (Shin et al., 2026). As a result, first-order guidance can require many steps. This direction-dependent curvature is exactly what the Gauss–Newton matrix of Sec. 2.2 encodes. With the residual $\mathbf{r}_{mj}$ of Sec. 3.2, the joint-j block of $\nabla_{\mathbf{p}_t} \mathcal{E}(\mathbf{p}_t)$ and the Gauss–Newton matrix of $\mathcal{E}$ are

$$\nabla_{\mathbf{p}_{t,j}} \mathcal{E} = \sum_{m=1}^{M} \omega_{mj} \big(\mathbf{J}_{mj}^{\mathrm{rep}}\big)^\top \mathbf{r}_{mj}, \qquad \mathbf{H}_j = \sum_{m=1}^{M} \omega_{mj} \big(\mathbf{J}_{mj}^{\mathrm{rep}}\big)^\top \mathbf{J}_{mj}^{\mathrm{rep}}, \tag{10}$$

with $\omega_{mj} = \alpha_{mj}\, \rho'(r_{mj})/r_{mj}$ and $\mathbf{J}_{mj}^{\mathrm{rep}} = \partial \mathbf{r}_{mj}/\partial \mathbf{p}_{t,j} \in \mathbb{R}^{2\times 3}$, the confidence- and influence-weighted form of Eq. 2. The gradient reflects the current reprojection mismatch, whereas $\mathbf{H}_j \in \mathbb{R}^{3\times 3}$ encodes how strongly each 3D direction is constrained across views. Since $\mathbf{H}_j$ is nearly singular along directions the cameras barely observe, inverting it directly would yield unbounded corrections there; we therefore damp it,

$$\mathbf{M}_{t,j} = \mathbf{H}_j + \kappa\, \mathrm{diag}(\mathbf{H}_j), \tag{11}$$

where κ is a constant damping coefficient (Levenberg, 1944; Marquardt, 1963). The resulting step $-\mathbf{M}_{t,j}^{-1} \nabla_{\mathbf{p}_{t,j}} \mathcal{E}$ is large where the cameras leave a direction free and small where they pin it down, and, because each residual involves one joint, $\mathbf{M}_t^{-1} \nabla_{\mathbf{p}_t} \mathcal{E}$ in Eq. 7 amounts to J independent 3×3 solves.

Identity-Jacobian approximation. Computing this correction in flow-state coordinates, as Eq. 5 does, would require pulling the endpoint gradient and curvature back through the flow network. Since $\mathbf{r}_{mj}$ depends on $\mathbf{x}_t$ through $\hat{\mathbf{x}}_1(\mathbf{x}_t)$ and Eq. 6, the chain rule factorizes as

$$\frac{\partial \mathbf{r}_{mj}}{\partial \mathbf{x}_t} = \underbrace{\frac{\partial \mathbf{r}_{mj}}{\partial \mathbf{p}_{t,j}}}_{\mathbf{J}_{mj}^{\mathrm{rep}}\ \text{(camera)}} \underbrace{\frac{\partial \mathbf{p}_{t,j}}{\partial \hat{\mathbf{x}}_1}}_{s\,\mathbf{S}_j} \underbrace{\frac{\partial \hat{\mathbf{x}}_1}{\partial \mathbf{x}_t}}_{\mathbf{J}_{\mathrm{net}}\ \text{(network)}}, \qquad \mathbf{J}_{\mathrm{net}} = \mathbf{I} + (1-t)\, \frac{\partial v_\theta(\mathbf{x}_t, t)}{\partial \mathbf{x}_t}, \tag{12}$$

where $\mathbf{S}_j \in \mathbb{R}^{3\times 3J}$ selects joint j and $\mathbf{J}_{\mathrm{net}} \in \mathbb{R}^{3J\times 3J}$ is the endpoint Jacobian of the flow network. Eq. 10 uses only the camera factor; the state-space gradient and curvature would be $s\, \mathbf{J}_{\mathrm{net}}^\top \nabla_{\mathbf{p}_t} \mathcal{E}(\mathbf{p}_t)$ and $s^2\, \mathbf{J}_{\mathrm{net}}^\top \mathbf{H}\, \mathbf{J}_{\mathrm{net}}$, with $\mathbf{H}$ the block-diagonal matrix of the $\mathbf{H}_j$, and obtaining them requires differentiating v_θ at every step. To avoid this cost, we follow QGF (Zhou et al., 2026) and set $\mathbf{J}_{\mathrm{net}} \approx \mathbf{I}$. The approximation keeps the gradient and curvature of Eq. 10 in endpoint coordinates and discards the cross-joint coupling of the network, so the 3×3 solves stay independent per joint; the endpoint-to-state scale is left to the transport step below.

Transport to the current flow time. The endpoint step $-\mathbf{M}_t^{-1} \nabla_{\mathbf{p}_t} \mathcal{E}$ is a displacement of the predicted clean endpoint, not a velocity at the current state, and has to be carried to $\mathbf{x}_t$ along the flow. We employ the clean-estimate rescaling of MPGD (He et al., 2024b): the source estimate $\hat{\mathbf{x}}_0 = \mathbf{x}_t - t\, v_\theta(\mathbf{x}_t, t)$ is held fixed and the next state is reconstructed from the corrected endpoint, which at $t' = t + \Delta t$ gives

$$\mathbf{x}_{t'} = (1-t')\, \hat{\mathbf{x}}_0 + t' \Big(\hat{\mathbf{x}}_1 - \tfrac{1}{s} \mathbf{M}_t^{-1} \nabla_{\mathbf{p}_t} \mathcal{E}(\mathbf{p}_t) \Big) = \mathbf{x}_t + \Delta t\, v_\theta(\mathbf{x}_t, t) - \tfrac{t'}{s}\, \mathbf{M}_t^{-1} \nabla_{\mathbf{p}_t} \mathcal{E}(\mathbf{p}_t),$$

which is Eq. 7; $1/s$ converts the metric correction to normalized coordinates, and the interpolant contributes only the factor t' in front of the correction. Algorithm 1 summarizes the pose update; the full procedure with the translation update and best-of-N selection is Algorithm 2 in Appendix B.2.

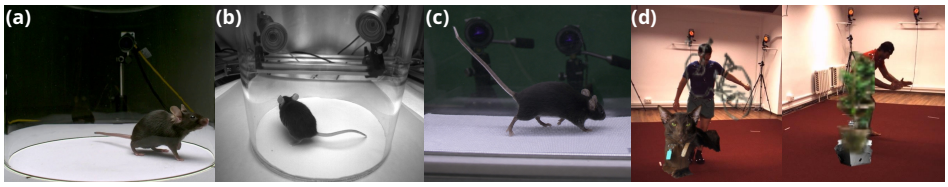


Figure 4: **Evaluation environments.** (a) DANNCE open-field, rendering-matched; (b) out-of-domain THM open-field; (c) our out-of-domain treadmill; (d) Human3.6M-Occluded.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Mouse. We train a DINOv2 (Oquab et al., 2024) 2D detector and its LT (Iskakov et al., 2019) and AdaFuse (Zhang et al., 2021) variants only on synthetic MoReMouse (Zhong et al., 2026) renderings. Evaluation uses three datasets (Fig. 4): a held-out DANNCE (Dunn et al., 2021) open-field set (in-domain), THM (Zhao et al., 2023) (out-of-domain), and a newly collected, manually annotated treadmill set (out-of-domain); no test-time adaptation is applied. Mouse-Occluded overlays cage bars and additional object occluders on each view of the original (clean) frames (Appendix B.1).

Human. We evaluate on Human3.6M (Ionescu et al., 2014) and Human3.6M-Occluded (Bragagnolo et al., 2025). The 2D detector is a SimpleBaseline ResNet-152 (Xiao et al., 2018) trained on MPII (Andriluka et al., 2014) and Human3.6M, again with LT and AdaFuse variants (Appendix B).

Implementation details. For each species we train one unconditional flow-matching prior, a graph network over the species' skeleton, on clean root-relative 3D poses. Unless stated otherwise, every result uses $S{=}10$ guidance steps, $N{=}4$ hypotheses, and damping $\kappa{=}0.1$, identically across species, datasets, and detectors (Appendix B.2). Occlusion enters only at evaluation. GenTri entries are three-seed means (Appendix B). Single mouse entries are macro averages over the three domains. The confidence α_{mj} is the heatmap peak for the base detectors, the algebraic confidence for LT, and the learned view weight for AdaFuse.

4.2 OCCLUSION-ROBUST 3D POSE ESTIMATION

Baselines. For mouse, we compare with confidence-weighted triangulation, RANSAC-based outlier rejection (Fischler & Bolles, 1981), Anipose's robust triangulation (Karashchuk et al., 2021), and learned multi-view methods LT (Iskakov et al., 2019) and AdaFuse (Zhang et al., 2021). Human baselines (Appendix C.3) include prior-based FMPose3D (Wang et al., 2026), DPoser-X (body) (Lu et al., 2025), NRDF (He et al., 2024a), and ScoreHMR (Stathopoulos et al., 2024). Multi-view baselines are RUMPL (Ghasemzadeh et al., 2026) and SkelSplat (Bragagnolo et al., 2026).

Mouse. To test whether robust guidance helps regardless of the detector, we apply GenTri to the DINOv2, LT, and AdaFuse detections in all three mouse domains (Table 2; qualitative results in Appendix A.1). GenTri lowers the error relative to each detector's baseline reconstruction for every detector in every domain and, with identical DINOv2 detections, outperforms RANSAC-based outlier rejection, including on the unseen treadmill. Averaged over the nine detector and domain pairs, error drops by 45% relative to the reconstruction method paired with each detector. The largest reductions occur out of domain, where detections degrade most: on the THM data set, error drops from 81.2 to 13.2 mm with LT and from 44.9 to 14.1 mm with AdaFuse, and the treadmill set, recorded with a different apparatus and behavior, follows the same trend. GenTri is therefore most effective where occlusion and domain shift combine, which is where changing the detector is least practical.

Human. To test transfer across skeletons, we train a human pose prior and keep every inference hyperparameter from the mouse experiments. On Human3.6M-Occluded, GenTri reaches 30.7 mm with ResNet-152 detections, outperforming the compared multi-view and prior-based methods, and improves the LT and AdaFuse detections (Table 3). On clean data, GenTri improves mouse reconstruction with DINOv2 and LT detections and human reconstruction with ResNet-152 and AdaFuse detections, and otherwise stays within 0.4 mm of the corresponding baseline (Appendices A.2–A.3). The same inference therefore transfers across species and detectors without retuning.

2D detector	3D reconstruction	**Open-field** *InD*	**Open-field** *OoD*	**Treadmill** *OoD*
DINOv2	Triangulation (Hartley & Zisserman, 2004)	8.3	70.6	16.0
DINOv2	RANSAC (Fischler & Bolles, 1981)	6.3	20.3	13.9
DINOv2	Anipose‡ (Karashchuk et al., 2021)	5.2‡	69.4	13.2‡
DINOv2	**Ours**	5.1	16.1	9.5
LT†	LT (Iskakov et al., 2019)	7.1	81.2	13.9
LT†	**Ours**	5.7	**13.2**	**9.2**
AdaFuse†	AdaFuse (Zhang et al., 2021)	5.2	44.9	15.6
AdaFuse†	**Ours**	**4.7**	14.1	9.8

Table 2: **Performance on Mouse-Occluded.** Absolute MPJPE (mm) ↓. † DINOv2-based detector. ‡ Uses temporal information; Open-field OoD is non-sequential.

2D detector	3D reconstruction	Human3.6M-Occluded
Image	ScoreHMR (Stathopoulos et al., 2024)	43.8
ResNet-152	Triangulation (Hartley & Zisserman, 2004)	57.0
ResNet-152	FMPose3D‡ (Wang et al., 2026)	51.6
ResNet-152	DPoser-X (body)‡ (Lu et al., 2025)	41.7
ResNet-152	RUMPL (Ghasemzadeh et al., 2026)	39.7
ResNet-152	SkelSplat (Bragagnolo et al., 2026)	36.4
ResNet-152	NRDF‡ (He et al., 2024a)	34.7
ResNet-152	**Ours**	30.7
LT†	LT (Iskakov et al., 2019)	71.5
LT†	**Ours**	32.9
AdaFuse†	AdaFuse (Zhang et al., 2021)	35.6
AdaFuse†	**Ours**	**29.8**

Table 3: **Performance on Human3.6M-Occluded.** Absolute MPJPE (mm) ↓. † ResNet-152-based detector. ‡Our multi-view extensions of the monocular FMPose3D, DPoser-X (body), and NRDF.

(a) Prior and guidance

	Mouse			**Human**		
Frame subset	w/o prior	w/o guid.	Ours	w/o prior	w/o guid.	Ours
All frames	15.5	17.9	**10.2**	37.2	45.7	**30.6**
Hardest 50%	19.5	22.6	**11.4**	48.9	56.4	**36.9**
Hardest 25%	25.2	27.8	**12.9**	66.8	71.9	**43.9**
Hardest 10%	31.6	38.3	**15.5**	104.7	104.1	**54.6**

(b) Robust penalty

Penalty	**Mouse**	**Human**
L2	15.6	48.1
Huber	11.6	34.7
IMQ	11.1	33.4
Geman–McClure	**10.0**	33.4
Cauchy (Ours)	10.2	**30.7**

Table 4: **Guidance ablations.** Absolute MPJPE (mm; ↓). Mouse results are averaged across three test sets. **(a)** Prior and guidance ablations on subsets ranked by initial reprojection error. Human results are averaged over frames. **(b)** Robust penalty ablations. Human results are averaged over actions.

4.3 Ablations

All ablations use the base 2D detectors (DINOv2 for mouse and ResNet-152 for human), without LT or AdaFuse. For the guidance comparisons in Figure 5, we adapt existing inverse-problem methods to the same pose flow prior, keeping the 2D detections, confidence weights, camera model, initialization, and translation solver fixed. Only the guidance objective and update rule vary across methods. Guidance constants are tuned on clean data; adaptation details are provided in Appendix C.1.

Role of the prior and guidance. To assess whether the pose prior and observation guidance are both needed, we remove each in Table 4(a). These variants respectively perform robust reprojection refinement from triangulation and unconditional prior sampling. Both increase error on every subset, with larger gaps under more challenging observations. Pose structure and observation constraints are therefore complementary, and each matters more as observations degrade.

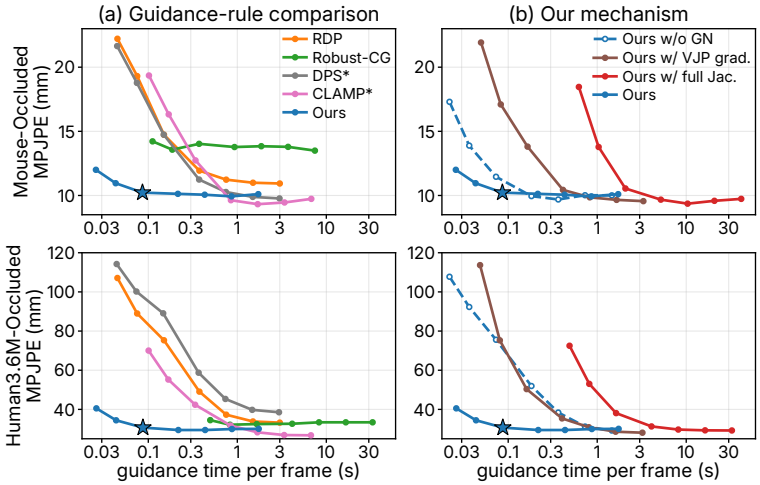


Figure 5: **Accuracy versus guidance time.** Absolute MPJPE (mm) ↓ per frame on Mouse-Occluded (top) and Human3.6M-Occluded (bottom); points mark guidance steps $S \in \{3, 5, \ldots, 200\}$ and ★ the deployed $S{=}10$ (per-step latency in Table 11). (a) Existing guidance rules; *Cauchy-matched variants. (b) Our correction without GN, with VJP, and with the full endpoint Jacobian.

Choice of robust penalty. To test whether robustness depends on Cauchy specifically, we compare reprojection penalties in Table 4(b). All robust penalties outperform L_2 on both species; Cauchy gives the lowest human error and stays within 0.3 mm of the best mouse result. Thus, limiting large-residual influence is broadly beneficial, and Cauchy provides a competitive default across species.

Comparison with existing guidance rules. To evaluate accuracy under limited guidance time, we compare GenTri with RDP (Yang et al., 2026) and Robust-CG (Zheng et al., 2026) in Figure 5(a). RDP has higher error than GenTri at every evaluated time budget. On Human3.6M-Occluded, GenTri achieves 30.7 mm in 0.09 s per frame at 10 steps. Robust-CG's best five-step setting gives 32.2 mm in 0.83 s, taking about $10\times$ longer due to its inner solve. To isolate the effects of correction and state-transition rules, we compare DPS (Chung et al., 2023) and CLAMP (Shin et al., 2026) using our Cauchy objective, as detailed in Appendix C.2. DPS* uses first-order gradients; CLAMP* uses Gauss–Newton guidance. GenTri has the lowest error among the compared rules on both species at guidance times up to 0.4 s per frame. To improve on GenTri's deployed accuracy, CLAMP* requires about $10\times$ the deployed guidance time on mouse and $20\times$ on human. DPS* requires about $17\times$ on mouse. GenTri is therefore the most accurate of the compared rules when guidance time is limited.

Ablating the correction rule. To assess the efficiency of Gauss–Newton (GN) and the identity Jacobian approximation, we ablate each in Figure 5(b). Among the plotted step budgets, matching GenTri's deployed error at 0.09 s per frame requires $2\times$ the guidance time on mouse and $17\times$ on human without GN. Vector–Jacobian product (VJP) gradients require $10\times$ and $19\times$, respectively, and the full endpoint Jacobian over $50\times$ on both species. Endpoint Gauss–Newton without network derivatives therefore gives the best accuracy per unit of guidance time.

5 Conclusion

We presented GenTri for inference-time multi-view 3D pose estimation under occlusion. Existing guidance lets a single confidently incorrect detection dominate an L_2 reprojection energy and needs many first-order steps to apply corrections stably. GenTri replaces the energy with an adaptive redescending Cauchy penalty, and the first-order step with a Gauss–Newton correction at the predicted endpoint sized by the calibrated geometry, enabling reconstruction in 10 steps. With one configuration, GenTri improves occluded reconstruction across three mouse domains and three detectors and transfers to Human3.6M-Occluded. We believe that robust observation modeling offers a practical route for applying generative pose priors to hours of behavioral recordings across species.

Reproducibility Statement

All inference settings of GenTri (guidance steps, damping, number of hypotheses, and the Cauchy scale rule), the evaluation protocol, and the variance across inference seeds are given in Appendix B, and the adaptation of every baseline guidance rule, together with how its constants are selected, in Appendix C.1. The construction of the three mouse domains, the synthetic occluders, and the annotation procedure are described in Appendix B.1; the Human3.6M-Occluded protocol, and our one deviation from it, are described in Appendix B.1. Per-step timings and hardware are in Table 11 (Appendix D). Code, the trained priors, and the mouse evaluation sets are available at `https://github.com/gentri-3dpose/gentri`.

AI Use Statement

Generative AI tools were used to assist with literature search and summarization, interpretation of experimental results, and manuscript editing. AI-assisted suggestions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

Ethics Statement

All animal experiments in this study were conducted under protocols approved by our institution's Institutional Animal Care and Use Committee (IACUC).

References

Liang An, Jilong Ren, Tao Yu, Tang Hai, Yichang Jia, and Yebin Liu. Three-dimensional surface motion capture of multiple freely moving pigs using MAMMAL. *Nature Communications*, 14: 7727, 2023.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Dan Biderman, Matthew R. Whiteway, Cole Hurwitz, Nicholas Greenspan, Robert S. Lee, Ankit Vishnubhotla, Richard Warren, Federico Pedraja, Dillon Noone, Michael M. Schartner, Julia M. Huntenburg, Anup Khanal, Guido T. Meijer, Jean-Paul Noel, Alejandro Pan-Vazquez, Karolina Z. Socha, Anne E. Urai, The International Brain Laboratory, John P. Cunningham, Nathaniel B. Sawtell, and Liam Paninski. Lightning Pose: Improved Animal Pose Estimation via Semi-Supervised Learning, Bayesian Ensembling and Cloud-Native Open-Source Tools. *Nature Methods*, 21(7):1316–1328, 2024. doi: 10.1038/s41592-024-02319-1.

Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19 (1):57–91, 1996.

Laura Bragagnolo, Matteo Terreran, Davide Allegro, and Stefano Ghidoni. Multi-view Pose Fusion for Occlusion-Aware 3D Human Pose Estimation. In *Computer Vision – ECCV 2024 Workshops*, Lecture Notes in Computer Science, pp. 117–133. Springer, 2025. doi: 10.1007/978-3-031-92591-7_8.

Laura Bragagnolo, Leonardo Barcellona, and Stefano Ghidoni. SkelSplat: Robust multi-view 3D human pose estimation with differentiable Gaussian rendering. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *International Conference on Learning Representations (ICLR)*, 2024.

Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. GFPose: Learning 3D human pose prior with gradient fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

Vandad Davoodnia, Saeed Ghorbani, Marc-André Carbonneau, Alexandre Messier, and Ali Etemad. UPose3D: Uncertainty-aware 3D human pose estimation with cross-view and temporal cues. In *European Conference on Computer Vision (ECCV)*, 2024.

Timothy W. Dunn, Jesse D. Marshall, Kyle S. Severson, Diego E. Aldarondo, David G. C. Hildebrand, Selmaan N. Chettih, William L. Wang, Amanda J. Gellis, David E. Carlson, Dmitriy Aronov, Winrich A. Freiwald, Fan Wang, and Bence P. Ölveczky. Geometric deep learning enables 3D kinematic profiling across species and environments. *Nature Methods*, 18(5):564–573, 2021. doi: 10.1038/s41592-021-01106-6.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. In *International Conference on Machine Learning (ICML)*, 2025.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Seyed Abolfazl Ghasemzadeh, Alexandre Alahi, and Christophe De Vleeschouwer. RUMPL: Ray-based transformers for universal multi-view 2D to 3D human pose lifting. *IEEE Transactions on Image Processing*, 2026.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. NRDF: Neural Riemannian Distance Fields for Learning Articulated Pose Priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1661–1671, 2024a.

Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *International Conference on Learning Representations (ICLR)*, 2024b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peter J. Huber. *Robust Statistics*. Wiley, 1981.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7718–7727, 2019.

Pierre Karashchuk, Katie L. Rupp, Evyn S. Dickinson, Sarah Walling-Bell, Elischa Sanders, Eiman Azim, Bingni W. Brunton, and John C. Tuthill. Anipose: A toolkit for robust markerless 3D pose estimation. *Cell Reports*, 36(13):109730, 2021. doi: 10.1016/j.celrep.2021.109730.

Kwanyoung Kim and Jong Chul Ye. Noise2Score: Tweedie's approach to self-supervised image denoising without clean images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Cuong Le, Pavlo Melnyk, Bastian Wandt, and Mårten Wadenbäck. Flow Matching for Probabilistic Monocular 3D Human Pose Estimation. *Transactions on Machine Learning Research*, 2026. arXiv:2601.16763.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.

Tianqing Li, Kyle S. Severson, Fan Wang, and Timothy W. Dunn. Improved 3D markerless mouse pose estimation using temporal semi-supervision. *International Journal of Computer Vision*, 131 (6):1389–1405, 2023. doi: 10.1007/s11263-023-01756-3.

Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, and Steven L. Waslander. Multiple view geometry transformers for 3D human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. Explicit occlusion reasoning for multi-person 3D human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Yulun Zhang, and Haoqian Wang. DPoser: Diffusion model as robust 3D human pose prior. *arXiv preprint arXiv:2312.05541*, 2023.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, and Ziwei Liu. DPoser-X: Diffusion model as robust 3D whole-body human pose prior. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jeffrey E. Markowitz, Winthrop F. Gillis, Maya Jay, Jeffrey Wood, Ryley W. Harris, Robert Cieszkowski, Rebecca Scott, David Brann, Dorothy Koveal, Tomasz Kula, Caleb Weinreb, Mohammed Abdal Monium Osman, Sandra Romero Pinto, Naoshige Uchida, Scott W. Linderman, Bernardo L. Sabatini, and Sandeep Robert Datta. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature*, 614(7946):108–117, 2023. doi: 10.1038/s41586-022-05611-2.

Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

Jesse D. Marshall, Diego E. Aldarondo, Timothy W. Dunn, William L. Wang, Gordon J. Berman, and Bence P. Ölveczky. Continuous whole-body 3D kinematic recordings across the rodent behavioral repertoire. *Neuron*, 109(3):420–437, 2021.

Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018. doi: 10.1038/s41593-018-0209-y.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D'Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods*, 19(4): 486–495, 2022. doi: 10.1038/s41592-022-01426-1.

Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3D human pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Artur Schneider, Christian Zimmermann, Mansour Alyahyay, Florian Steenbergen, Thomas Brox, and Ilka Diester. 3D pose estimation enables virtual head fixation in freely moving rats. *Neuron*, 110(13):2080–2093.e10, 2022. doi: 10.1016/j.neuron.2022.04.019.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3D human pose estimation with multi-hypothesis aggregation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Seunghyeok Shin, Minwoo Kim, Dabin Kim, and Hongki Lim. Geometry-correct diffusion posterior sampling with denoiser-pullback curvature guidance and manifold-aligned damping. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.27990.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3D human recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice (ICCV Workshop)*, pp. 298–372, 2000.

Ti Wang, Xiaohang Yu, and Mackenzie Weygandt Mathis. FMPose3D: Monocular 3D pose estimation via flow matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Weiquan Wang, Jun Xiao, Chunping Wang, Wei Liu, Zhao Wang, and Long Chen. Di^2Pose: Discrete diffusion model for occluded 3D human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Caleb Weinreb, Jonah E. Pearl, Sherry Lin, Mohammed Abdal Monium Osman, Libby Zhang, Sidharth Annapragada, Eli Conlin, Red Hoffmann, Sofia Makowska, Winthrop F. Gillis, Maya Jay, Shaokai Ye, Alexander Mathis, Mackenzie W. Mathis, Talmo Pereira, Scott W. Linderman, and Sandeep Robert Datta. Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *Nature Methods*, 21(7):1329–1339, 2024. doi: 10.1038/s41592-024-02318-2.

Joshua Wu, Hari Koneru, James Ravenel, Anshuman Sabath, James Roach, Shaun Lim, Michael Tadross, Alex Williams, and Timothy W. Dunn. Disentangling 3D animal pose dynamics with scrubbed conditional latent variables. In *International Conference on Learning Representations (ICLR)*, 2025.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

Yiming Yang, Xiaoyuan Cheng, Yi He, Kaiyu Li, Wenxuan Yuan, and Zhuo Sun. Outlier-robust diffusion posterior sampling for Bayesian inverse problems. *arXiv preprint arXiv:2602.02045*, 2026.

Shaokai Ye, Anastasiia Filippova, Jessy Lauer, Steffen Schneider, Maxime Vidal, Tian Qiu, Alexander Mathis, and Mackenzie Weygandt Mathis. SuperAnimal pretrained pose estimation models for behavioral analysis. *Nature Communications*, 15:5165, 2024. doi: 10.1038/s41467-024-48792-2.

Harry Zhang and Luca Carlone. CHAMP: Conformalized 3D Human Multi-Hypothesis Pose Estimators. In *International Conference on Learning Representations (ICLR)*, 2025.

Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7376–7385, 2020.

Xinyi Zhang, Qinpeng Cui, Qiqi Bao, Wenming Yang, and Qingmin Liao. Geometry-guided diffusion model with masked transformer for robust multi-view 3D human pose estimation. In *ACM International Conference on Multimedia (ACM MM)*, 2024. doi: 10.1145/3664647.3681265.

Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. AdaFuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129:703–718, 2021.

Jiachen Zhao, Tao Yu, Liang An, Yipeng Huang, Fang Deng, and Qionghai Dai. Triangulation residual loss for data-efficient 3D pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 12721–12732, 2023.

Yang Zheng, Jiahua Liu, Tongyao Pang, Wen Li, and Zhaoqiang Liu. Outlier-robust diffusion solvers for inverse problems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yuan Zhong, Jingxiang Sun, Zhongbin Zhang, Liang An, and Yebin Liu. MoReMouse: Monocular reconstruction of laboratory mouse. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Zhiyuan Zhou, Andy Peng, Charles Xu, Qiyang Li, Jost Tobias Springenberg, Kevin Frans, and Sergey Levine. Test-time gradient guidance of flow policies in reinforcement learning. *arXiv preprint arXiv:2606.11087*, 2026.

APPENDIX

A Additional Results **15**

A.1 Mouse: Qualitative Results 15

A.2 Mouse: Clean and Occluded Results 16

A.3 Human3.6M: Clean and Occluded Results 16

B Experimental and Implementation Details **17**

B.1 Datasets and Training 17

B.2 Inference and Hyperparameters 19

C Baseline Adaptations **21**

C.1 Adaptation of Guidance Rules to the Flow Prior 21

C.2 Published Objective vs. Our Cauchy Objective 23

C.3 Human Pose Baselines 23

D Computational Cost **23**

E Limitations **24**

F Choice of Generative Prior and Sampler **24**

A ADDITIONAL RESULTS

A.1 MOUSE: QUALITATIVE RESULTS

Figure 6 compares reconstructions on occluded treadmill and open-field frames across in-domain and out-of-domain settings.

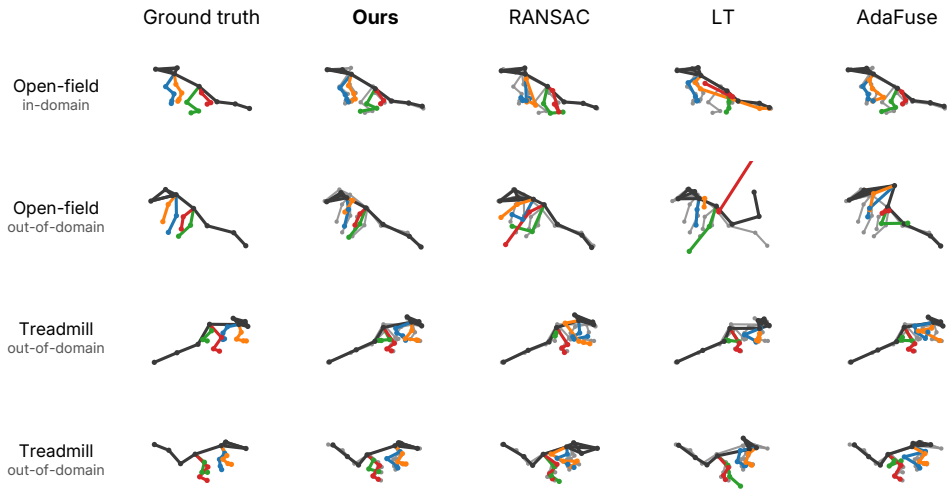


Figure 6: **Qualitative results on Mouse-Occluded.** The ground-truth pose is overlaid in gray.

A.2 MOUSE: CLEAN AND OCCLUDED RESULTS

(a) Mouse Clean

		Open-field *InD*		Open-field *OoD*		Treadmill *OoD*	
2D detector	3D reconstruction	Abs	Rel	Abs	Rel	Abs	Rel
DINOv2	Triangulation (Hartley & Zisserman, 2004)	3.9	4.8	6.7	7.5	6.5	9.1
DINOv2	RANSAC (Fischler & Bolles, 1981)	2.8	3.8	5.3	6.3	7.3	9.9
DINOv2	Anipose‡ (Karashchuk et al., 2021)	2.7‡	3.6‡	5.0	5.3	7.6‡	10.3‡
DINOv2	**Ours**	2.4	3.1	4.6	5.4	4.7	7.7
LT†	LT (Iskakov et al., 2019)	3.4	4.2	22.4*	23.6*	4.8	7.8
LT†	**Ours**	2.6	3.1	4.9	6.3	**4.3**	7.6
AdaFuse†	AdaFuse (Zhang et al., 2021)	**2.2**	3.1	**4.4**	**5.3**	5.1	7.7
AdaFuse†	**Ours**	2.3	**3.0**	4.4	5.3	4.4	**7.3**

(b) Mouse Occluded

		Open-field *InD*		Open-field *OoD*		Treadmill *OoD*	
2D detector	3D reconstruction	Abs	Rel	Abs	Rel	Abs	Rel
DINOv2	Triangulation (Hartley & Zisserman, 2004)	8.3	10.4	70.6	103.4	16.0	21.3
DINOv2	RANSAC (Fischler & Bolles, 1981)	6.3	7.6	20.3	29.2	13.9	17.5
DINOv2	Anipose‡ (Karashchuk et al., 2021)	5.2‡	**6.0**‡	69.4	97.2	13.2‡	26.3‡
DINOv2	**Ours**	5.1	6.5	16.1	22.8	9.5	**12.0**
LT†	LT (Iskakov et al., 2019)	7.1	11.5	81.2	88.0	13.9	25.3
LT†	**Ours**	5.7	8.0	**13.2**	**17.0**	**9.2**	12.8
AdaFuse†	AdaFuse (Zhang et al., 2021)	5.2	7.7	44.9	55.5	15.6	21.3
AdaFuse†	**Ours**	**4.7**	6.2	14.1	19.3	9.8	13.5

Table 5: **Performance on Mouse, clean and occluded.** Absolute and Root-relative MPJPE (mm) ↓. † DINOv2-based detector. ‡ Uses temporal information; Open-field OoD is non-sequential. *Mean inflated by rare single-view 2D outliers.

Table 5 reports absolute and Root-relative MPJPE on clean and occluded mouse data across the three domains. With DINOv2 and LT detections, GenTri also improves reconstruction on clean observations across the three domains. With the same DINOv2 detections, its three-domain mean absolute MPJPE is approximately 3.9 mm, compared with 5.1 mm for RANSAC, 5.1 mm for Anipose, and 5.7 mm for triangulation: relative reductions of about 24%, 23%, and 32%, respectively. On the out-of-domain treadmill set, the error falls from 7.3 mm with RANSAC and 7.6 mm with Anipose to 4.7 mm, reductions of about 35% and 38%. The improvement extends to other detector outputs. On clean treadmill frames, GenTri reduces absolute MPJPE from 4.8 to 4.3 mm with LT detections and from 5.1 to 4.4 mm with AdaFuse detections, corresponding to reductions of approximately 10% and 12%, respectively. These results show that the reconstruction gains extend beyond the synthetically occluded setting.

A.3 HUMAN3.6M: CLEAN AND OCCLUDED RESULTS

On clean Human3.6M, GenTri improves absolute MPJPE with ResNet-152 and AdaFuse detections and stays within 0.4 mm of LT (Table 6). The largest gain is with ResNet-152 detections, where it reduces error from 23.5 to 22.3 mm relative to confidence-weighted triangulation and outperforms all other reconstruction methods using the same detections. With LT and AdaFuse detections, clean accuracy remains comparable: absolute MPJPE changes from 20.2 to 20.6 mm for LT and decreases from 22.2 to 21.9 mm for AdaFuse. Root-relative error follows the same trend, with improvements for ResNet-152 and AdaFuse and a 0.4 mm increase for LT. Together with the occluded results, these

		Human3.6M			
		Clean		Occluded	
2D detector	3D reconstruction	Abs	Rel	Abs	Rel
Image	ScoreHMR (Stathopoulos et al., 2024)	35.1	36.8	43.8	43.8
ResNet-152	Triangulation (Hartley & Zisserman, 2004)	23.5	25.1	57.0	63.6
ResNet-152	FMPose3D‡ (Wang et al., 2026)	38.0	37.4	51.6	50.4
ResNet-152	DPoser-X (body)‡ (Lu et al., 2025)	31.5	32.9	41.7	42.9
ResNet-152	RUMPL (Ghasemzadeh et al., 2026)	26.4	27.1	39.7	41.4
ResNet-152	SkelSplat (Bragagnolo et al., 2026)	22.9	24.7	36.4	41.7
ResNet-152	NRDF‡ (He et al., 2024a)	27.2	29.0	34.7	38.0
ResNet-152	**Ours**	22.3	24.2	30.7	34.9
LT†	LT (Iskakov et al., 2019)	**20.2**	**23.0**	71.5	153.3
LT†	**Ours**	20.6	23.4	32.9	38.1
AdaFuse†	AdaFuse (Zhang et al., 2021)	22.2	24.0	35.6	40.8
AdaFuse†	**Ours**	21.9	23.7	**29.8**	**34.4**

Table 6: **Performance on Human3.6M, clean and occluded.** Absolute and Root-relative MPJPE (mm) ↓. † ResNet-152-based detector. ‡Our multi-view extensions of the monocular FMPose3D, DPoser-X (body), and NRDF.

findings suggest that GenTri improves robustness to occlusion while largely preserving reconstruction accuracy on clean observations as well.

B Experimental and Implementation Details

B.1 Datasets and Training

Mouse data and occlusion protocol. We train the mouse 2D detector on synthetic multi-view images rendered with MoReMouse's Gaussian avatar (Zhong et al., 2026) and the flow prior on the corresponding 3D poses. We use the articulated mouse-model fits released with MoReMouse, produced using the MAMMAL fitter (An et al., 2023) on the DANNCE `markerless_mouse_1` sequence (Dunn et al., 2021). We obtain 22-joint coordinates for both training poses and in-domain test labels from these meshes using a mesh-to-joint mapper that we fit to DANNCE's manually annotated 3D keypoints. Detector training uses 1,100 poses sampled every tenth frame from source frames 0–10,990, rendered from 36 views to produce 39,600 images. The flow prior uses 990 poses from the early part of the same sequence. The in-domain (InD) open-field test set contains 200 frames from six views, sampled every tenth frame from source frames 15,500–17,490. These test poses are excluded from both prior training and all training renderings. The two out-of-domain (OoD) sets are THM (Zhao et al., 2023) (99 frames, six views), recorded in a different open-field arena, and our newly collected and manually annotated treadmill set, which differs in apparatus, camera configuration, and behavior. We perform no fine-tuning or adaptation on either OoD set.

All mouse sets use a common 22-joint skeleton rooted at the front spine. The treadmill set contains 99 labeled frames from two female C57BL/6J mice, with 47 and 52 frames from the respective animals, recorded with four synchronized cameras. We annotate visible joints using Label3D (Dunn et al., 2021) and triangulate them across labeled views to obtain 3D ground truth. THM uses the released six-view annotations for 16 of the 22 joints (excluding wrists, elbows, and ankles), and errors are computed only over labeled joints (Zhao et al., 2023).

Following prior work on synthetic-occlusion evaluation (Zhang et al., 2020; Liu et al., 2022; Wang et al., 2024), we construct Mouse-Occluded by overlaying cage bars and one object occluder (e.g., feed, bedding, or a water bottle) per view on the original test images. Because the tail can extend the bounding box to include substantial empty space, we place object occluders around joint keypoints to cover the animal rather than sampling positions uniformly within the box. We retain the original 3D annotations for paired clean/occluded evaluation.

Human data and occlusion protocol. Human3.6M (Ionescu et al., 2014) follows the standard multi-view protocol (Iskakov et al., 2019; Zhang et al., 2021) (train S1, S5, S6, S7, S8; test S9, S11). Human3.6M-Occluded (Bragagnolo et al., 2025) overlays PASCAL VOC (Everingham et al., 2012) instances on three of the four views, with the occluder size set relative to the person bounding box. Our only change to the corruption protocol is to use tight ground-truth keypoint boxes to set the occluder scale. All methods compared on Human3.6M-Occluded in Sec. 4 are rerun on the same corrupted frames using a common evaluation script.

Detectors and confidences. For each species, we evaluate the base per-view detector and its LT and AdaFuse variants on the same backbone. LT uses differentiable triangulation (Iskakov et al., 2019), while AdaFuse learns cross-view fusion (Zhang et al., 2021). The human detector is a SimpleBaseline ResNet-152 (Xiao et al., 2018) trained on MPII (Andriluka et al., 2014) and Human3.6M. The mouse 2D detector is a DINOv2 (Oquab et al., 2024) heatmap head trained on the synthetic renderings with domain randomization, RandAugment (Cubuk et al., 2020), and novel-view augmentation. Its LT and AdaFuse variants are trained on the same renderings with the same backbone, so the three detectors of Table 2 differ only in how views are combined. The confidence α_{mj} of Sec. 3.2 is the heatmap peak for the base detectors, the algebraic confidence for LT, and the learned view weight for AdaFuse, normalized over views for each joint.

(a) Mouse Clean

	Open-field *InD*		**Open-field** *OoD*		**Treadmill** *OoD*	
Augmentation	Abs	Rel	Abs	Rel	Abs	Rel
none (no aug)	5.9	7.2	56.4	58.4	37.5	46.0
+ domain randomization	**2.4**	**2.9**	5.2	6.4	4.7	7.9
+ novel-view aug (adopted)	2.4	3.1	**4.6**	**5.4**	**4.7**	**7.7**

(b) Mouse Occluded

	Open-field *InD*		**Open-field** *OoD*		**Treadmill** *OoD*	
Augmentation	Abs	Rel	Abs	Rel	Abs	Rel
none (no aug)	42.8	71.1	101.8	98.5	73.4	94.0
+ domain randomization	5.9	7.2	18.2	22.9	10.8	13.7
+ novel-view aug (adopted)	**5.1**	**6.5**	**16.1**	**22.8**	**9.5**	**12.0**

Table 7: **DINOv2 detector augmentation ablation.** Absolute and Root-relative MPJPE (mm) ↓ of ours (DINOv2) on (a) Mouse Clean and (b) Mouse-Occluded data; each row adds one training ingredient (no augmentation → +domain randomization → +novel-view augmentation, deployed).

Mouse detector augmentation ablation. Table 7 ablates the synthetic-render augmentation recipe of the DINOv2 detector. The second row combines render-time domain randomization with RandAugment, including its geometric operations and excluding horizontal flips. Domain randomization with RandAugment provides the largest reduction in error: without augmentation, absolute MPJPE reaches 38–102 mm under occlusion or domain shift. Novel-view augmentation expands training coverage from the six original views to 36 rendered viewpoints and further reduces absolute MPJPE on Mouse-Occluded from 5.9 to 5.1 mm on open-field, 18.2 to 16.1 mm on THM, and 10.8 to 9.5 mm on treadmill. On clean data, absolute MPJPE with and without novel-view augmentation differs by at most 0.7 mm.

Flow prior. We use a time-conditioned residual graph network as the flow prior v_θ, trained on normalized root-relative 3D poses with the flow-matching objective of Eq. 3. The same architecture is used for both species, with the graph adapted to each skeleton (J=17 human, J=22 mouse). Table 8 summarizes the prior and detector training settings.

Setting	Flow prior (both species)	Mouse 2D detector
Architecture	Graph network, width 384, depth 10	Frozen DINOv2-B/14 + heatmap head
Input / output	Root-relative 3D pose	448^2 image / 256^2 heatmaps
Optimizer	AdamW	AdamW
Learning rate	2×10^{-3}, cosine decay	7×10^{-4}, $\times 0.1$ at epoch 25
Batch / epochs	4096 / 2500	32 / 50
Weight decay / EMA	10^{-4} / 0.999	10^{-4} / —

Table 8: **Training settings.**

Algorithm 2 GenTri Inference (detailed, N hypotheses)

Require: detections $(\mathbf{y}_{mj}, \alpha_{mj})$, cameras Π_m, prior v_θ, scale s, steps S, damping κ, hypotheses N

$\Delta t \leftarrow 1/S$; for $k = 1, \ldots, N$: $\mathbf{x}_0^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\tau}_0^{(k)} \leftarrow$ confidence-weighted root triangulation
for $t = 0, \Delta t, \ldots, 1{-}\Delta t$ (each hypothesis k in parallel; index k omitted) **do**
  $\hat{\mathbf{x}}_1 \leftarrow \mathbf{x}_t + (1{-}t)\, v_\theta(\mathbf{x}_t, t)$; $\mathbf{p}_t \leftarrow s\, \hat{\mathbf{x}}_1 + \boldsymbol{\tau}_t$ ▷ predicted endpoint, absolute pose
  $\mathbf{r}_{mj} \leftarrow \mathbf{y}_{mj} - \Pi_m(\mathbf{p}_{t,j})$; $C_t \leftarrow \mathrm{median}_{k,m,j} \|\mathbf{r}_{mj}\|_2$ ▷ residuals, Cauchy scale (Eq. 15)
  $\omega_{mj} \leftarrow \alpha_{mj}\, \rho'(r_{mj})/r_{mj}$
  $\mathbf{g}_j := \nabla_{\mathbf{p}_{t,j}} \mathcal{E} \leftarrow \sum_m \omega_{mj} \big(\mathbf{J}_{mj}^{\mathrm{rep}}\big)^\top \mathbf{r}_{mj}$; $\mathbf{H}_j \leftarrow \sum_m \omega_{mj} \big(\mathbf{J}_{mj}^{\mathrm{rep}}\big)^\top \mathbf{J}_{mj}^{\mathrm{rep}}$ ▷ Eq. 10
  $\mathbf{M}_{t,j} \leftarrow \mathbf{H}_j + \kappa\, \mathrm{diag}(\mathbf{H}_j)$ ▷ Eq. 11
  $\mathbf{x}_{t+\Delta t} \leftarrow \mathbf{x}_t + \Delta t\, v_\theta(\mathbf{x}_t, t) - (t{+}\Delta t)/s\, \mathbf{M}_t^{-1}\mathbf{g}$ ▷ Eq. 7
  $\boldsymbol{\tau}_{t+\Delta t} \leftarrow \boldsymbol{\tau}_t - \mathbf{H}_\tau^{-1}\mathbf{g}_\tau$ ▷ translation, Eq. 13
end for
$\mathbf{p}_1^{(k)} \leftarrow s\mathbf{x}_1^{(k)} + \boldsymbol{\tau}_1^{(k)}$ for each k ▷ final absolute poses
$k^\star \leftarrow \arg\min_k \mathcal{E}(\mathbf{p}_1^{(k)})$ ▷ best of N
return $\mathbf{p}_1^{(k^\star)}$

B.2 Inference and Hyperparameters

Algorithm 2 details inference, including the translation update and hypothesis selection.

Inference settings. Unless stated otherwise, every experiment uses $S = 10$ guidance steps, Marquardt damping $\kappa = 0.1$ (Eq. 11), $N = 4$ hypotheses, and the Cauchy penalty with the per-step median scale C_t, for both species and all detectors; no hyperparameter is set per dataset. Figure 7 reports the sensitivity to κ, the only constant of the guidance step. We initialize global translation by confidence-weighted triangulation of the root.

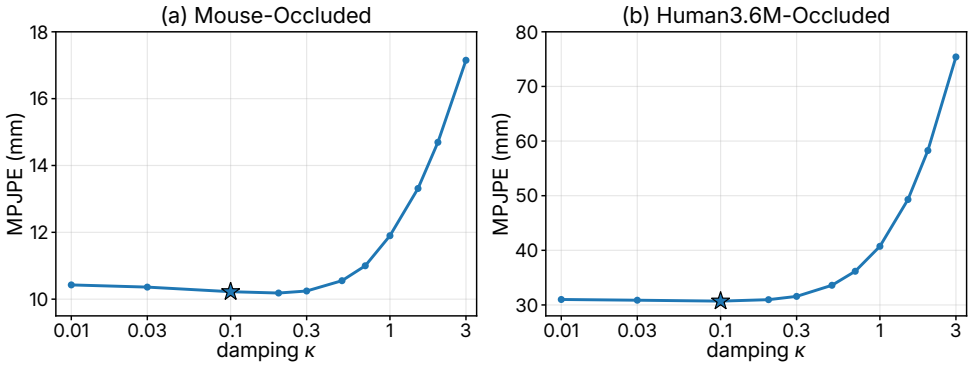


Figure 7: **Sensitivity to the damping κ.** Absolute MPJPE (mm) ↓ on the occluded test sets as κ varies at the deployed S=10; ★: deployed κ=0.1, selected on mouse clean data and transferred to human.

Global translation. Since the flow prior is trained on root-relative poses, the global translation $\boldsymbol{\tau}_t$ is optimized separately in absolute 3D space. Because translation moves all joints equally, we

aggregate the joint-wise terms as $\nabla_{\boldsymbol{\tau}_t}\mathcal{E}(\mathbf{p}_t) = \sum_{j=1}^{J} \nabla_{\mathbf{p}_{t,j}}\mathcal{E}(\mathbf{p}_t)$ and $\mathbf{H}_\tau = \sum_{j=1}^{J} \mathbf{H}_j$, and take one Gauss–Newton step at $t' = t + \Delta t$:

$$\boldsymbol{\tau}_{t'} = \boldsymbol{\tau}_t - \mathbf{H}_\tau^{-1}\nabla_{\boldsymbol{\tau}_t}\mathcal{E}(\mathbf{p}_t). \tag{13}$$

For numerical stability, the implementation replaces $\mathbf{H}_\tau^{-1}$ with $(\mathbf{H}_\tau + \lambda_\tau \mathbf{I})^{-1}$, where $\lambda_\tau = \max(10^{-3}\,\mathrm{tr}(\mathbf{H}_\tau)/3, 10^{-12})$.

Multi-hypothesis inference and selection. For each frame, we initialize N pose hypotheses from independent Gaussian noise samples and run guided sampling in parallel using the same 2D observations. Inspired by reprojection-based selection in JPMA (Shan et al., 2023), we use the final observation energy $\mathcal{E}$ as a frame-level hypothesis score and select

$$k^\star = \arg\min_{k\in\{1,\dots,N\}} \mathcal{E}\big(\mathbf{p}_1^{(k)}\big). \tag{14}$$

Here $\mathbf{p}_{1,j}^{(k)} = s\mathbf{x}_{1,j}^{(k)} + \boldsymbol{\tau}_1^{(k)}$ denotes the final absolute joint positions. The final prediction is $\mathbf{p}_1^{(k^\star)}$.

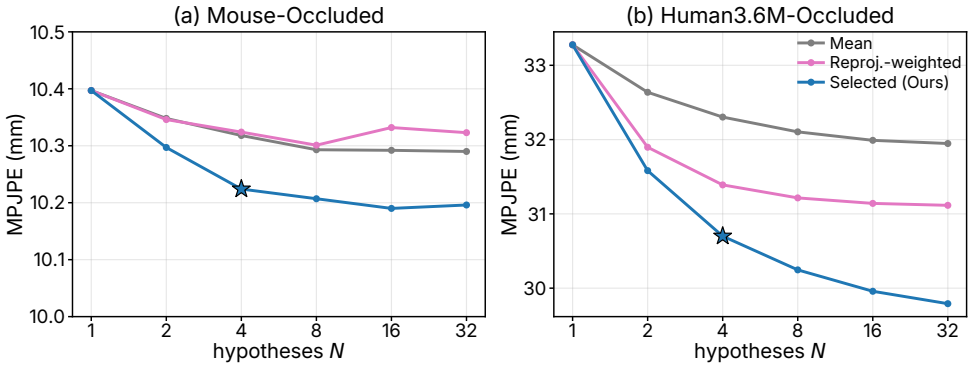


Figure 8: **Selection rule versus the number of hypotheses N.** Absolute MPJPE (mm) ↓ of three ways of turning N hypotheses into one pose. The star (★) marks the deployed setting $N{=}4$.

Figure 8 compares three ways of reducing N hypotheses to one pose. Selected (Eq. 14, ours) keeps the complete hypothesis with the lowest final robust energy; Mean averages them; Reproj.-weighted averages them per joint with softmax weights on the reprojection error (Wang et al., 2026). Increasing the number of hypotheses provides little additional benefit on mouse, whereas Human3.6M benefits from multiple hypotheses. We use $N{=}4$ for both species in the reported experiments.

Scale rule	**Mouse**	**Human**
Global median	12.3	35.0
Per-step median (Ours)	10.2	30.7

Table 9: **Adaptivity of the Cauchy scale.** Absolute MPJPE (mm) ↓ with the deployed solver; only the rule that sets the scale changes. Per-step median: Eq. 15, re-estimated for each frame at every step. Global median: one constant per test set, the median of all initial reprojection residuals of that set.

Adaptive Cauchy scale. The Cauchy scale is determined directly from reprojection residuals computed at inference time, without separate tuning of the scale parameter. Specifically, at flow step t we set, independently for each frame,

$$C_t = \mathrm{median}_{k,m,j}\; r_{mj}\big(\mathbf{p}_t^{(k)}\big), \tag{15}$$

the median over the pool of $N{\times}M{\times}J$ reprojection residuals of all hypotheses k, views m, and joints j at the current step, re-estimated at every step. We hold C_t fixed during each update. Replacing it with a global median increases MPJPE by 2.1 mm on mouse and 4.3 mm on human (Table 9).

Variance across seeds. All stochastic inference results in the appendix are averaged over three seeds. We vary inference seeds while keeping the trained prior and detectors fixed. GenTri entries in

Tables 2–4 are three-seed means. The sample standard deviation (1σ) is at most 0.1 mm in Tables 2, 3, and 4(b), 0.2 mm in Table 4(a), and, per point of Figure 5, 0.2 mm on mouse and 0.6 mm on human. Every curve of Figure 5 is a three-seed mean. Baseline reconstructions are deterministic given the 2D detections, except FMPose3D‡ and DPoser-X (body)‡, whose samplers draw noise at inference. For these two, we also report three-seed means; they differ from the corresponding single-seed runs by at most 0.01 mm on both the clean and occluded splits.

C Baseline Adaptations

C.1 Adaptation of Guidance Rules to the Flow Prior

Figure 5 compares guidance rules adapted to the same pose flow prior and multi-view reprojection operator. We compare guidance methods whose correction mechanisms can be applied to multi-view reprojection, with the required flow-parameterization adaptations detailed below. All settings share 2D detections, confidence weights, initialization, and the translation solver of Eq. 13. Mouse experiments use base DINOv2 detections, and human experiments use base ResNet-152 detections. For each number of guidance steps, numerical guidance hyperparameters are selected on the clean counterparts of the evaluation frames and kept fixed under occlusion. GenTri transfers the mouse-selected κ to human experiments. Table 10 summarizes the baseline implementations in our framework.

Method	**Loss**	**Network derivatives**	**Correction**	**State update**
Robust guidance				
Robust-CG (Zheng et al., 2026)	Huber	None	Gradient (inner loop)	$t'(\hat{\mathbf{x}}_1 + \Delta\hat{\mathbf{x}}_1) + (1-t')\boldsymbol{\epsilon}$
RDP (Yang et al., 2026)	IMQ	VJP	Gradient	$\mathbf{b}_t + \Delta\mathbf{x}_t$
Effect of the correction rule				
CLAMP (Shin et al., 2026)	L_2	VJP	Gauss–Newton	Langevin
DPS (Chung et al., 2023)	L_2	VJP	Gradient	$\mathbf{b}_t + \Delta\mathbf{x}_t$
Our mechanism				
Ours w/o GN	Cauchy	None	Gradient	$\mathbf{b}_t + t'\,\Delta\hat{\mathbf{x}}_1$
Ours w/ VJP grad.	Cauchy	VJP	Gauss–Newton	$\mathbf{b}_t + t'\,\Delta\mathbf{x}_t$
Ours w/ full Jac.	Cauchy	Full Jacobian	Gauss–Newton	$\mathbf{b}_t + \Delta\mathbf{x}_t$
Ours	Cauchy	None	Gauss–Newton	$\mathbf{b}_t + t'\,\Delta\hat{\mathbf{x}}_1$

Table 10: **Implementation of the guidance rules.** DPS* and CLAMP* change only the listed loss to Cauchy. VJP: vector–Jacobian product through the flow network. $t' = t + \Delta t$; $\mathbf{b}_t = \mathbf{x}_t + \Delta t\, v_\theta(\mathbf{x}_t, t)$: unguided Euler step; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$: fresh noise.

Adapting diffusion guidance to flow matching. Our flow prior uses a standard Gaussian source, independent source–target coupling, and the affine path $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. These satisfy the uncoupled affine Gaussian path assumption of Feng et al. (2025), under which diffusion guidance can be transferred to flow matching with appropriate schedule factors. We therefore express the baseline corrections in flow coordinates; the method-specific corrections and sampling transitions are detailed below.

DPS and RDP. Unlike GenTri's endpoint Gauss–Newton correction without network derivatives, DPS (Chung et al., 2023) and RDP (Yang et al., 2026) apply first-order corrections to the flow state through the network Jacobian. For the observation energy $\mathcal{E}_\rho$ (Eq. 9 with penalty ρ), the correction is

$$\Delta\mathbf{x}_t = -a_t\, s\, \mathbf{J}_{\mathrm{net}}^{\top} \nabla_{\mathbf{p}_t} \mathcal{E}_\rho(\mathbf{p}_t), \qquad a_t = \frac{\lambda\,\Delta t}{\sqrt{\mathcal{E}_\rho}}\, c_t,$$

with $\mathbf{J}_{\mathrm{net}}$ the endpoint Jacobian. Both methods pull the gradient back through the flow network with one vector–Jacobian product and differ in the observation penalty. Both correct the state, so porting their VP-diffusion step to the linear interpolant multiplies it by $c_t = t^2 + (1-t)^2$ (the change of variables $\mathbf{z}_\xi = \mathbf{x}_t / \sqrt{t^2 + (1-t)^2}$). DPS uses an L_2 loss and a VJP through the predicted endpoint.

RDP retains the DPS update and changes the observation model to inverse multi-quadric weighting:

$$\rho(r) = \frac{r^2}{2\sqrt{1 + r^2/C_t^2}}, \qquad C_t = \mathrm{Quantile}_{0.75}\big(\{r_{mj}(\mathbf{p}_t^{(k)})\}\big).$$

The quantile is computed per frame over valid views, joints, and hypotheses. Like GenTri, RDP uses a robust observation model, but applies first-order VJP guidance instead of endpoint Gauss–Newton correction.

Robust-CG. Like GenTri, our Robust-CG adaptation (Zheng et al., 2026) corrects the endpoint without network derivatives. We retain the Huber loss with a fixed initial-residual scale and the original observation correction $\bar{\mathbf{y}}$, replacing GenTri's Cauchy loss with per-step scale C_t. The endpoint is corrected by an iterative proximal solve instead of GenTri's single damped GN step. The corrected endpoint is then combined with fresh Gaussian noise, $\mathbf{x}_{t'} = t'(\hat{\mathbf{x}}_1 + \Delta\hat{\mathbf{x}}_1) + (1 - t')\boldsymbol{\epsilon}$, replacing the current source estimate $\hat{\mathbf{x}}_0$ preserved by GenTri.

Using the common observation energy $\mathcal{E}_\rho$ of Eq. 9, the endpoint correction approximately solves

$$\Delta\hat{\mathbf{x}}_1 \approx \arg\min_{\mathbf{d}} \left\{ \frac{\|\mathbf{d}\|_2^2}{2q_t} + \frac{\lambda_{\mathrm{CG}}\sigma_t^2}{\delta^2}\mathcal{E}_\rho(\mathbf{p}_t + s\mathbf{d}) \right\}, \qquad \rho(r) = \begin{cases} \frac{1}{2}r^2, & r \le \delta, \\ \delta(r - \frac{1}{2}\delta), & r > \delta, \end{cases}$$

where $q_t = (1-t)^2/\big((1-t)^2 + t^2\big)$ and λ_{CG} balances the two terms. We set the Huber threshold to $\delta = 1.345 \cdot 1.4826 \cdot \mathrm{MAD}(\{r_{mj}^{\mathrm{init}}\})$ (Huber, 1981), using the conventional Huber tuning constant and Gaussian-consistent MAD scaling. We solve with the conjugate-gradient rule of Algorithm 2 of Zheng et al. (2026), refreshing the Huber weights at each iteration, with K tuned to 20 for mouse and 100 for human.

CLAMP. Our CLAMP adaptation retains the correction and sampling transition of Algorithm 1 of Shin et al. (2026) in flow coordinates. CLAMP retains the denoiser Jacobian and corrects the noisy state, whereas GenTri uses joint-wise endpoint GN without network derivatives. CLAMP also differs from GenTri in its curvature approximation and damping. After this correction, CLAMP applies its variance-preserving Langevin transition in place of GenTri's deterministic transition that preserves the current source estimate.

For reproduction, in EDM coordinates $\tilde{\mathbf{x}}_t = \mathbf{x}_t/t$ we use $\mathbf{J}_F = t\mathbf{J}_{\mathrm{net}}$ and

$$\Delta\tilde{\mathbf{x}}_t = -\big(s^2\mathbf{J}_F^\top\mathbf{H} + \beta_t\mathbf{R}_t\big)^{-1} s\,\mathbf{J}_F^\top\nabla_{\mathbf{p}_t}\mathcal{E}_\rho(\mathbf{p}_t),$$

$$\mathbf{R}_t = 2\mathbf{I} + \frac{\mathbf{d}_t\mathbf{d}_t^\top}{\|\mathbf{d}_t\|_2^2}, \qquad \beta_t = \frac{\sigma_n^2\,\|\mathbf{d}_t\|_2}{\|s\,\mathbf{J}_F^\top\nabla_{\mathbf{p}_t}\mathcal{E}_\rho(\mathbf{p}_t)\|_2\,\sigma_{t'}^2},$$

where $\mathbf{d}_t = \tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_1$, and the gradient and curvature use the chosen observation penalty. We retain the one-sided $s^2\mathbf{J}_F^\top\mathbf{H}$ operator, not the two-sided $s^2\mathbf{J}_F^\top\mathbf{H}\mathbf{J}_F$ system. The inverse is applied with five matrix-free GMRES iterations, each using one VJP; the measurement scale σ_n is selected on clean data. After the Langevin transition to $\sigma_{t'}$, we map back by $\mathbf{x}_{t'} = t'\tilde{\mathbf{x}}_{t'}$.

GenTri and its step-rule variants. Ours w/o GN. We replace the Gauss–Newton solve by the first-order rule of Eq. 5 applied to our energy, i.e. a scalar step along the raw endpoint gradient,

$$\Delta\hat{\mathbf{x}}_1 = -\lambda\,\Delta t\,s\,\nabla_{\mathbf{p}_t}\mathcal{E}(\mathbf{p}_t),$$

where $s\nabla_{\mathbf{p}_t}\mathcal{E}$ is the normalized endpoint gradient. We keep everything else, including the t' transport of Eq. 7.

Ours w/ VJP grad. We pull back only the gradient through the flow network using a single vector–Jacobian product, while retaining the endpoint curvature and damping of Eq. 11:

$$\Delta\mathbf{x}_t = -\frac{1}{s}\mathbf{M}_t^{-1}\mathbf{J}_{\mathrm{net}}^\top\nabla_{\mathbf{p}_t}\mathcal{E}(\mathbf{p}_t).$$

The curvature (endpoint) and the gradient (state) are therefore expressed in different coordinates; the displacement enters the state with the same t' factor as GenTri.

Ours w/ full Jac. We pull back both the gradient and curvature to form the state-space Gauss–Newton system,

$$\Delta\mathbf{x}_t = -\frac{1}{s}\left(\mathbf{J}_{\mathrm{net}}^\top\mathbf{H}\mathbf{J}_{\mathrm{net}} + \kappa\,\mathrm{diag}(\mathbf{H})\right)^{-1}\mathbf{J}_{\mathrm{net}}^\top\nabla_{\mathbf{p}_t}\mathcal{E}(\mathbf{p}_t).$$

Our implementation assembles $\mathbf{J}_{\mathrm{net}}$ of Eq. 12 with $3J$ backward passes per frame and solves the coupled system; its displacement is a state-space step and is added directly, without a t' or Δt factor.

C.2 Published Objective vs. Our Cauchy Objective

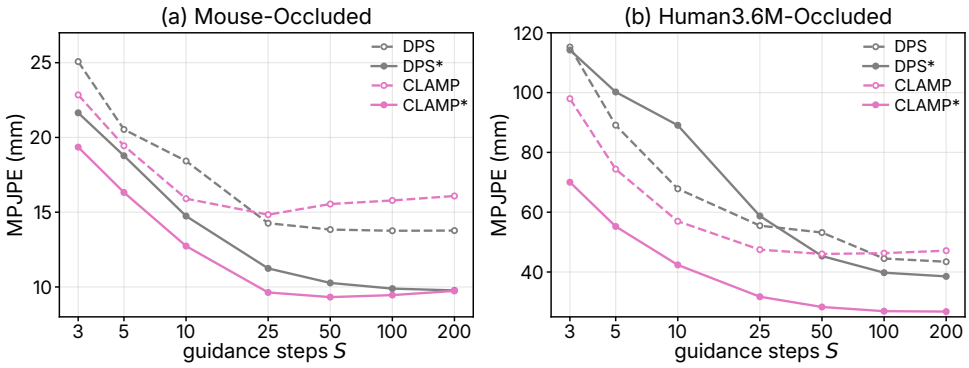


Figure 9: **Effect of the robust objective on the adapted baselines.** Absolute MPJPE (mm) ↓ versus guidance steps S for DPS and CLAMP with their published L_2 likelihood (dashed) and with our Cauchy objective (solid, *).

Figure 5(a) reports DPS and CLAMP with our Cauchy objective to compare correction rules under a shared observation objective. Figure 9 shows that replacing the published L_2 likelihood with Cauchy reduces error for both methods on Mouse-Occluded across the evaluated step budgets. On Human3.6M-Occluded, CLAMP also improves across the evaluated budgets, whereas the effect on DPS depends on the number of steps. The Cauchy-matched variants therefore serve as stronger baselines for both methods on Mouse-Occluded and for CLAMP on Human3.6M-Occluded.

C.3 Human Pose Baselines

The triangulation baseline in Tables 3 and 6 uses confidence-weighted DLT with the per-view 2D detector confidences. Baselines with a native multi-view pipeline (LT, AdaFuse, RUMPL, SkelSplat) are run with their own initialization and settings. Our multi-view extensions (‡) of FMPose3D, DPoser-X (body), and NRDF build on the released single-view code of Wang et al. (2026), Lu et al. (2023; 2025), and He et al. (2024a), respectively, and start from the same confidence-weighted triangulation as GenTri. FMPose3D selects the cross-view reprojection-consistent hypothesis. For DPoser, we use the body model from the released DPoser-X implementation (Lu et al., 2023; 2025) and extend its reprojection term to multiple views. DPoser-X (body) and NRDF retain their released SMPL objectives, schedules, and hyperparameters, aggregating the reprojection term over the four calibrated views. Both start from the same confidence-weighted triangulation as ours, with their AMASS priors fine-tuned on Human3.6M training poses. RUMPL (Ghasemzadeh et al., 2026) is adapted to our protocol: its lifter is pretrained on AMASS and fine-tuned on the Human3.6M training split with the same ResNet-152 detections as the other rows, and we report its final checkpoint. SkelSplat (Bragagnolo et al., 2026) has no learned weights; we run its per-scene test-time optimization on the same occluded frames with the same ResNet-152 detections, initialized as in its released pipeline by plain DLT triangulation of those detections, and aggregate its output over all evaluation frames. ScoreHMR (Stathopoulos et al., 2024) uses the released HMR2 and diffusion checkpoints to estimate and refine SMPL poses from occluded image crops. The absolute position is triangulated from the same ResNet-152 detections. ScoreHMR additionally uses image-based HMR2 predictions, so its inputs differ from those of the shared-detector baselines. All rows of Table 3 use the same occluded frames and the same evaluation script.

D Computational Cost

Table 11 reports per-step costs for the methods in Figures 5 and 9 on an otherwise idle RTX 4090, processing one frame with $N{=}4$ hypotheses. Using precomputed detections, we time the inference loop with a wall-clock timer and CUDA synchronization before and after it. Timing includes prior evaluation, reprojection and derivatives, pose correction, and translation updates, but excludes model loading and 2D detection. We estimate incremental step cost as $\ell = (T_{50} - T_{10})/40$, where T_S is

Method	Mouse	Human
RDP	15.0	15.0
Robust-CG	37.2	165.5
DPS	13.8	13.6
DPS*	14.9	14.7
CLAMP	32.4	32.3
CLAMP*	33.9	33.6
Ours w/o GN	7.3	7.3
Ours w/ VJP grad.	16.5	16.1
Ours w/ full Jac.	207.4	162.9
Ours	8.6	8.7

Table 11: **Per-step guidance latency.** Milliseconds per guidance step on an RTX 4090, one frame with $N{=}4$ hypotheses, excluding fixed loop overhead; measurement procedures are described in the text. The time axis of Figure 5 is S times these values. *Our Cauchy-matched variants of DPS and CLAMP.

the S-step loop time, to remove fixed overhead. We report the median of three estimates from one session, interleaving methods and species. Our full-Jacobian variant explicitly constructs the endpoint Jacobian, increasing its computational cost.

E Limitations

Our occlusion benchmarks are synthetic, so they control the corruption rate but not the correlation between occluders and behavior found in real enclosures. Because each joint's correction is invariant to a common scaling of its view weights, the Cauchy energy suppresses views that disagree with the others but not a joint that is detected consistently wrong in every view; such joints rely on the pose prior. The damped correction also assumes calibrated cameras.

F Choice of Generative Prior and Sampler

Comparison setup. To assess how the generative prior and sampler affect reconstruction, we compare flow matching with Euler integration against diffusion with DDIM (Song et al., 2021a) and DDPM (Ho et al., 2020) under the same Gauss–Newton guidance. The flow-matching and cosine-VP diffusion priors (Nichol & Dhariwal, 2021) share the architecture, training data, and optimization schedule. Detections, initialization, Cauchy observation energy, translation solver, and evaluation seeds are fixed.

Diffusion start times. For diffusion, we set $t_0 = 0.0065$ by converting the terminal SNR of the continuous-time linear-β VP schedule of Song et al. (2021b) to affine time. We additionally choose $t_0 = 0.05$ to test whether a later start improves stability at small step budgets, using the same value across samplers and species.

Results. Table 12 reports results at $S = 10$, and Figure 10 varies S from 5 to 50. On Mouse-Occluded, flow matching stays within 10.2–10.9 mm across the evaluated step budgets. Diffusion is more sensitive to the start time at small budgets. At $S = 10$, changing t_0 from 0.0065 to 0.05 reduces DDIM error from 375.5 to 10.8 mm and DDPM error from 94.4 to 10.8 mm, compared with 10.2 mm for flow matching. On Human3.6M-Occluded, the two start times yield similar results. At $S = 10$, flow matching achieves 30.7 mm, DDIM 30.8 mm, and DDPM 32.9–33.1 mm. At $S = 50$, DDPM reaches 29.5 mm, compared with 30.2 mm for flow matching.

Choice of default. These results show that diffusion's sensitivity to the start time differs between the mouse and human priors. Flow matching performs reliably on both, while diffusion achieves comparable accuracy with a suitable start time. We use flow matching with Euler integration as the default based on its consistent performance in these evaluations.

(a) Mouse

Method	Clean			Occluded		
	Open-field *InD*	Open-field *OoD*	Treadmill *OoD*	Open-field *InD*	Open-field *OoD*	Treadmill *OoD*
Diffusion / DDIM	436.8	465.5	207.0	438.8	476.6	211.1
Diffusion / DDIM ($t_0 = 0.05$)	**2.2**	4.6	5.1	**4.7**	16.6	10.9
Diffusion / DDPM	92.4	112.6	53.0	94.8	130.2	58.3
Diffusion / DDPM ($t_0 = 0.05$)	2.4	4.7	5.4	5.1	16.3	10.9
FM / Euler (Ours)	2.4	**4.6**	**4.7**	5.1	**16.1**	**9.5**

(b) Human3.6M

Method	Human3.6M	Human3.6M-Occluded
Diffusion / DDIM	22.4	30.8
Diffusion / DDIM ($t_0 = 0.05$)	22.4	30.8
Diffusion / DDPM	23.9	33.1
Diffusion / DDPM ($t_0 = 0.05$)	23.8	32.9
FM / Euler (Ours)	**22.3**	**30.7**

Table 12: **Flow-matching and diffusion priors with Gauss–Newton guidance** ($S = 10$)**.** Absolute MPJPE (mm) ↓.

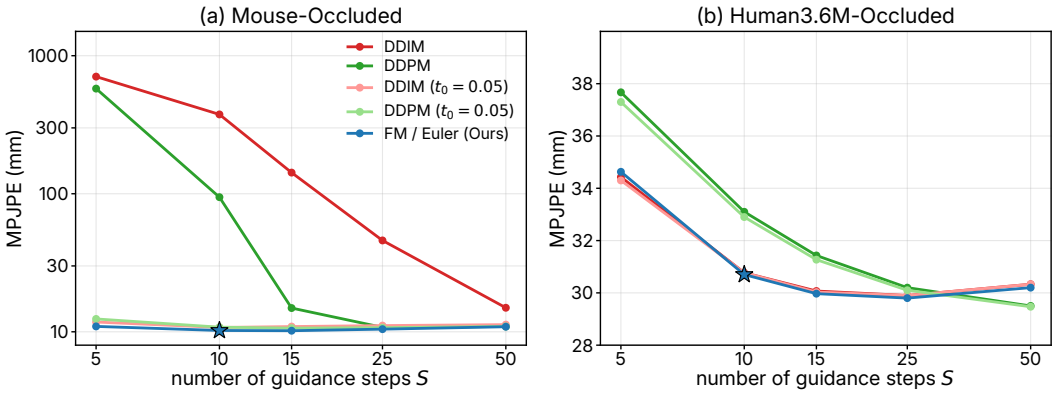


Figure 10: **Priors and samplers across step budgets.** Absolute MPJPE (mm; ↓) on occluded data. The star (★) marks the deployed setting S=10. **(a)** Mouse results averaged across three domains, shown on a logarithmic scale. **(b)** Human3.6M results averaged across actions using ResNet-152 detections.