MoDiTac: Modality-Decoupled Diffusion Transformer for Preference-Controlled Tactile Generation

Anonymous authors
Paper under double-blind review

Abstract

Tactile data capture local object deformation and contact properties, but their collection relies on specialized sensors and physical interaction, making large-scale acquisition more difficult than for visual data. Visual-to-tactile generation can alleviate this bottleneck through scalable synthesis of tactile data. Existing tactile generators typically emphasize paired reconstruction fidelity, yielding reconstruction-oriented outputs whose generation characteristics cannot be readily adjusted to different downstream task requirements. MoDiTac introduces a continuous preference coordinate, enabling a single generator to span a continuum from reconstruction-oriented to perception-oriented tactile generation. Within a unified diffusion transformer, tactile-guided visual selection extracts tactile-relevant visual evidence from the full scene according to the current generative state, alleviating the spatial-support mismatch between scene-level visual input and local tactile response; modality-decoupled projections reduce interference between heterogeneous representations before joint self-attention; and preference-controlled global-context injection maps the continuous preference coordinate to different reconstruction–perception operating points. Experiments show that material recognition and Press–Sliding recognition favor different regions of the reconstruction–perception continuum, demonstrating the need for task-selectable tactile generation.

1 Introduction

Tactile perception is essential for embodied intelligence, multimodal understanding, and physical interaction (Calandra et al., 2018). Unlike vision, which primarily conveys global semantics, tactile sensors directly capture local surface deformation (Yuan et al., 2017; Lambeta et al., 2020). Therefore, in tasks such as robotic manipulation, material recognition, and contact reasoning, tactile sensing can provide information on local contact deformation and surface microstructure that is difficult to obtain directly from visual observations. However, acquiring high-quality tactile data relies on direct physical interaction and specialized sensing hardware, making large-scale collection expensive. This limitation has motivated visual-to-tactile generation, which synthesizes tactile signals from visual observations (Li et al., 2019; Lee et al., 2019b; Yang et al., 2023; Song et al., 2026).

However, a single generative objective centered on paired reconstruction does not necessarily suit all downstream tasks. The perception–distortion (P-D) trade-off in image restoration arises in visual-to-tactile generation (Blau & Michaeli, 2018); under mean-squared-error-dominated optimization, diffusion models tend toward conservative posterior-mean solutions, yielding lower paired distortion while often weakening the perceptual statistics of tactile textures. Different downstream tasks place different emphasis on paired fidelity to the recorded tactile target (reconstruction-oriented) and perceptual realism in tactile texture and distributional statistics (perception-oriented). In the preference sweep on the evaluated datasets, material recognition favors the reconstruction-oriented region, whereas Press–Sliding recognition favors the perception-oriented region (Figure 1). These contrasting preferences expose a fundamental limitation of a single generative operating point: a fixed output is difficult to satisfy different downstream tasks simultaneously. Visual-to-tactile generation therefore calls for reconstruction–perception preference to be modeled as a continuous control variable, enabling a single generator to cover multiple operating points that can be selected according to task

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.

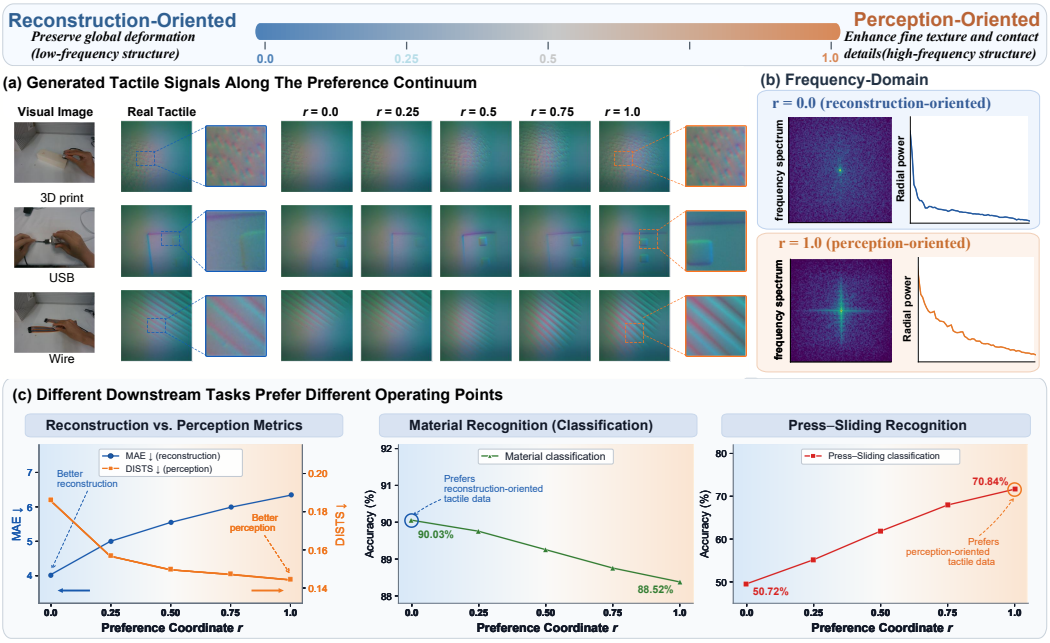


Figure 1: Preference-controlled tactile generation and task-dependent utility. (a) Outputs from one model across preference coordinates r. (b) Endpoint frequency characteristics. (c) Reconstruction–perception trade-off and task-dependent operating points.

requirements. An architecture that adjusts these operating points must address differences in spatial support and representation conflicts between vision and touch. Visual observations cover the background, manipulated object, and sensor, whereas the target tactile response is associated with only a limited local region. Existing cross-modal tactile generation methods fuse visual, textual, and tactile conditions into a shared representation space (Fang et al., 2024; Dou et al., 2024); because tactile signals characterize local surface deformation and fine-grained texture, whereas natural images contain spatial context and semantic information, prematurely sharing these representations can bias the generation process toward scene-level visual features and weaken fine-grained tactile structures. Existing methods establish cross-modal correspondence through hardware calibration, explicit 3D reconstruction, or segmentation priors (Li et al., 2019; Fu et al., 2024; Zhong et al., 2023; 2025; Song et al., 2026), but such approaches are resource-intensive and difficult to scale.

MoDiTac addresses the vision–tactile spatial-support mismatch, conflicts between heterogeneous representations, and reconstruction–perception preferences associated with downstream task requirements within a unified diffusion Transformer. MoDiTac selects visual evidence according to the tactile generation state and uses modality-decoupled projections to reduce interference between heterogeneous conditions before joint self-attention; the preference coordinate modulates the contribution of global visual context to the generative trajectory. Consequently, a single model can cover a range of operating points from reconstruction-oriented to perception-oriented generation. The main contributions are summarized as follows:

- MoDiTac combines tactile-guided full-frame visual selection with modality-decoupled projections to address the vision–tactile spatial-support mismatch and pre-fusion representation conflicts within a unified diffusion Transformer.
- Reconstruction–perception preference is formulated as an explicit continuous control variable, enabling a single model to cover operating points with different paired reconstruction fidelity and perceptual realism rather than being restricted to a single fixed configuration.
- Downstream evaluations show that material recognition and Press–Sliding recognition favor different regions of the reconstruction–perception continuum, demonstrating task-dependent preferences for generated tactile characteristics.

2 Related Work

2.1 Visual-to-Tactile Generation

Visual-to-tactile generation reduces the cost of physical data acquisition by predicting tactile responses from readily available visual observations. The emergence of high-resolution vision-based tactile sensors (Yuan et al., 2017; Taylor et al., 2022; Lambeta et al., 2020) has led to several visual–tactile datasets and benchmarks that support cross-modal learning (Calandra et al., 2018; Lee et al., 2019b; Kerr et al., 2023; Yang et al., 2022; Fu et al., 2024; Cheng et al., 2025; Rodriguez et al., 2025). Early work typically formulated tactile synthesis as cross-modal translation and used conditional GANs to generate tactile signals from visual inputs (Li et al., 2019). Subsequent methods introduced residual fusion or feature-matching constraints to recover fine texture geometry (Lee et al., 2019a; Cai et al., 2021; Fang et al., 2024). Sim-to-real approaches have meanwhile sought to reduce the domain gap between simulated and physical environments (Wang et al., 2022; Si & Yuan, 2022; Higuera et al., 2023). Spatial correspondence remains challenging because visual scenes and tactile responses differ in support. Geometry-based approaches use calibration or explicit 3D priors, including NeRF representations of objects and sensor locations (Fu et al., 2024; Zhong et al., 2025; 2023; Dou et al., 2024; Stefani et al., 2025; Gao et al., 2024). Alternative approaches use hand segmentation (Yang et al., 2023) or shared VAE representations (Yang et al., 2023; Fang et al., 2024; Luo et al., 2025). MoDiTac instead learns generation-relevant visual support from the evolving tactile state and adapts visual and tactile projections before joint attention.

2.2 The Perception–Distortion Trade-off in Generative Restoration

Pixel distortion and structural similarity (Wang et al., 2004) complement perceptual and distributional measures such as DISTS and FID (Heusel et al., 2017; Zhang et al., 2018; Ding et al., 2020). The perception–distortion trade-off (Blau & Michaeli, 2018) formalizes their competing objectives. Under squared-error distortion, the optimal estimator is the posterior mean, which averages feasible solutions and can smooth fine structure (Blau & Michaeli, 2018; Ohayon et al., 2024); posterior sampling can improve distributional agreement at the cost of paired error (Adrai et al., 2023; Ohayon et al., 2024). This trade-off informs image restoration (Blau et al., 2018; Ledig et al., 2017; Wang et al., 2018; Saharia et al., 2022; Wang et al., 2024; Lin et al., 2024), including adversarial approaches (Goodfellow et al., 2014; Wang et al., 2018; Mentzer et al., 2020) and controllable diffusion priors (Zhou et al., 2025; Wang et al., 2025). MoDiTac examines it in visual-to-tactile generation, linking preference-controlled global-context injection to output statistics and task-dependent utility.

3 Controllable Visual-to-Tactile Generation

Given a full-scene visual observation x_v, a text condition x_c, and a preference coordinate r, MoDiTac generates a tactile image $\hat{x}_t$ through flow matching. The visual input is encoded through two complementary pathways. A shared VAE produces the visual latent used by the diffusion backbone, while a frozen visual encoder provides spatial visual features for visual-evidence selection and a pooled representation for global context. At each denoising step, the current noisy tactile state updates a set of relevance queries, which then search the spatial visual features for regions relevant to the current tactile generation. The resulting soft relevance map is introduced as a bias on visual keys in joint self-attention. Within MM-DiT-Tactile, text, visual, and tactile representations are jointly modeled after modality-decoupled projections. The updated tactile features further cross-attend to the pooled global visual representation, producing a global-context residual whose contribution is scaled by r. Varying r therefore moves the generation between reconstruction-oriented and perception-oriented operating points. Finally, the prediction head estimates the flow-matching velocity, and the sampled tactile latent is decoded by the shared VAE decoder to obtain $\hat{x}_t$. Figure 2 summarizes the complete generation flow.

3.1 Conditional Flow Matching

Let $z_t = E_{\mathrm{VAE}}(x_t)$ be the clean latent of tactile target x_t, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ Gaussian noise, and $\sigma \in [0, 1]$ the noise coefficient. The linear flow path and target velocity are

$$z_\sigma = (1 - \sigma)z_t + \sigma\epsilon, \qquad v^\star = \epsilon - z_t. \tag{1}$$

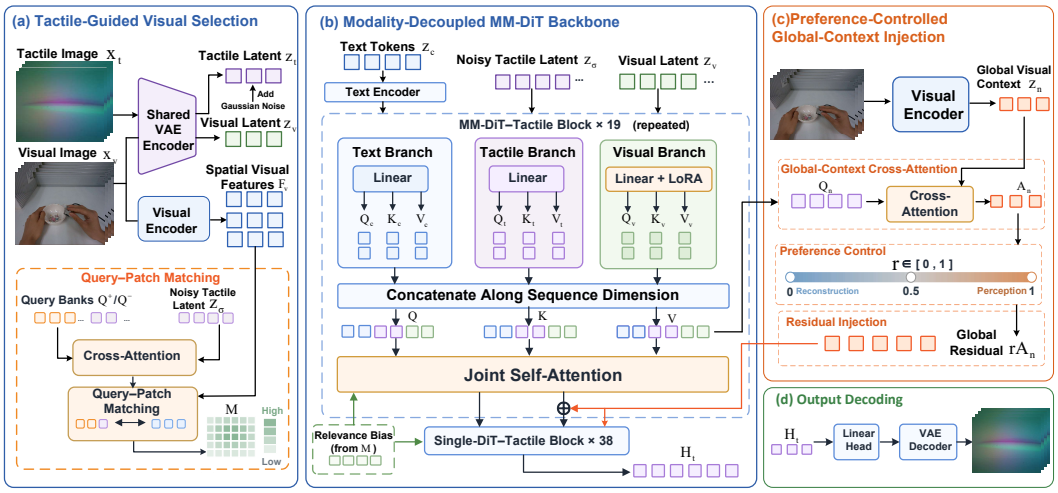


Figure 2: Conditioning and generation flow of MoDiTac. (a) The evolving tactile state produces a soft visual relevance map. (b) The relevance bias and modality-decoupled projections operate in the MM-DiT-Tactile backbone. (c) Global visual context is injected as a residual scaled by r. (d) A linear head and the VAE decoder reconstruct the tactile image from the tactile features. During training, the noisy tactile state is constructed from the tactile target by adding noise; inference starts directly from noise.

The conditional velocity field v_θ takes z_σ, σ, and condition $\mathcal{C}$ as input, where $\mathcal{C}$ includes the text condition x_c, the full-frame visual latent with relevance bias $B(M)$, and global visual context controlled by r. It is optimized by

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{z_t,\epsilon,\sigma}\left[w(\sigma)\,\|v_\theta(z_\sigma,\sigma,\mathcal{C}) - v^\star\|_2^2\right]. \tag{2}$$

During sampling, the generator integrates from the high-noise state toward the tactile-data endpoint and estimates $\widehat{z}_t = z_\sigma - \sigma v_\theta$. Tactile-guided visual selection is updated from the current z_σ, while r remains fixed throughout sampling.

3.2 Tactile-Guided Visual Selection

The visual observation covers a complete scene, whereas the tactile response corresponds to a limited local region. As shown in Fig. 2(a), MoDiTac uses the evolving tactile state to select visual evidence relevant to the current generation. A visual encoder extracts spatial visual features while preserving the visual patch grid:

$$F_v = E_{\text{vis}}(x_v)_{1:} \in \mathbb{R}^{N_v \times d_v}, \qquad N_v = h_v w_v, \tag{3}$$

where (h_v, w_v) denotes the grid dimensions.

MoDiTac uses two learnable query banks: eight relevance queries Q^+ and four irrelevance queries Q^-, with $N_+ = 8$, $N_- = 4$, and $Q^0 = [Q^+; Q^-]$. The relevance queries capture visual patterns that support the current tactile generation, whereas the irrelevance queries model competing patterns that should contribute less. The current noisy tactile latent z_σ first conditions both query banks through cross-attention:

$$\widetilde{Q} = Q^0 + g(\sigma)\,\text{MHA}(Q^0, z_\sigma W_t, z_\sigma W_t), \qquad g(\sigma) = \text{clip}(1-\sigma, 0, 1). \tag{4}$$

As the noise level decreases, the gate $g(\sigma)$ gradually strengthens the conditioning of the queries on the tactile state. With $Q^{(0)} = \widetilde{Q}$, the conditioned queries then retrieve evidence from visual patches through L layers of cross-attention:

$$Q^{(l+1)} = Q^{(l)} + \text{MHA}(Q^{(l)}, F_v W_v, F_v W_v), \quad l = 0, \ldots, L-1, \qquad Q^\star = \text{LN}(Q^{(L)}). \tag{5}$$

For visual location j, the similarity between an updated query and the corresponding visual patch is

$$s_{qj} = \frac{\langle Q_q^\star W_q, (F_v)_j W_p \rangle}{\sqrt{d_q}}. \tag{6}$$

MoDiTac softly aggregates the strongest query responses within each query bank, with normalization by the number of queries. The difference between the relevance and irrelevance responses determines the soft visual relevance map:

$$a_j^{\pm} = \log\left(\frac{1}{N_{\pm}} \sum_{q \in Q^{\pm}} \exp(s_{qj})\right), \qquad M_j = \mathrm{sigmoid}(a_j^{+} - a_j^{-}). \tag{7}$$

Normalization by the number of queries places the two banks on a comparable scale, preventing the larger bank from receiving a higher aggregate response solely because it contains more queries. The resulting map M represents generation-relevant visual support for the current tactile state.

After alignment with the visual latent grid, M is converted into the relevance bias $B(M)$:

$$\mathrm{Attn}(Q, K, V; M) = \mathrm{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}} + B(M)\right) V, \qquad B_{ij}(M) = \begin{cases} \log \widehat{M}_j, & j \in \Omega_v, \\ 0, & j \notin \Omega_v. \end{cases} \tag{8}$$

Here, Ω_v denotes the visual-key indices, and $\widehat{M}$ is the aligned and numerically clipped relevance map. The bias scales the unnormalized attention weight of each visual key by $\widehat{M}_j$, increasing the weight assigned to generation-relevant visual evidence while retaining the full spatial layout.

Because visual selection is learned without spatial supervision, the soft visual relevance map can degenerate into a diffuse or nearly all-one response, while different queries can collapse to similar patterns. A spatial compactness constraint and a query-diversity constraint discourage these two failure modes, respectively; their definitions and hyperparameters are provided in Appendix A.

3.3 Modality-Decoupled Representations in MM-DiT-Tactile

Visual and tactile representations differ substantially in spatial extent and information granularity: full-frame visual conditions contain scene-level context and global appearance, whereas tactile states primarily represent local deformation and fine-grained texture. To reduce the pre-fusion representation conflict, MM-DiT-Tactile uses modality-decoupled projection parameterization for vision and touch within a unified backbone, as shown in Figure 2(b).

A shared VAE encodes the visual observation as $z_v = E_{\mathrm{VAE}}(x_v)$. At noise level σ, the tactile latent z_σ and visual latent z_v are concatenated along the height dimension:

$$z_{\mathrm{img}} = \mathrm{Concat}_h(z_\sigma, z_v). \tag{9}$$

The concatenated result is flattened into a unified image-token sequence. Vision and touch use distinct positional coordinates while retaining their respective spatial layouts; implementation details are provided in Appendix A.2. Let H_t and H_v denote the tactile and visual hidden states, respectively. For $X \in \{Q, K, V\}$, MM-DiT-Tactile parameterizes distinct effective projections for the two modalities:

$$X_t = H_t W_X, \qquad X_v = H_v(W_X + \Delta W_X), \qquad \Delta W_X = B_X A_X. \tag{10}$$

Here, W_X is the pretrained projection base inherited by both modalities. Tactile representations use this base mapping, whereas visual representations obtain a distinct effective projection through the low-rank term $\Delta W_X = B_X A_X$. This parameterization reduces forced sharing of the same Q/K/V geometry without duplicating full projection matrices, thereby mitigating interference from global visual representations with local tactile representations.

Text-condition hidden states H_c are independently projected as $X_c = H_c W_X^c$. For each $X \in \{Q, K, V\}$, the three representations are concatenated as $X = [X_c; X_t; X_v]$ and passed to joint self-attention in MM-DiT-Tactile. The relevance bias $B(M)$ obtained in Section 3.2 is applied only to visual keys, incorporating tactile-guided visual selection into the same attention operation. The soft visual relevance map determines which visual locations receive greater attention weight, while modality-decoupled projections control the representation geometry through which vision and touch enter joint attention. They address spatial-support mismatch and pre-fusion representation conflict, respectively, and jointly support tactile generation within the unified MM-DiT-Tactile backbone.

3.4 PREFERENCE-CONTROLLED GLOBAL-CONTEXT INJECTION

Tactile-guided visual selection and modality-decoupled representations address spatial-support mismatch and pre-fusion representation conflict, respectively, but the generator would otherwise remain tied to a fixed operating point along the reconstruction–perception continuum and could not adjust its output properties to downstream task requirements. MoDiTac therefore introduces preference-controlled global-context injection, as shown in Figure 2(c). The pooled global visual representation supplies contextual information about the object and surrounding scene, complementing the local visual evidence emphasized by tactile-guided selection, while the preference coordinate r controls its contribution to the tactile generation state.

The pooled visual representation is projected into a global-context representation Z_n. In the ℓ-th MM-DiT-Tactile block, the tactile hidden state $H_t^{(\ell)}$ acts as an independent query to the global context:

$$A_n^{(\ell)} = \mathrm{Softmax}\left(\frac{Q_n^{(\ell)} K_n^{(\ell)\top}}{\sqrt{d}}\right) V_n^{(\ell)}, \qquad H_t^{(\ell)\prime} = H_t^{(\ell)} + r A_n^{(\ell)}, \tag{11}$$

where

$$Q_n^{(\ell)} = H_t^{(\ell)} W_{Q,n}^{(\ell)}, \qquad K_n^{(\ell)} = Z_n W_{K,n}^{(\ell)}, \qquad V_n^{(\ell)} = Z_n W_{V,n}^{(\ell)}. \tag{12}$$

Each block uses independent projection parameters, allowing the global visual context to interact with tactile representations as they evolve across layers. The preference coordinate $r \in [0, 1]$ scales the context response before residual injection: $r = 0$ disables this global-context residual, $r = 1$ injects the response without preference scaling, and intermediate values continuously adjust its contribution. During training, preference supervision combines an image-space L1 reconstruction objective $\mathcal{L}_D$ with a high-frequency wavelet L1 objective $\mathcal{L}_P$ that biases supervision toward tactile texture and microstructure. Both objectives are defined in Appendix B:

$$\mathcal{L}_{\mathrm{pref}}(r) = (1 - r)\mathcal{L}_D + r\mathcal{L}_P, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{flow}} + 0.1\mathcal{L}_{\mathrm{pref}} + \mathcal{L}_{\mathrm{sel}}. \tag{13}$$

As r increases, the supervision shifts from image-space paired reconstruction toward high-frequency texture preservation. Global-context scaling provides the inference-time control pathway, while preference-dependent supervision establishes a stable correspondence between r and the reconstruction–perception characteristics of the generated tactile signals.

During training, one preference coordinate r is sampled independently for each example and jointly controls global-context injection and preference supervision. At inference, r remains fixed throughout the complete sampling trajectory; changing its value selects different operating points under the same model parameters. Successive MM-DiT-Tactile blocks progressively update the preference-modulated tactile state, and a linear prediction head maps the final tactile representation to the flow-matching velocity prediction. After sampling, the VAE decoder converts the final tactile latent into the tactile image $\hat{x}_t$.

4 EXPERIMENTS

The experiments evaluate MoDiTac along four dimensions: generation quality under different data partitions; the distinct contributions of tactile-guided visual selection and modality-decoupled projections; controllability of reconstruction–perception operating points within a single generator; and the downstream utility of different operating points.

4.1 EXPERIMENTAL SETUP

The evaluation covers TacQuad (Feng et al., 2025), Touch-and-Go (Yang et al., 2022), and ObjectsGrasping (Han et al., 2025). TacQuad uses frame-level and object-level partitions to assess reconstruction on familiar objects and transfer to unseen objects, respectively, whereas Touch-and-Go and ObjectsGrasping are partitioned by interaction trajectory. Comparisons include TextToucher (Tu et al., 2025), GVST (Yang et al., 2023), and VTTac (Song et al., 2026). All methods are retrained on the same training and test partition within each setting while retaining their native conditioning interfaces. Generation quality is measured by MAE, PSNR, SSIM, DISTS, and FID. MoDiTac is trained on two NVIDIA A100 (80GB) GPUs and uses FLUX.1-dev with 25 sampling steps at 512×512 resolution. Appendices C, D, E, and F detail the partitions, baselines, implementation, and metrics.

Table 1: Tactile-generation comparison across four evaluation settings. The two MoDiTac endpoints within each setting share model parameters. The frame-level TacQuad split evaluates reconstruction within the acquisition distribution; the remaining settings hold out complete objects or trajectories for testing.

Dataset / partition	Method	MAE ↓	PSNR ↑	SSIM ↑	DISTS ↓	FID ↓
TacQuad / frame	GVST	15.759	22.656	0.833	0.208	157.06
	TextToucher	9.033	26.633	0.867	0.157	97.66
	VTTac	4.590	**31.362**	0.907	0.174	85.32
	MoDiTac ($r = 0$)	**4.151**	29.651	**0.910**	0.185	90.08
	MoDiTac ($r = 1$)	6.255	27.578	0.893	**0.147**	**51.75**
TacQuad / object	GVST	16.979	21.023	0.831	0.263	256.21
	TextToucher	10.584	23.459	0.857	0.231	223.84
	VTTac	7.883	**30.744**	0.872	0.226	217.77
	MoDiTac ($r = 0$)	**7.708**	28.396	**0.880**	0.229	220.09
	MoDiTac ($r = 1$)	10.221	24.585	0.877	**0.212**	**172.48**
Touch-and-Go / trajectory	GVST	25.458	11.653	0.598	0.286	144.05
	TextToucher	23.966	14.865	0.614	0.235	122.32
	VTTac	20.944	16.427	0.636	0.218	114.73
	MoDiTac ($r = 0$)	**20.705**	**16.439**	**0.640**	0.229	117.89
	MoDiTac ($r = 1$)	22.539	15.881	0.623	**0.204**	**95.08**
ObjectsGrasping / trajectory	GVST	38.248	9.752	0.378	0.270	188.47
	TextToucher	34.398	12.532	0.405	0.226	146.83
	VTTac	31.445	**15.877**	0.430	0.208	134.23
	MoDiTac ($r = 0$)	**30.943**	15.423	**0.435**	0.215	157.69
	MoDiTac ($r = 1$)	31.622	14.070	0.412	**0.192**	**97.43**

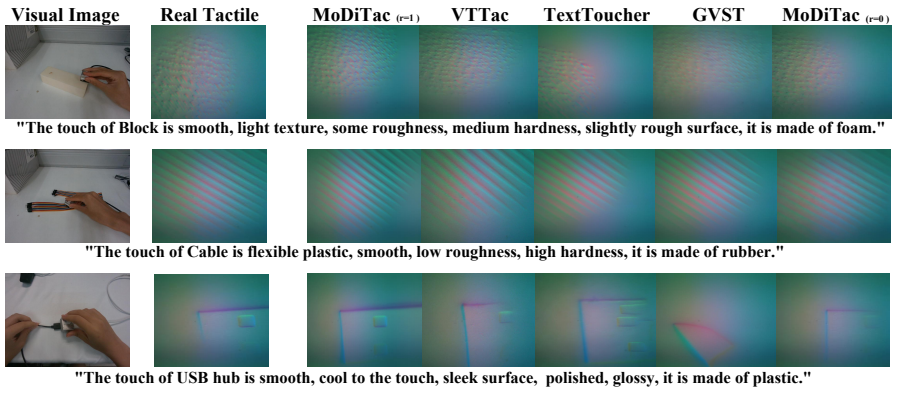


Figure 3: Qualitative comparison with baseline generators. MoDiTac preserves regular texture and contact geometry, including cable ridges and USB hub outlines.

4.2 Generation Quality across Objects and Interactions

Table 1 compares generators across four settings, reporting the two endpoints of a single MoDiTac generator with shared parameters. Table 1 shows that $r = 0$ achieves the lowest MAE and highest SSIM in all four settings, whereas $r = 1$ achieves the lowest DISTS and FID throughout. Relative to VTTac, $r = 1$ reduces FID from 85.32 to 51.75 on frame-level TacQuad and from 134.23 to 97.43 on ObjectsGrasping. Object-level TacQuad and Touch-and-Go show the same trend. These results show that a single generator provides two distinct operating points: $r = 0$ favors paired reconstruction fidelity, whereas $r = 1$ favors perceptual similarity and distributional fidelity.

Figure 3 illustrates these advantages: MoDiTac more faithfully reproduces the cable's regular ridges and the USB hub's rectangular outline and raised feature, preserving texture and contact shape.

Table 2: Ablations on the frame-level TacQuad split. The final row is the same-backbone MM-DiT baseline with trainable shared Q/K/V projections; backbone and pretrained initialization are matched to MoDiTac.

Variant	MAE ↓	PSNR ↑	SSIM ↑	DISTS ↓	FID ↓
Full MoDiTac	6.255	27.578	0.893	0.147	51.75
w/o tactile-guided selection ($M = 1$)	6.833	26.927	0.882	0.167	61.46
Shared modality projections	7.215	26.432	0.873	0.169	68.26
w/o global context pathway	4.151	29.651	0.910	0.185	90.08
Same-backbone MM-DiT baseline	7.644	25.983	0.861	0.186	84.34

4.3 Tactile-Guided Visual Selection and Modality-Decoupled Representation

Ablations use the frame-level TacQuad split (Table 2). Full MoDiTac, w/o tactile-guided selection, and Shared modality projections are evaluated at $r = 1$; here, $M = 1$ replaces the learned selection with a uniform relevance map. Because the global-context pathway carries preference injection, w/o global context pathway corresponds to the $r = 0$ operating point. The final row is the same-backbone MM-DiT baseline: removing tactile-guided visual selection, modality-decoupled projections, and global-context injection recovers standard shared-projection MM-DiT, with the original attention's shared Q/K/V projections trained. It uses the same FLUX.1-dev backbone, pretrained initialization, TacQuad frame-level split, training steps, and optimization budget as MoDiTac, with no additional baseline-specific modules.

Why does tactile-guided selection help? In the retrained ablation, replacing tactile-guided visual selection with the uniform map $M = 1$ increases MAE from 6.255 to 6.833 and FID from 51.75 to 61.46 (Table 2). Complementarily, an inference-time spatial intervention fixes the full model and its generation conditions: permuting the predicted map preserves its values but changes their spatial locations, increasing MAE to 7.084 and FID to 67.90; setting the map to all ones yields 6.741 and 62.23, respectively (Table 6). The advantage therefore arises not only from relevance modulation itself but also from assigning relevance to appropriate visual positions. Figure 4 shows support concentrated on the manipulated object and nearby interaction context across denoising stages.

Why decouple projections before joint attention? Visual appearance and tactile deformation play different roles in conditional generation. Sharing their projections increases MAE from 6.255 to 7.215 and FID from 51.75 to 68.26, with corresponding degradation in PSNR, SSIM, and DISTS, directly validating the generation utility of modality-decoupled projections. Linear centered kernel alignment (CKA) further characterizes the projected visual and tactile Q/K/V representations. Compared with shared projections, modality-decoupled projections produce a clearer separation between the two modalities, indicating more differentiated representation geometries before joint attention. The complete CKA analysis is provided in Appendix H, while the same-backbone ablation in Table 2 further shows the effect of this design on generation performance.

The same-backbone MM-DiT baseline yields an FID of 84.34, compared with 51.75 for MoDiTac. With the backbone, pretrained initialization, and training budget matched, MoDiTac uses slightly fewer trainable parameters (Appendix D.1), indicating that the gains arise from the proposed conditioning design rather than a larger trainable parameter count.

4.4 Task-Dependent Operating Points along the Reconstruction–Perception Continuum

On frame-level TacQuad, r varies over $\{0, 0.25, 0.5, 0.75, 1\}$ under the same model parameters. Larger r shifts generation toward the perception-oriented region: DISTS decreases from 0.185 to 0.147 and FID from 90.08 to 51.75. Smaller r favors paired reconstruction fidelity, with the lowest MAE and highest PSNR and SSIM at $r = 0$. Intermediate values of r produce a continuous but nonlinear transition between reconstruction- and perception-oriented characteristics; Appendix I reports the complete results. This controllability arises from two complementary roles. Global-context scaling provides the inference-time control pathway, while preference-dependent supervision establishes a stable correspondence between r and the resulting reconstruction–perception characteristics;

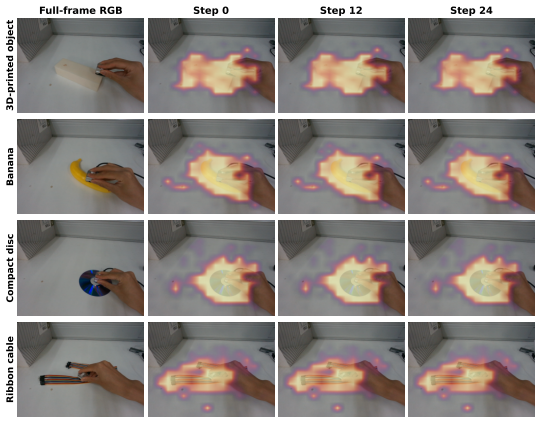


Figure 4: Soft visual relevance maps at early, middle, and late stages of flow matching. High-response regions are mainly concentrated on the manipulated object and its local interaction context.

Table 3: Downstream comparison. Material recognition uses TacQuad; Press–Sliding uses ObjectsGrasping. Source denotes augmentation data and frozen-classifier test input, respectively.

Task	Source	Acc. ↑	F1 ↑
Material recognition	Real only	83.47	83.27
	GVST	86.55	86.34
	TextToucher	87.99	87.86
	VTTac	88.49	88.36
	MoDiTac ($r = 0$)	**90.03**	**89.94**
	MoDiTac ($r = 1$)	88.52	88.35
Press–Sliding	Real tactile	90.00	90.00
	GVST	49.03	48.83
	TextToucher	52.32	51.78
	VTTac	55.26	54.64
	MoDiTac ($r = 0$)	50.72	49.86
	MoDiTac ($r = 1$)	**70.84**	**70.17**

Appendix I.2 analyzes their contributions. Frequency-domain analysis further shows that lower-r outputs better preserve the low-frequency structure of paired tactile observations, whereas increasing r introduces more high-frequency content and improves agreement with real high-frequency spectral statistics (Appendix J).

These systematic changes in generated tactile characteristics motivate examining whether different downstream tasks prefer different regions of the continuum. Material recognition on frame-level TacQuad evaluates the utility of generated tactile data for preserving material-discriminative cues and synthetic-data augmentation, whereas Press–Sliding recognition on trajectory-disjoint ObjectsGrasping evaluates the preservation of action-state cues. Detailed protocols and operating-point analyses are provided in Appendix K.

The two tasks exhibit distinct preferences. Material recognition reaches its highest accuracy of 90.03% at $r = 0$, exceeding VTTac by 1.54 points, whereas Press–Sliding accuracy increases from 50.72% to 70.84% as r increases to 1. At the task-preferred operating points, MoDiTac also achieves 70.17% Macro-F1 on Press–Sliding, compared with VTTac's 54.64% Macro-F1 (Table 3). These results indicate that different downstream tasks benefit from different reconstruction–perception characteristics of the generated tactile data. Because the two tasks above use different datasets, Appendix K.4 further compares fruit identity and Press–Sliding recognition on the same ObjectsGrasping setting. With the generator, trajectory split, preference coordinates, and test samples fixed, fruit identity favors lower r, whereas Press–Sliding improves toward higher r. The contrasting trends under the same data setting further support that the preferred generation characteristics depend on the downstream task.

5 Conclusion

Visual-to-tactile generation does not admit a single output that is uniformly optimal across downstream tasks. MoDiTac turns the reconstruction–perception preference into a continuously controllable generation dimension, allowing a single generator to provide tactile data with different statistical properties according to task requirements. Experiments across three datasets show that the preference coordinate consistently shifts the reconstruction and perceptual characteristics of generated tactile signals, while material recognition and Press–Sliding recognition favor different regions of this continuum; a same-dataset controlled comparison further supports that this difference arises from task requirements. MoDiTac therefore shifts visual-to-tactile generation from searching for a single optimal solution toward learning a task-selectable generation space.

Reproducibility Statement

The model architecture, training objectives, and implementation details are provided in Appendices A, B, and E. Dataset partitions and evaluation protocols are described in Appendices C and K. Additional ablations and analyses are reported in Appendices D and G–J. The code is included in the Supplementary Material.

Ethics Statement

This work uses publicly available datasets or data collected in accordance with applicable requirements. It does not involve human-subject experiments or personally sensitive information. Generated tactile data may nevertheless be mistaken for physical measurements or used without sufficient validation in safety-critical applications. Users should preserve synthetic-data provenance, assess dataset biases, and validate generated outputs against physical measurements before deployment.

AI Use Statement

Generative AI tools were used to assist with literature search, translation, and language editing of the manuscript. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

References

Theo Adrai, Guy Ohayon, Michael Elad, and Tomer Michaeli. Deep optimal transport: A practical algorithm for photo-realistic image restoration. In *37th Advances in Neural Information Processing Systems (NEURIPS)*, volume 36, pp. 61777–61791, 2023.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 6228–6237, 2018.

Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *2018 IEEE Conference on European Conference on Computer Vision (ECCV) workshops*, 2018.

Shaoyu Cai, Kening Zhu, Yuki Ban, and Takuji Narumi. Visual-tactile cross-modal data generation using residue-fusion gan with feature-matching and perceptual losses. *IEEE Robotics and Automation Letters*, 6(4):7525–7532, 2021.

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

Ning Cheng, Jinan Xu, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, and Wenjuan Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *Information Fusion*, 124:103305, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 26529–26539, 2024.

Yu Fang, Xuehe Zhang, Wenqiang Xu, Gangfeng Liu, and Jie Zhao. Bidirectional visual-tactile cross-modal generation using latent feature space flow model. *Neural Networks*, 172:106088, 2024.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In *2025 International Conference on Learning Representations(ICLR)*, volume 2025, pp. 31265–31285, 2025.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *2024 International Conference on Machine Learning (ICML)*, volume 235, pp. 14080 – 14101, 2024.

Ruihan Gao, Kangle Deng, Gengshan Yang, Wenzhen Yuan, and Jun-Yan Zhu. Tactile dreamfusion: Exploiting tactile sensing for 3d generation. In *38th Advances in Neural Information Processing Systems (NEURIPS)*, volume 37, pp. 29839–29863, 2024.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *28th Advances in Neural Information Processing Systems (NEURIPS)*, volume 27, 2014.

Yunhai Han, Kelin Yu, Rahul Batra, Nathan Boyd, Chaitanya Mehta, Tuo Zhao, Yu She, Seth Hutchinson, and Ye Zhao. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. *IEEE/ASME Transactions on Mechatronics*, 30(1):554–566, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *31st Advances in Neural Information Processing Systems (NEURIPS)*, volume 30, 2017.

Carolina Higuera, Byron Boots, and Mustafa Mukadam. Learning to read braille: Bridging the tactile reality gap with diffusion models. *arXiv preprint arXiv:2304.01182*, 2023.

Justin Kerr, Huang Huang, Albert Wilcox, Ryan I Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Proceedings of Robotics: Science and Systems*, 2023.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 4681–4690, 2017.

Jet-Tsyn Lee, Danushka Bollegala, and Shan Luo. "touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4276–4282, 2019a.

Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8943–8950, 2019b.

Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 10609–10618, 2019.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *2024 IEEE Conference on European Conference on Computer Vision(ECCV)*, pp. 430–448, 2024.

Dongyu Luo, Kelin Yu, Amir-Hossein Shahidzadeh, Cornelia Fermüller, Yiannis Aloimonos, and Ruohan Gao. Controltac: Force-and position-controlled tactile data augmentation with a single reference image. *arXiv preprint arXiv:2505.20498*, 2025.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *34th Advances in Neural Information Processing Systems (NEURIPS)*, volume 33, pp. 11913–11924, 2020.

Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum mse photo-realistic image restoration. *arXiv preprint arXiv:2410.00418*, 2024.

Samanta Rodriguez, Yiming Dou, William van den Bogert, Miquel Oller, Kevin So, Andrew Owens, and Nima Fazeli. Contrastive touch-to-touch pretraining. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5857–5863, 2025.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022.

Rui Song, Yang Xu, Zhangzheng Tu, Hongzhou Wang, Hongchen Tan, and Huchuan Lu. Visual–textual information-driven tactile data generation method. *IEEE Transactions on Image Processing*, 35:6317–6330, 2026.

Antonio Luigi Stefani, Niccolò Bisagno, Nicola Conci, and Francesco De Natale. Splattouch: Explicit 3d representation binding vision and touch. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 118–127, 2025.

Ian H Taylor, Siyuan Dong, and Alberto Rodriguez. Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10781–10787, 2022.

Jiahang Tu, Hao Fu, Fengyu Yang, Hanbin Zhao, Chao Zhang, and Hui Qian. Texttoucher: Fine-grained text-to-touch generation. In *2025 AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 7455–7463, 2025.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):3930–3937, 2022.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *2018 IEEE Conference on European Conference on Computer Vision (ECCV) workshops*, 2018.

Yuhan Wang, Suzhi Bi, Ying-Jun Angela Zhang, and Xiaojun Yuan. Traversing distortion-perception tradeoff using a single score-based generative model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 2377–2386, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8081–8103, 2022.

Fengyu Yang, Jiacheng Zhang, and Andrew Owens. Generating visual scenes from touch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22070–22080, 2023.

Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 586–595, 2018.

Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *2023 Conference on Robot Learning(CoRL)*, pp. 1618–1628, 2023.

Shaohong Zhong, Alessandro Albini, Perla Maiolino, and Ingmar Posner. Tactgen: Tactile sensory data generation via zero-shot sim-to-real transfer. *IEEE Transactions on Robotics*, 41:1316–1328, 2025.

Chuqin Zhou, Guo Lu, Jiangchuan Li, Xiangyu Chen, Zhengxue Cheng, Li Song, and Wenjun Zhang. Controllable distortion-perception tradeoff through latent diffusion for neural image compression. In *2025 AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 10725–10733, 2025.

APPENDIX

A ADDITIONAL ARCHITECTURAL DETAILS

A.1 REGULARIZATION AND IMPLEMENTATION OF TACTILE-GUIDED VISUAL SELECTION

Section 3.2 defines tactile-conditioned queries, visual retrieval, and relevance aggregation. This appendix specifies the two constraints used to learn visual selection without spatial annotations and gives the corresponding training configuration.

Regularization without spatial supervision. Without spatial annotations, the soft visual relevance map can become diffuse or nearly all-one, while different queries can collapse to similar patterns. Let $p_j \in [0,1]^2$ denote the normalized coordinate of visual patch j, and let μ be the relevance-weighted soft center. The compactness term uses the normalized second spatial moment:

$$\mu = \frac{\sum_j M_j p_j}{\sum_j M_j + \varepsilon}, \qquad \mathcal{L}_{\mathrm{cmp}} = \frac{\sum_j M_j \|p_j - \mu\|_2^2}{\sum_j M_j + \varepsilon}. \tag{14}$$

Here, $\varepsilon > 0$ is a numerical-stability constant. Minimizing $\mathcal{L}_{\mathrm{cmp}}$ encourages generation-relevant visual support to remain concentrated rather than disperse across the scene. To discourage query collapse, let $d_{ik} = \sqrt{\|Q_i^\star - Q_k^\star\|_2^2/d_q}$ be the root-mean-square distance between LayerNorm-normalized queries, and let $\mathcal{P} = \{(i,k) : i < k\}$ be the set of query pairs. Then

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{|\mathcal{P}|} \sum_{(i,k)\in\mathcal{P}} \max(0, m - d_{ik}), \qquad \mathcal{L}_{\mathrm{sel}} = \beta_{\mathrm{cmp}}\mathcal{L}_{\mathrm{cmp}} + \beta_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}, \tag{15}$$

where $m = 1$ and $\beta_{\mathrm{cmp}} = \beta_{\mathrm{div}} = 0.01$. The diversity term penalizes query pairs separated by less than the margin m, encouraging the queries to capture complementary visual patterns. The combined term $\mathcal{L}_{\mathrm{sel}}$ is the visual-selection regularizer in the total training objective.

Map dropout. To prevent the backbone from relying exclusively on the learned selection, training replaces M with an all-one map with probability 0.2. In this case, $B(M) = 0$ for the visual keys, preserving a direct full-frame conditioning path alongside the selected visual support.

Dynamic selection along the flow trajectory. Because the tactile state changes along the flow trajectory, the soft visual relevance map is recomputed at every denoising step. The CLIP spatial visual features and full-frame VAE condition remain fixed within a trajectory. During training, the selector receives the noisy tactile state z_σ rather than the clean target latent. The map is aligned with and flattened in visual-token order, numerically clipped before taking its logarithm, and applied only to visual keys through the relevance bias $B(M)$ defined in Sections 3.2 and 3.3.

Configuration. The visual-selection parameters are optimized jointly with the other trainable MoDiTac parameters, while the CLIP visual encoder remains frozen. The spatial visual features F_v are projected to 256 dimensions, and packed tactile tokens are projected from the 64-dimensional FLUX input width into the same query space. Selection uses eight relevance queries, four irrelevance queries, one visual cross-attention layer, and eight attention heads. Tactile sequences longer than 64 tokens are reduced by one-dimensional adaptive average pooling before query attention.

A.2 TOKEN LAYOUT AND LOW-RANK PROJECTION SCALING

Tactile tokens use rotary positional coordinates $\pi_t(u,v) = (u,v)$, whereas visual tokens use $\pi_v(u,v) = (u + \delta_p, v)$ with $\delta_p = 32$. This offset assigns distinct positional indices to the two modalities while preserving their respective spatial layouts.

For each visual Q/K/V projection, let s_X denote the fixed adapter scaling coefficient. Incorporating this coefficient into the up-projection factor, $B_X \leftarrow s_X B_X$, allows the complete low-rank update to be expressed directly as $\Delta W_X = B_X A_X$ in Section 3.3, without an additional multiplier.

B PREFERENCE OBJECTIVES

Training image estimate. The preference objectives compare the paired tactile target with the clean image estimated at the sampled noise level. Following Section 3.1,

$$\widehat{z}_t = z_\sigma - \sigma v_\theta(z_\sigma, \sigma, \mathcal{C}), \qquad \widetilde{x}_t = D_{\mathrm{VAE}}(\widehat{z}_t). \tag{16}$$

where D_{VAE} includes the latent rescaling required by the decoder. Both $\widetilde{x}_t$ and the target x_t are mapped from $[-1, 1]$ to $[0, 1]$ and clipped to that range before computing the losses. Below, y and x denote these processed prediction and target images, each of size $C \times H \times W$.

Distortion objective. The image-space L1 loss measures the mean absolute error over channels and spatial positions:

$$\mathcal{L}_D(y, x) = \frac{1}{CHW} \sum_{c,h,w} |y_{c,h,w} - x_{c,h,w}|. \tag{17}$$

High-frequency texture objective. The high-frequency wavelet loss compares the three detail bands of a single-level orthonormal Haar transform. For each channel and non-overlapping 2×2 block, let a, b, c, d be its top-left, top-right, bottom-left, and bottom-right values. The detail coefficients are

$$W_1(x) = \frac{a - b + c - d}{2}, \qquad W_2(x) = \frac{a + b - c - d}{2}, \qquad W_3(x) = \frac{a - b - c + d}{2}. \tag{18}$$

Each band has size $C \times H' \times W'$, with $H' = H/2$ and $W' = W/2$ for even image dimensions. Odd dimensions are padded by replicating the boundary before decomposition. The loss averages the absolute coefficient differences across all three detail bands:

$$\mathcal{L}_P(y, x) = \frac{1}{3CH'W'} \sum_{k=1}^{3} \sum_{c,h,w} |W_k(y)_{c,h,w} - W_k(x)_{c,h,w}|. \tag{19}$$

This objective emphasizes paired high-frequency tactile texture and microstructure; the low-frequency approximation band is excluded.

Preference weighting. For a batch of B examples, the two per-sample mean losses are combined directly according to the preference coordinate of each example and then averaged:

$$\mathcal{L}_{\mathrm{pref}} = \frac{1}{B} \sum_{i=1}^{B} \left[(1 - r_i)\mathcal{L}_D^{(i)} + r_i \mathcal{L}_P^{(i)} \right]. \tag{20}$$

Lower r_i emphasizes image-space paired reconstruction, whereas increasing r_i shifts supervision toward high-frequency texture preservation. This auxiliary objective enters the total loss in Section 3.4 with weight 0.1.

C DATASET PARTITIONS AND EVALUATION PROTOCOL

TacQuad is a multi-sensor aligned tactile dataset containing 151 registered multimodal samples from 99 objects. Its coarse-grained indoor subset captured with a GelSight sensor is used; this subset primarily contains everyday objects under controlled indoor conditions. Touch-and-Go contains human-collected RGB–GelSight observations of diverse objects and materials in natural indoor and outdoor scenes. ObjectsGrasping contains synchronized external RGB and GelSight sequences collected during robotic interactions with deformable objects, including pinching, sliding, and grasping actions.

The four main settings cover familiar objects, unseen objects, and unseen interaction trajectories. For TacQuad, both the frame-level and object-level partitions use a 9:1 training-to-test ratio; they evaluate in-distribution reconstruction and transfer to unseen objects, respectively. Touch-and-Go and ObjectsGrasping use a 20:1 training-to-test ratio and are partitioned by complete interaction trajectory to evaluate transfer to unseen interactions.

Table 4: Four data-partition settings used in the main experiments.

Dataset	Partition unit	Evaluation scope
TacQuad	frame	Familiar-object observations
TacQuad	physical object	Unseen objects
Touch-and-Go	interaction trajectory	Unseen trajectories
ObjectsGrasping	interaction trajectory	Unseen robotic trials

All methods use the same training and test partitions within each setting. Preference comparisons additionally fix the conditioning inputs, initial noise, and sampler configuration, thereby isolating the effect of r within a fixed generator.

D Baseline Conditioning and Comparison

TextToucher (Tu et al., 2025) is a fine-grained text-to-touch generator built on a diffusion transformer. It represents object-level tactile texture and shape together with sensor-level gel information through dual-granularity text conditioning, and therefore provides a text-only reference. GVST (Yang et al., 2023) is a latent-diffusion framework for bidirectional visual–tactile synthesis; its published image-to-touch direction is instantiated to condition tactile generation on visual-scene representations learned through contrastive visual–tactile pretraining. VTTac (Song et al., 2026) combines visual and multi-granularity textual conditions through cascaded cross-attention and injects an uncontacted tactile background as a low-frequency sensor prior before wavelet-based fusion. Under the evaluation protocol, this background contains no target contact deformation.

All three baselines are retrained for every target dataset and partition using their published architectures and native conditioning interfaces. Training and test examples, target-image preprocessing, output resolution, applicable textual semantics, and metric implementations are held constant within each setting. These comparisons assess complete tactile-generation systems, whereas the same-backbone controls below isolate the architectural choices introduced by MoDiTac.

D.1 Parameter Budget and Same-Backbone Comparison

The same-backbone MM-DiT baseline (Base MM-DiT in Table 5) and MoDiTac use the identical 11.90B-parameter FLUX.1-dev backbone, pretrained initialization, TacQuad frame-level split, training steps, and optimization budget. Base MM-DiT trains the standard shared Q/K/V projections in the original MM-DiT attention, together with the visual input projector common to both systems. No additional baseline-specific modules are introduced. MoDiTac reallocates the trainable parameter budget to modality-decoupled projections, tactile-guided visual selection, and global visual conditioning.

Table 5: Parameter budgets for the same-backbone MM-DiT baseline and MoDiTac. Full-system counts include the backbone and all registered conditioning and encoder parameters, including frozen parameters; trainable counts include parameters optimized during training. B denotes billions.

Method	Backbone	Backbone Params	Full-system Params	Trainable Params
Base MM-DiT	FLUX.1-dev	11.90B	17.50B	1.817B
MoDiTac	FLUX.1-dev	11.90B	19.09B	1.787B

The additional conditioning branches increase the full-system size of MoDiTac, but its trainable parameter count remains slightly below that of the shared-projection MM-DiT baseline (1.787B vs.

1.817B). Together with the matched backbone, initialization, and training budget, this comparison indicates that the observed gains cannot be explained simply by a larger optimized parameter budget. The CKA diagnostic in Appendix H uses the same matched shared-projection control protocol as the generation ablation in Section 4.3.

E Implementation Details

TacQuad uses the selected aligned subset and its corresponding tactile descriptions. Touch-and-Go prompts are constructed from material labels, whereas ObjectsGrasping conditions describe the observed object and interaction without using test tactile measurements. MoDiTac initializes MM-DiT-Tactile from FLUX.1-dev and operates at 512×512 resolution. Training uses two NVIDIA A100 (80GB) GPUs for 50 epochs with BF16 mixed precision. The per-device batch size is 1 and gradient accumulation is 1, giving an effective batch size of 2. AdamW uses a learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}, a constant learning-rate schedule with 500 warmup steps, and a maximum gradient norm of 1. The random seed is 42. For flow matching, a timestep index is sampled uniformly from the pretrained scheduler's discrete training timesteps, and the corresponding noise coefficient σ is used with a constant loss weight $w(\sigma) = 1$. LoRA rank and alpha are both 128, and inference uses 25 flow-matching steps.

The pretrained VAE, two text encoders, CLIP visual encoder, and base FLUX transformer weights remain frozen. Trainable components comprise the attention LoRA adapters, visual-condition K/V projections and global visual projection, and the tactile-guided visual-selection module. The adapters are installed in all 19 double-stream and 38 single-stream transformer blocks. The visual pathway receives the resized full RGB frame, from which the frozen CLIP encoder supplies spatial features and a pooled global representation. These features and the full-frame VAE latent are computed once per example and reused, while the soft visual relevance map is updated from the evolving tactile state. Tactile-guided visual selection uses $N_+ = 8$, $N_- = 4$, 256-dimensional queries, eight attention heads, one visual cross-attention layer, and at most 64 pooled tactile tokens. During training, each endpoint is sampled with probability 0.2; otherwise, r is sampled uniformly. At inference, a selected r directly controls global-context injection and remains fixed throughout the sampling trajectory.

F Metric Interpretation

Following the multi-metric evaluation practice used by VTTac (Song et al., 2026), complementary aspects of tactile-image quality are assessed. Mean absolute error (MAE) averages the absolute pixel-wise difference between a generated image and its paired target; lower values indicate smaller paired distortion. Peak signal-to-noise ratio (PSNR) expresses the ratio between the image intensity range and mean squared reconstruction error on a logarithmic scale, with larger values indicating greater pixel fidelity. Structural similarity (SSIM) compares local luminance, contrast, and structural patterns, and therefore complements purely pixel-wise errors; larger values indicate closer paired structure.

DISTS (Ding et al., 2020) compares structure and texture statistics in learned deep features. It is less tied to exact pixel alignment than MAE, PSNR, and SSIM, and lower values indicate greater perceptual similarity. FID (Heusel et al., 2017) measures the distance between Gaussian approximations to the feature distributions of generated and real tactile images; lower values indicate closer distribution-level statistics. All methods use identical preprocessing and the same reference set within each setting. Because FID depends on the reference distribution and sample count, it is used only for method comparisons within the same dataset and partition.

The paired metrics emphasize correspondence with a particular recorded contact, whereas DISTS and FID can reward plausible texture and distributional statistics even when they do not reproduce that contact pixel by pixel. Since a full-scene observation may admit multiple locally plausible tactile responses, the paired and perceptual evidence should be interpreted jointly rather than collapsed into a single quality criterion.

Additional main-result analysis. Relative to VTTac within each partition, the perception-oriented MoDiTac endpoint reduces FID from 85.32 to 51.75 on frame-level TacQuad and from 217.77 to

172.48 on object-level TacQuad. The corresponding within-partition reductions are 114.73 to 95.08 on Touch-and-Go and 134.23 to 97.43 on ObjectsGrasping. The object-level TacQuad split also exhibits a clear transfer gap in paired reconstruction: compared with the frame-level setting, the reconstruction-oriented endpoint changes from 4.151 to 7.708 in MAE, from 29.651 to 28.396 in PSNR, and from 0.910 to 0.880 in SSIM. These paired-metric changes, rather than a direct comparison of FID across partitions, indicate the greater difficulty of transfer to unseen objects.

G Additional Analysis of Tactile-Guided Visual Selection

All reported statistics of the learned soft visual relevance maps are computed with map dropout disabled.

Spatial support during generation. The early, middle, and late responses in Figure 4 cover the manipulated surface and neighboring interaction context. This behavior reflects the use of object-level visual evidence and surrounding context to condition a local tactile response. The examples show persistent object-level support during sampling.

Spatial intervention. Spatial permutation preserves the predicted map's values and marginal distribution but changes the visual positions to which those values are assigned. This control distinguishes location-specific evidence selection from a nonspecific attention bias. The all-one condition retains the same full-frame information while removing selection. These inference-time interventions complement the retrained ablations in Table 2: the former test whether a trained model depends on the learned spatial allocation, while the latter test whether training can compensate for the missing mechanism.

Table 6: Inference-time interventions on the visual spatial support in MoDiTac. Generator parameters and condition inputs remain fixed.

Condition	MAE ↓	PSNR ↑	SSIM ↑	DISTS ↓	FID ↓
Predicted map	6.255	27.578	0.893	0.147	51.75
Spatially permuted map	7.084	26.658	0.875	0.168	67.90
All-one map	6.741	27.023	0.884	0.158	62.23

Query conditioning. Table 7 separates tactile-state conditioning from the two-query-bank design. The w/o tactile-state query conditioning variant removes conditioning on the evolving tactile state, increasing MAE from 6.255 to 6.553 and FID from 51.75 to 57.33. Relevance-only queries retain tactile conditioning but remove the irrelevance-query bank and its contrastive response, yielding an MAE of 6.464 and FID of 55.68; PSNR, SSIM, and DISTS also degrade under both controls. Thus, tactile-state conditioning and relevance–irrelevance contrast jointly improve visual selection.

Table 7: Ablation of query conditioning for tactile-guided visual selection.

Variant	MAE ↓	PSNR ↑	SSIM ↑	DISTS ↓	FID ↓
Full tactile-guided selection	6.255	27.578	0.893	0.147	51.75
w/o tactile-state query conditioning	6.553	27.244	0.887	0.154	57.33
Relevance-only queries	6.464	27.350	0.889	0.152	55.68

H Visual–Tactile Representation-Geometry Diagnostic

Linear centered kernel alignment (CKA) serves as a representation-geometry diagnostic for the projected visual and tactile Q/K/V representations on TacQuad. The shared-projection control and modality-decoupled model follow the same matched comparison protocol used by the generation ablation, including identical evaluation samples, preference values, sampler settings, and per-example

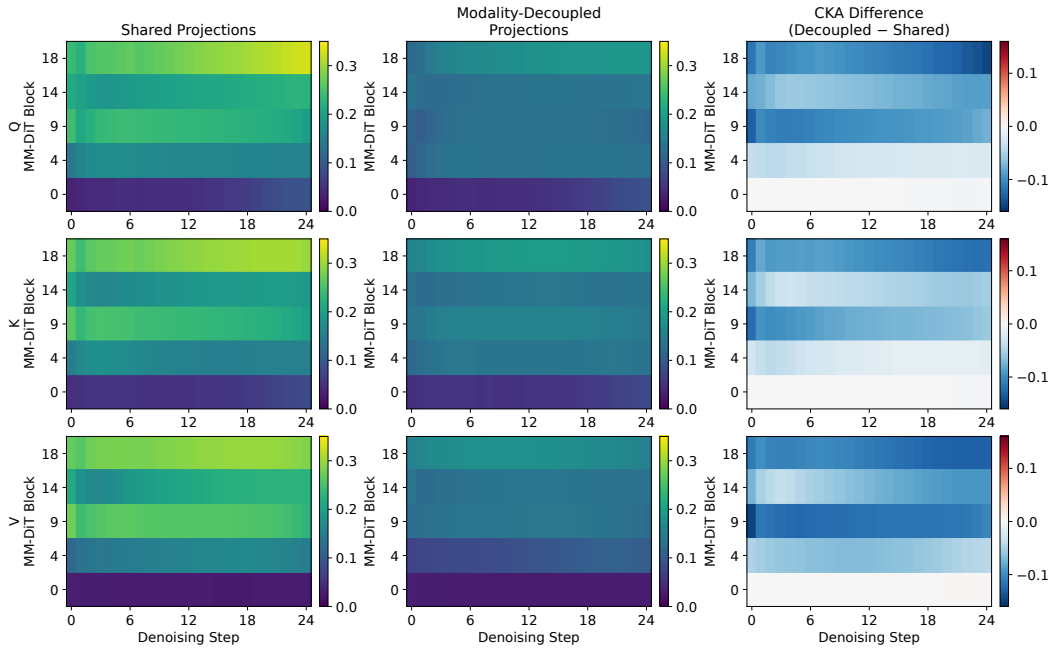


Figure 5: CKA similarity between projected visual and tactile representations. Rows show Q, K, and V; columns compare shared projections, modality-decoupled projections, and their CKA difference (decoupled minus shared). Values are averaged over test objects and preference values. Modality-decoupled projections consistently produce lower cross-modal similarity, indicating more differentiated representation geometries; negative differences denote this reduction.

noise seeds. Q/K/V activations are recorded before QK normalization and rotary position embedding, and before visual and tactile representations interact through joint self-attention. The evaluation covers blocks $\{0, 4, 9, 14, 18\}$ over all 25 denoising steps and the five preference values $r \in \{0, 0.25, 0.5, 0.75, 1\}$. For each activation, up to 256 equally spaced positions are selected from the equal-sized tactile and visual grids, and linear CKA is computed over paired spatial indices. This index correspondence is an analysis convention rather than an assertion of exact physical alignment.

Figure 5 reports CKA across MM-DiT blocks and denoising steps. Mean Q-, K-, and V-CKA decrease from 0.186, 0.183, and 0.185 under shared projections to 0.124, 0.132, and 0.111 under modality-decoupled projections. The differences are most pronounced in the middle and later sampled blocks. Figure 6 shows the same reduction in the per-object averages for Q, K, and V across every sampled object.

These results show that modality-decoupled projections consistently produce more differentiated visual and tactile representation geometries across Q/K/V projections and object instances. CKA characterizes the representation change induced by the projection design, while the corresponding generation gains are evaluated independently by the same-backbone ablations in Section 4.3.

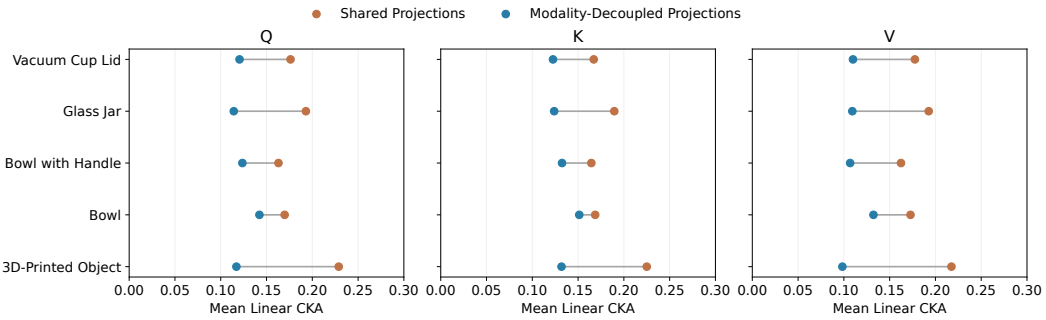


Figure 6: Cross-object consistency of the CKA representation-geometry diagnostic. Connected markers compare shared and modality-decoupled projections for each test object, averaging CKA over preference values, sampled blocks, and denoising steps. Modality-decoupled projections yield lower mean Q-, K-, and V-CKA for every sampled object, showing consistently more differentiated visual and tactile representation geometries across object instances.

I Additional Analysis of Preference-Controlled Operating Points

This section addresses two questions underlying the preference-control results in Section 4.4: whether one checkpoint covers distinct reconstruction–perception operating points as r varies, and how preference-dependent supervision and global-context scaling contribute to this behavior.

I.1 Operating-Point Sweep of a Single Checkpoint

Table 8: Preference-controlled operating points on frame-level TacQuad. Model parameters, inputs, and per-example noise seeds are fixed.

r	MAE ↓	PSNR ↑	SSIM ↑	DISTS ↓	FID ↓
0.00	**4.15**	**29.65**	**0.910**	0.185	90.08
0.25	5.04	28.65	0.907	0.158	64.82
0.50	5.54	28.18	0.899	0.152	59.49
0.75	5.94	27.83	0.894	0.149	55.91
1.00	6.26	27.58	0.893	**0.147**	**51.75**

Table 8 reports the five operating points summarized in the main paper, and Figure 7 visualizes their trend. As the preference coordinate r increases, MAE rises and PSNR falls, while DISTS and FID decrease. The decline in SSIM reflects reduced paired structural consistency. With model parameters, conditional inputs, and per-example noise seeds fixed, the complete preference-control design maps different values of r to operating points with distinct paired-fidelity and perceptual characteristics.

Although the global-context residual in each MM-DiT-Tactile block is linearly scaled by r, the end-to-end metric response is nonlinear. The change from $r = 0$ to $r = 0.25$ accounts for 70.2% of the endpoint reduction in DISTS and 65.9% of the reduction in FID, after which adjacent increments become smaller. Thus, r selects continuous operating points but does not constitute a linear interpolation in metric space.

I.2 Disentangling Preference Supervision and Global-Context Scaling

Three independently trained settings separate the inference-time control pathway from its training supervision. Full retains both preference-dependent supervision and preference-controlled global-context injection. W/o preference supervision removes the preference objective while retaining global-context scaling. W/o preference injection retains preference-dependent supervision but removes r from the generator forward path. All settings share the backbone initialization, TacQuad

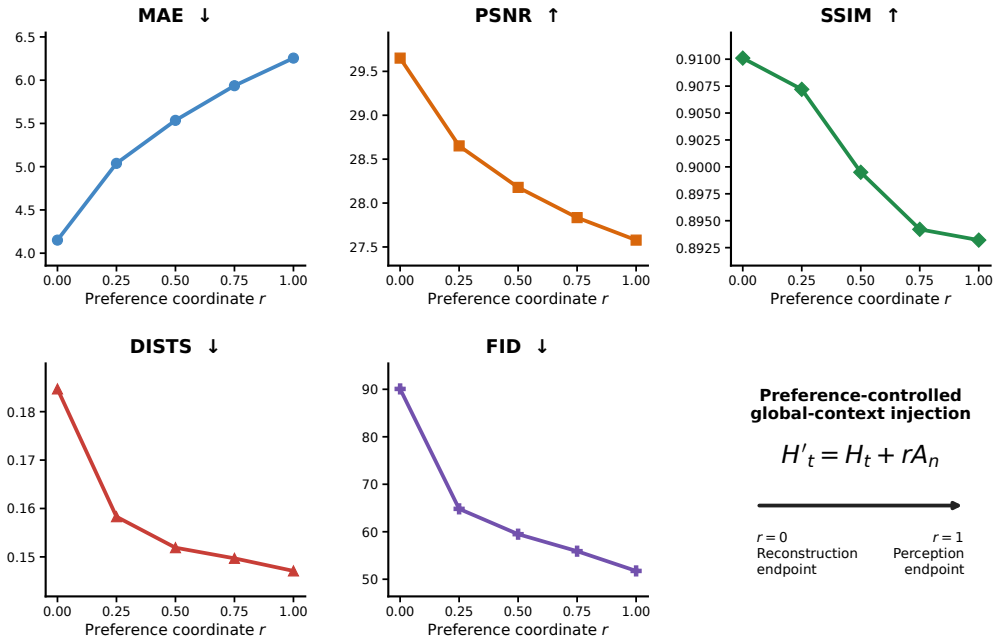


Figure 7: Preference sweep of a single MoDiTac model on TacQuad. Increasing r moves the operating point from lower paired distortion toward lower DISTS and FID. All operating points use the same inputs and per-example noise seeds.

frame-level split, training budget, and evaluation protocol. Full and w/o preference supervision each sweep five inference coordinates using one checkpoint and matched inputs and per-example noise seeds. W/o preference injection samples r only to weight the training objectives; because the generator does not observe r, the sampled preferences are absorbed into a single learned solution.

Table 9: Preference-mechanism controls on the TacQuad frame-level split. Full uses both mechanisms; w/o pref. supervision removes preference supervision, and w/o pref. injection removes preference-dependent injection. Each setting is independently trained with matched backbone initialization, training budget, and evaluation protocol. Each five-point sweep uses one checkpoint; w/o pref. injection is reported once because its forward path does not receive r.

Setting	r	MAE ↓	PSNR ↑	SSIM ↑	DISTS ↓	FID ↓
Full	0.00	4.15	29.65	0.910	0.185	90.08
	0.25	5.04	28.65	0.907	0.158	64.82
	0.50	5.54	28.18	0.899	0.152	59.49
	0.75	5.94	27.83	0.894	0.149	55.91
	1.00	6.26	27.58	0.893	0.147	51.75
w/o pref. supervision	0.00	4.38	29.21	0.905	0.185	88.72
	0.25	4.97	28.72	0.901	0.168	69.88
	0.50	5.31	28.41	0.897	0.173	74.14
	0.75	5.86	27.98	0.893	0.156	61.26
	1.00	6.02	27.81	0.895	0.151	55.06
w/o pref. injection	—	4.24	28.36	0.899	0.186	88.90

Table 9 shows that all five metrics vary monotonically with r under Full: MAE increases, while PSNR, SSIM, DISTS, and FID decrease. W/o preference supervision retains clear endpoint shifts (MAE: 4.38 to 6.02; DISTS: 0.185 to 0.151; FID: 88.72 to 55.06), confirming that global-context scaling provides an effective inference-time control pathway. Its intermediate response is less stable: DISTS and FID reverse at $r = 0.50$, and SSIM reverses over the final interval.

Because w/o preference injection does not expose r to the generator, preference-weighted training is absorbed into one solution rather than producing selectable inference-time operating points. Together, the controls identify complementary roles: global-context scaling provides the inference-time control pathway, while preference-dependent supervision stabilizes the correspondence between r and the opposing trends in paired-fidelity and perceptual metrics.

J Frequency-Domain Characterization of Preference-Controlled Operating Points

Section 4.4 shows that lower preference coordinates emphasize paired reconstruction fidelity, whereas higher coordinates emphasize perceptual realism. This appendix further characterizes the image statistics associated with these operating points in the frequency domain. Two complementary analyses are used: high-frequency ratio and energy quantify how much high-frequency content is present, while band-wise spectral gaps measure agreement with paired real tactile observations.

High-frequency content across operating points. The full-test-set analysis measures both the relative contribution and absolute strength of high-frequency content. Each generated or real tactile image is resized to 256×256, converted to Rec. 601 luminance, mean-centered, and multiplied by a separable 2-D Hann window w. Let $F_r = \mathcal{F}[w \odot (Y_r - \bar{Y}_r)]$ denote the resulting spectrum at preference coordinate r. The discrete Fourier coordinates are $f_x, f_y \in \{-128/256, \ldots, 127/256\}$ cycles per resized pixel, and the radial spatial frequency is $\rho = \sqrt{f_x^2 + f_y^2}$. The high-frequency threshold is fixed to $\tau = 0.25$ cycles per resized pixel, i.e., half the one-axis Nyquist frequency of 0.5 cycles per pixel. The DC bin ($\rho = 0$) is excluded; the high-frequency set contains every 2-D Fourier bin satisfying $\rho \geq 0.25$, up to the corner radius $\sqrt{0.5^2 + 0.5^2}$. The measures are

$$E_H(r) = \frac{\sum_{\rho \geq \tau} |F_r|^2}{\sum w^2}, \qquad E_L(r) = \frac{\sum_{0 < \rho < \tau} |F_r|^2}{\sum w^2}, \qquad h(r) = \frac{E_H(r)}{E_H(r) + E_L(r)}. \tag{21}$$

HFR is $100h(r)$ in percent and measures the relative high-frequency contribution, whereas E_H measures absolute high-frequency energy. Both statistics are computed per image and then aggregated with equal object weighting.

Table 10: Full paired test-set frequency statistics. Values are object-weighted means; the final column reports $10^6 E_H$.

Reference / r	HFR (%)	$10^6 E_H$
Real tactile	0.16876	12.6505
0.00	0.10293	7.4542
0.25	0.11380	8.5140
0.50	0.12243	9.0824
0.75	0.12247	9.9257
1.00	0.13138	10.8358

Table 10 and Figure 8 show that both quantities generally increase with r, although HFR remains nearly unchanged between $r = 0.5$ and $r = 0.75$. From $r = 0$ to $r = 1$, HFR increases by 27.64% and high-frequency energy by 45.37%, demonstrating progressively stronger high-frequency content at higher preference coordinates.

Band-wise agreement with paired real tactile observations. To measure frequency-specific agreement, each generated image and its paired real tactile observation are resized to 512×512, converted to Rec. 601 luminance, mean-centered, multiplied by a 2-D Hann window, and transformed into a power spectrum. Unlike the HFR computation above, this analysis retains only the Nyquist circle,

$$0 < \rho \leq 0.5 \text{ cycles/pixel}, \qquad \rho_{\text{norm}} = \frac{\rho}{0.5} \in (0, 1]. \tag{22}$$

The radial power spectral density is computed on 128 shared radial bins, and each image is normalized by its total AC energy to obtain the spectral-shape distribution $S(\rho)$. For a diagnostic band B,

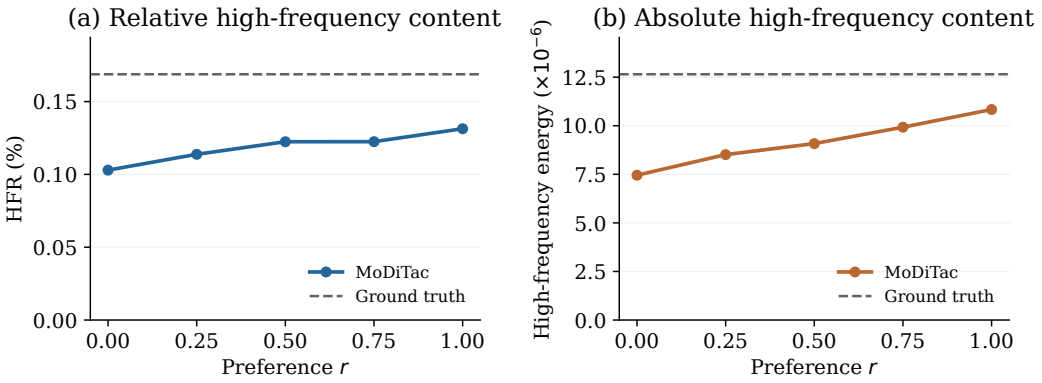


Figure 8: High-frequency statistics across operating points. Markers show equal-object means at the five evaluated operating points, and dashed lines denote real tactile references. Both relative and absolute high-frequency content move toward the reference values as r increases.

the paired log-spectral gap is

$$D_B(r) = \underset{\rho \in B}{\text{mean}} \left| \log(S_{\text{gen},r}(\rho) + \epsilon) - \log(S_{\text{real}}(\rho) + \epsilon) \right|, \qquad \epsilon = 10^{-12}. \tag{23}$$

The normalized bands are Low: $0 < \rho_{\text{norm}} < 0.25$ (0–0.125 cycles/pixel), Mid: $0.25 \leq \rho_{\text{norm}} < 0.50$ (0.125–0.25 cycles/pixel), and High: $0.50 \leq \rho_{\text{norm}} \leq 1.00$ (0.25–0.50 cycles/pixel). Lower gaps indicate closer agreement within the corresponding image-space frequency band; these bands are not calibrated physical material scales. Table 11 reports equal-object means with object-level bootstrap 95% confidence intervals.

Table 11: Band-wise log-spectral gap to paired real tactile observations. Entries are equal-object means with object-level bootstrap 95% confidence intervals; lower is better.

r	Low Gap ↓	Mid Gap ↓	High Gap ↓
0.00	**0.697126** [0.619012, 0.777853]	0.542388 [0.502794, 0.581195]	0.591138 [0.581307, 0.600721]
0.25	0.699129 [0.620914, 0.780153]	0.473944 [0.430018, 0.515128]	0.587663 [0.578024, 0.597709]
0.50	0.705946 [0.616001, 0.799383]	**0.469174** [0.422722, 0.512670]	0.564917 [0.555153, 0.575243]
0.75	0.721414 [0.627851, 0.814751]	0.475407 [0.427329, 0.518812]	0.556494 [0.536497, 0.577296]
1.00	0.880609 [0.784100, 0.979286]	0.482091 [0.434803, 0.524909]	**0.544257** [0.534845, 0.553862]

Table 11 and Figure 9 indicate a frequency-selective trade-off rather than uniform spectral improvement. Low Gap is smallest at $r = 0$ and rises from 0.697126 to 0.880609 by $r = 1$, showing that lower preference coordinates better preserve the low-frequency structure of paired real tactile observations. High Gap decreases across the five evaluated means, from 0.591138 to 0.544257, and is smallest at $r = 1$, indicating progressively closer high-frequency spectral statistics as r increases.

Together, the two analyses separate spectral amount from spectral agreement. The increases in HFR and E_H show that increasing r adds high-frequency content, while the decrease in High Gap indicates that its high-frequency spectral shape becomes closer to paired real tactile statistics. The concurrent increase in Low Gap shows that preference control shifts the output from low-frequency structure preservation toward high-frequency statistical matching rather than improving all frequency bands uniformly.

Qualitative and object-level validation. Figure 10 places generated outputs beside their paired real tactile observations, while Figure 11 compares the two endpoints for 3D print and USB. Under shared spectral color and power scales, both objects exhibit higher mid-to-high-frequency power at $r = 1$ than at $r = 0$. The radial profiles show this increase across a range of frequencies, complementing the aggregate statistics with object-level examples of how r changes the generated texture spectrum.

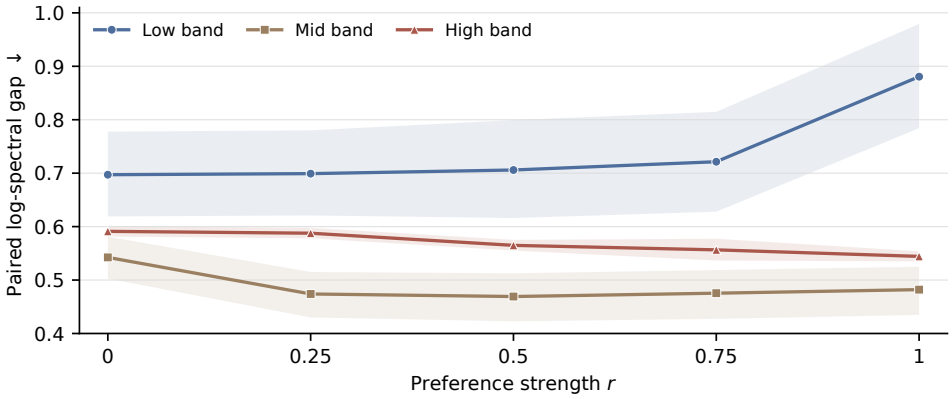


Figure 9: Frequency-selective spectral agreement with paired real tactile observations. Curves show equal-object means, and shaded regions show object-level bootstrap 95% confidence intervals. Low Gap is smallest at $r = 0$, whereas High Gap decreases across the evaluated operating points and is smallest at $r = 1$; lower is better for every gap.

K Downstream Evaluation Protocols and Task-Specific Operating Points

This appendix specifies three downstream evaluations and reports how their utility changes across the five operating points of a single MoDiTac generator.

K.1 Material-Recognition Augmentation Protocol

This section provides the material-classification protocol. The experiment tests whether generated tactile images preserve physically meaningful cues beyond pixel-level similarity. Following standard evaluation practice in tactile representation learning (Feng et al., 2025), generated tactile data are assessed through material classification. This task measures whether synthetic examples retain low-level physical cues such as surface roughness, local deformation, texture regularity, and contact patterns related to friction.

The experiment uses an $A + B$ augmentation protocol on TacQuad to simulate a realistic low-resource setting. The dataset is first divided into training and test sets at a 9:1 ratio. The training portion is then split into a real subset A, containing 30% of the training data, and an augmentation subset B, containing the remaining 70%. Samples are grouped into 13 material classes based on the original object descriptions, and generated samples inherit labels from their paired visual references. The real-data baseline trains the classifier only on real tactile images from subset A. Each augmented model trains on $A + B$, with the samples in B synthesized by the corresponding generator. All models are evaluated only on the same held-out real test set, which remains isolated from generation, augmentation, and classifier training. The downstream classifier follows a linear-evaluation protocol. The pretrained AnyTouch (Feng et al., 2025) base encoder is frozen as the feature extractor, and only a newly initialized linear classification head is optimized with AdamW. All training hyperparameters follow the original AnyTouch evaluation setting. Accuracy measures overall recognition, while Macro-F1 gives equal weight to every class.

K.2 ObjectsGrasping Action-State Recognition Protocol

Press–Sliding action-state recognition on ObjectsGrasping evaluates whether generated touch preserves information that distinguishes interaction modes. The original Grasping and Pinching actions are mapped to Press, while Sliding remains Sliding. The dataset follows the same 20:1 trajectory-disjoint train–test partition used for tactile generation, with all frames from one interaction trajectory assigned to the same partition.

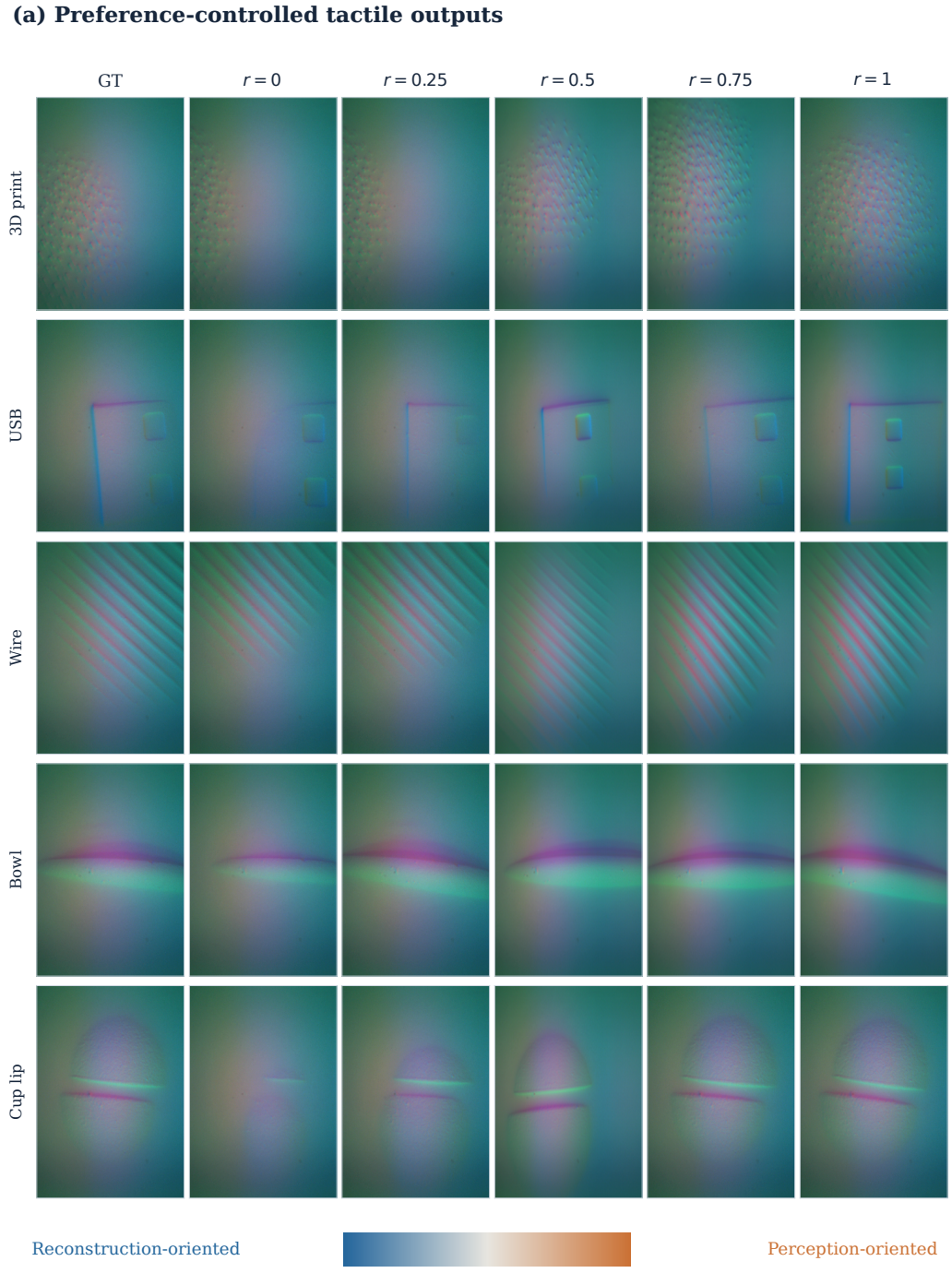


Figure 10: Paired tactile examples from the preference sweep, including real tactile references and outputs at five coordinates. Identical display scaling is used across tactile images.

A lightweight tactile-only TimeSformer is trained exclusively on real tactile training trajectories and then frozen for evaluation. Each input contains five consecutive three-channel GelSight frames resized to 150×150; no visual image, text condition, object identity, or action metadata is provided to the classifier. For generated evaluation, the corresponding consecutive visual frames are processed

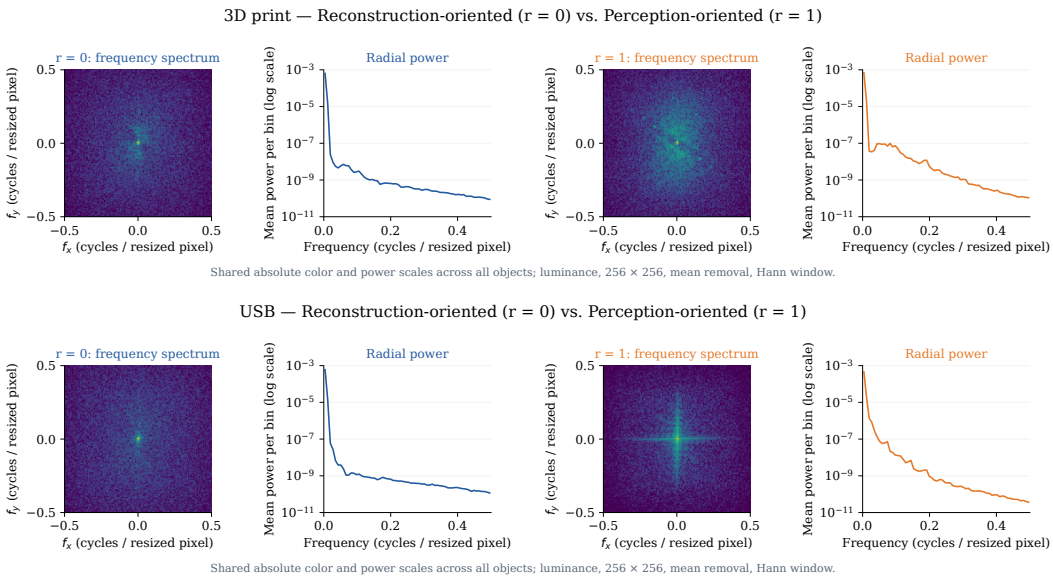


Figure 11: Object-level frequency comparisons for 3D print (top) and USB (bottom). Each row shows the two-dimensional spectrum and radial power profile at $r = 0$ (left pair) and $r = 1$ (right pair). The analyses use luminance images resized to 256×256, mean removal, and a Hann window. Spectral color scales and logarithmic power axes are shared across objects and endpoints; frequency is measured in cycles per resized pixel. Both examples show stronger mid-to-high-frequency power at the perception-oriented endpoint.

independently by the frozen visual-to-tactile generator and restored to their original temporal order before the same five-frame windows are formed. Real and generated tactile inputs use identical preprocessing, held-out trajectories, and window positions. The same frozen classifier is applied to every generator and operating point, and Accuracy and Macro-F1 measure the preservation of action-state discrimination cues.

K.3 Operating-Point Preferences across Downstream Tasks

For $r \in \{0, 0.25, 0.5, 0.75, 1\}$, generation uses the same conditioned examples and per-example noise initialization and changes only the preference coordinate, without retraining the generator. Every operating point follows the material-recognition and Press–Sliding protocols defined above.

Table 12: Task utility across the preference continuum of a single generator.

	Material recognition		Press–Sliding	
r	Accuracy ↑	Macro-F1 ↑	Accuracy ↑	Macro-F1 ↑
0.00	**90.03**	**89.94**	50.72	49.86
0.25	89.76	89.61	55.84	55.13
0.50	89.31	89.16	61.92	61.25
0.75	88.86	88.69	67.48	66.91
1.00	88.52	88.35	**70.84**	**70.17**

For material recognition, each operating point provides a synthetic subset with the same size and label set, and every classifier uses the same real subset, optimization budget, and independent real test set. For Press–Sliding recognition, the real-trained tactile classifier remains frozen and every operating point uses the same held-out trajectories and five-frame windows.

K.4 SAME-DATASET CONTROLLED COMPARISON

Two downstream tasks on ObjectsGrasping are further compared using the same trajectory-disjoint split, trained MoDiTac generator, preference sweep, and held-out trajectories. Fruit identity classification uses a five-class ResNet18 trained only on the real tactile training split and then frozen for evaluation; synthetic tactile data are not used to train this classifier. The classifier is applied to generated tactile images from every operating point, with all values of r using the same test samples. Accuracy and Macro-F1 are reported, while Press–Sliding follows the protocol described in Appendix K.2.

Table 13: Controlled task-specific preference study on ObjectsGrasping. Both tasks use the same dataset, trajectory-disjoint split, trained generator, preference sweep, and held-out trajectories.

	Fruit identity		Press–Sliding	
r	Accuracy ↑	Macro-F1 ↑	Accuracy ↑	Macro-F1 ↑
0.00	**71.15**	**64.52**	50.72	49.86
0.25	69.43	62.67	55.84	55.13
0.50	66.83	58.59	61.92	61.25
0.75	64.33	55.81	67.48	66.91
1.00	62.00	54.00	**70.84**	**70.17**

Table 13 shows that fruit identity Accuracy and Macro-F1 decrease monotonically as r increases, favoring reconstruction-oriented operating points, whereas both Press–Sliding metrics improve monotonically toward the perception-oriented endpoint. These opposite trends under the same dataset, trajectory split, generator, and held-out samples provide controlled evidence that the preferred operating point depends on the downstream task.

The preference sweep exhibits opposing task-dependent preferences. From $r = 0$ to $r = 1$, fruit-identity accuracy decreases by 9.15 percentage points, whereas Press–Sliding accuracy increases by 20.12 points; Macro-F1 decreases by 10.52 points and increases by 20.31 points, respectively. Under the evaluated setting, both metrics show that fruit identity favors the reconstruction-oriented endpoint and Press–Sliding favors the perception-oriented endpoint.

L LIMITATIONS AND BROADER IMPACTS

Limitations. Evaluation covers three datasets and three downstream recognition tasks. Evaluation across additional tactile sensors and tasks would further assess transfer. MoDiTac generates frames independently, so the Press–Sliding experiment evaluates action-state cues preserved by sequences of frame-wise predictions rather than explicitly modeled tactile dynamics. Temporal generation and integration with physical measurements remain directions for future work.

Broader impacts. Synthetic tactile data can supplement costly contact acquisition and reduce the data barrier to tactile-learning research. Visually plausible outputs may nevertheless be mistaken for physically accurate measurements, and dataset biases may propagate to downstream models. Synthetic provenance should therefore be retained, and generated examples should be validated against physical measurements before use in safety-critical applications. Training and sampling costs should also be considered against the reduction in physical data collection.