Language-Guided Multimodal Texture Authoring via Generative Models

Wanli Qian
Department of Computer Science
University of Southern California
Los Angeles, USA
wanliqia@usc.edu

Aiden Chang
Department of Computer Science
University of Southern California
Los Angeles, USA
aidencha@usc.edu

Shihan Lu
Center for Robotics and Biosystems
Northwestern University
Evanston, USA
shihanlu@northwestern.edu

Michael Gu
Department of Computer Science
University of Southern California
Los Angeles, USA
mxgu@usc.edu

Heather Culbertson
Department of Computer Science
University of Southern California
Los Angeles, USA
hculbert@usc.edu

***Abstract*—Authoring realistic haptic textures typically requires low-level parameter tuning and repeated trial-and-error, limiting speed, transparency, and creative reach. We present a language-driven authoring system that turns natural-language prompts into multimodal textures: two coordinated haptic channels—sliding vibrations via force/speed-conditioned autoregressive (AR) models and tapping transients—and a text-prompted visual preview from a diffusion model. A shared, language-aligned latent links modalities so a single prompt yields semantically consistent haptic and visual signals; designers can write goals (e.g., "gritty but cushioned surface," "smooth and hard metal surface") and immediately see and feel the result through a 3D haptic device. To verify that the learned latent encodes perceptually meaningful structure, we conduct an anchor-referenced, attribute-wise evaluation for roughness, slipperiness, and hardness. Participant ratings are projected to the interpretable line between two real-material references, revealing consistent trends—asperity effects in roughness, compliance in hardness, and surface-film influence in slipperiness. A human-subject study further indicates coherent cross-modal experience and low effort for prompt-based iteration. The results show that language can serve as a practical control modality for texture authoring: prompts reliably steer material semantics across haptic and visual channels, enabling a prompt-first, designer-oriented workflow that replaces manual parameter tuning with interpretable, text-guided refinement.**

***Index Terms*—Haptic texture generation, multimodal synthesis, language-conditioned generation, texture perception, variational autoencoder (VAE).**

I. Introduction

Advances in haptic technologies and multimodal generation have increased the demand for methods to deliver realistic and congruent multisensory feedback, particularly centered around haptic signals, through intuitive interfaces [1], [2]. Previous work in virtual haptic textures has relied on the use of data recorded during interactions with real objects [3]. While these methods can create realistic and multimodal interactions, they rely on collected texture samples and therefore cannot easily replicate new, unmodeled textures. Increasingly, users seek systems that use simple commands to seamlessly generate haptic cues that are aligned with other content, a process commonly referred to as *texture authoring*, for use in applications like digital design, virtual prototyping, gaming, and medical simulation and training [4]–[6].

Early authoring approaches focused on recreating real-world textures [7] by searching libraries of pre-recorded physical samples [8], [9]. While effective for faithful reproduction, this *data-bounded* paradigm inherently limits creative exploration to physically available materials, preventing the synthesis of novel or imaginary sensations.

To move beyond direct mimicking and reproducing sensations from real-world textured materials, recent research has advanced toward correlating texture properties with sensory signals. This enables the creation of experiences that are difficult to capture directly or can only be inferred and grounded through other modalities (e.g., semantic meaning, vision, or language) [10]–[12].

Such approaches make it possible to synthesize new virtual sensations by manipulating the underlying texture properties, rather than relying solely on pre-recorded datasets. However, key challenges arise at both ends of this correlation. On the one hand, texture properties are often perceived subjectively [10]. Although measurable descriptors such as roughness, compliance, or friction can be defined, users may interpret them differently depending on context and prior experience [13], making it difficult to ground texture properties with consistent perceptual or semantic mappings. Conversely, accurately capturing texture properties requires action-conditioned, multimodal, and crossmodally redundant sensory signals (e.g., tapping force to convey hardness, vibrations to represent roughness, and visual cues to suggest gloss or microstructure). Incorporating multiple modalities not only increases the richness of the signal representation but also introduces greater

This research was supported by the National Science Foundation under Grant No. 2504241.

data complexity, more demanding collection procedures, and most critically, the challenge of aligning information across modalities.

Early efforts in this direction established the concept of texture authoring by linking the affective space derived from human perceptual ratings with the haptic model space, from which new textures can be rendered by interpolation [14]. With the rise of generative models, researchers began to leverage the traversals in latent spaces of these models to synthesize and tune haptic signal outputs, guided by adjective-based interpolation [14], application needs [15], or user preference [16]. Notably, to maintain the active exploration and interactivity of the haptic experience, several works used parametric haptic texture models as the generation target, rather than the raw haptic signals [14], [16], emphasizing the importance of representations for real-time rendering and user-centered control and expressivity in the authoring process. However, while these approaches allow systems to generate haptic cues that extend beyond what has been physically measured, the grounding of generated textures (e.g., their correspondence to human perception and semantic meaning) has not been systematically established. Additionally, multimodal generation and alignment across haptic and other modalities are still underexplored, with most existing work focusing on a single output modality related to textures [11], [12], [17]. These gaps limit the applicability and generalizability of free-form open-vocabulary texture authoring and constrain the potential for intuitive and crossmodal interactions with virtual textures.

To achieve this, we propose a *multimodal variational autoencoder (VAE)* with a *language-aligned shared latent space* that spans both haptic modalities above: *tapping transients* and *sliding-produced vibrations*. A text encoder (CLIP-style Transformer) maps language prompts into this latent space and is trained via contrastive and conditional objectives to align linguistic semantics with action-conditioned haptic structure [18]. In parallel, a *decoupled diffusion pipeline* generates *texture images* conditioned on the same latent and text prompt, leveraging advances in text-to-image diffusion models. The shared, language-aligned latent thus serves as a hub for text-to-trimodal generation, cross-modal completion (e.g., encode any subset, decode the rest), and counterfactual composition (e.g., "tap like X but slide like Y"). While our system uses a stylus-based interface, this modality is critical for professional domains such as digital industrial design (e.g., prototyping the feel of consumer electronics), virtual surgery (where tool-tissue interaction is fundamental), and digital artistry (enhancing the tactile feedback of digital brushes and sculpting tools).

The key contributions are as follows:

1) **Language-guided, tri-modal generation:** Proposed a VAE that jointly encodes *tapping* and *sliding* when interacting with textured surfaces into a shared, *language-aligned* latent space, enabling text-to-haptics (tap & texture) generation and fine-tuned, diffusion-based text-to-image generation with cross-modal coherence.
2) **Visual-haptic alignment via decoupled diffusion:** Conditioned image diffusion models on the shared, language-aligned latent space to generate texture images whose appearance is perceptually consistent with their rendered haptic responses.
3) **Evaluation of perceptual coherence and language-guided usability:** Conducted two-phase human-subject studies assessing (i) perceptual realism of VAE-generated textures through latent interpolations, and (ii) the usability test of language-based texture authoring in end-to-end interaction.

II. Related Work

A. Multimodal Texture Rendering

Multimodal interaction with virtual environments has recently received increasing attention, particularly within the fields of haptics, virtual reality, and human–robot interaction, and continues to be an active topic of research [19], [20]. Vision shapes expectations, while touch verifies them through actions [21]–[24]. Visual feedback provides rich spatial and material cues such as gloss, roughness, and microstructure that strongly influence users' expectations of touch [25]. In virtual textured surfaces, rendering techniques such as normal mapping [26], reflectance modeling [27], and photorealistic shading [28] are employed to convey fine surface detail and material appearance. Beyond realism, recent work explores how visual cues can be parameterized or learned from multimodal datasets to maintain consistency with haptic feedback [29].

For haptic feedback, the bidirectional nature in interaction determines that a variety of texture properties are best understood through active exploration, either through a tool or through bare fingers. For example, short impacts shape hardness impressions, while sliding excites frequency-dependent vibrations that carry fine roughness [10], [30].

Extensive research has sought to reproduce such transient and vibrational sensations, which arise from impact and sliding, respectively, in virtual environments using standardized tools and setups, with approaches progressing from physics-based model to data-driven techniques [31], [32]. Physics-based models create the signal outputs following the contact dynamics between the tool/finger and the target surfaces to be modeled [33], [34]. Although these models build explicit relations between signal outputs with physical parameters such as surface geometry, material properties, and interaction forces, they often rely on simplified assumptions about contact mechanics and require precise knowledge of boundary conditions. As a result, while they offer interpretability and control, their generalization and rendering precision to complex real-world textures and exploratory behaviors remain limited.

Compared to physics-based methods, data-driven approaches directly map human actions to haptic signals in a black-box manner, relying on statistical models [35], neural networks [36]–[38], and time-series processes [39]. These methods greatly reduce the model tuning and enhance the rendering realism. However, their black-box nature often limits interpretability and physical consistency, for instance, excessive parameter tuning can lead to model collapse and

disrupt causal relationships between input actions and resulting sensory outcomes.

A notable hybrid approach is the event-based force feedback for tapping proposed in [40], which combines both physics-based and data-driven approaches: the tapping transients are modeled as decaying sinusoids inspired by physical observations, with parameters derived from recorded acceleration profiles. Similarly, for sliding-produced vibrations, [3] used autoregressive (AR) processes to capture high-frequency vibrations produced under specific force-speed conditions, with both force-speed selection and AR parameters computed from free-motion recordings of real textured surfaces. These works set the foundation for the haptic models adopted in this work, upon which we build multimodal texture generation from natural language with aligned visual outputs.

While tapping transients and vibrations form the core of our approach, other modalities also contribute to texture perception. Frictional cues modulate the perceived stick–slip dynamics [41]–[43] while auditory feedback correlates with vibrations and impact to reinforce judgments of roughness and stiffness [44], [45]. Thermal signals convey information on thermal conductivity and surface finish quality [46], [47]. Proprioceptive and cutaneous feedback further supports perception of macroscale surface features and overall object geometry [48], [49]. In this work, we focus on tapping transients and sliding-produced vibrations paired with visual previews as the basis for our multimodal texture generation and authoring.

B. Texture Generation and Authoring

Texture generation and authoring initially drew inspiration from methods in image generation and authoring [50], [51]. Over time, these methods have evolved from search-in-the-loop [52], which relied on user-guided retrieval and matching, to exploration within learned latent priors that enable free-form synthesis and semantic control [53].

Linearly interpolating the space derived either from affective descriptions or learned latent priors provides a straightforward way to create novel textures with intermediate perceptual qualities or semantics. [14] created an authoring space by correlating affective attributes of physical textures with changes in haptic signals, allowing users to synthesize new textures that continuously vary along affective (perceptual) dimensions. [54] fine-tuned the vibrotactile signals through linear sampling in the latent space of generative adversarial networks (GAN), showing that latent interpolation enables controlled manipulation of user perception on textures. However, the interpolation process in these methods is often unidimensional and requires subtle adjustments to reach the desired sensations, thereby hindering efficient texture authoring.

To effectively align the rendered haptic sensations with users' perceptual intent, preference-driven authoring leverages user's preference during the interactions with a set of texture candidates to navigate the GAN's latent space via an evolutionary strategy, converging toward the desired feel without tedious manual sampling or interpolation [16]. To ensure interactivity during this process, this work uses the haptic texture models as the generation target, rather than the raw haptic signal. Subsequent research has built upon this idea by integrating human judgment or preference into the haptic authoring process, known as *human-in-the-loop* framework, to produce outputs that better reflect subjective human perception [55]

Design-oriented systems have increasingly linked visual and auditory modalities to expected tactile outcomes [56]. By leveraging cross-sensory correspondences and multimodal priors, these frameworks enable the generation of haptic feedback conditioned on perceptual cues across modalities, spanning generations of image→friction or vibration [57]–[59], audio→vibration [60], and video+audio→friction [61]. Among these works, VAEs, GANs, and Transformer-based architectures dominate the design space, demonstrating strong cross-modal alignment.

Language has emerged as an effective authoring interface across 2D images [62]–[64], 3D scenes [65], [66], and audio [67]–[69]. Early efforts in haptics domain demonstrate the feasibility of text-to-haptic generation [12], [70], [71], but remain limited in authoring scope, typically focusing on a single haptic channel (often vibrations), and performing one-way translation rather than *co-generation*. We address these gaps by learning a unified and physically grounded latent space that *co-generates* both primary haptic channels, tapping transients and sliding-produced vibrations, paired with semantically aligned visual previews, all conditioned on natural language descriptions.

III. Methodology

Our system turns a short text description of a texture into two haptic signals – (i) a set of force/speed–conditioned autoregressive (AR) models to provide virtual texture vibrations during *sliding* and (ii) a bank of event-based *tapping* transients to provide hardness information – and, in parallel, an *independent* text-to-image generation for visual contexts. We first describe the end-to-end inference path (Fig. 1), then how the latent is learned (Fig. 2), followed by data, objectives, and runtime rendering.

A. Haptic Model Components

a) Sliding-produced vibrations: Following the data-driven texture modeling framework proposed in [7], [72], we represent each material using a set of stable, low-order autoregressive (AR) processes indexed by normal forces f and tangential speeds v, which constitute a 2D $f-v$ grid. These force-speed pairs are segmented, and their corresponding AR parameters are computed from data collected during free-motion tool-surface interactions. At runtime, we interpolate the AR coefficients according to the current force and speed within this grid and run the following inference procedure.

$$\begin{aligned} y[n] &= \sum_{k=1}^{p} a_k(f,v)\, y[n-k] + \varepsilon[n], \\ \varepsilon[n] &\sim \mathcal{N}\big(0, \sigma^2(f,v)\big) \end{aligned} \tag{1}$$

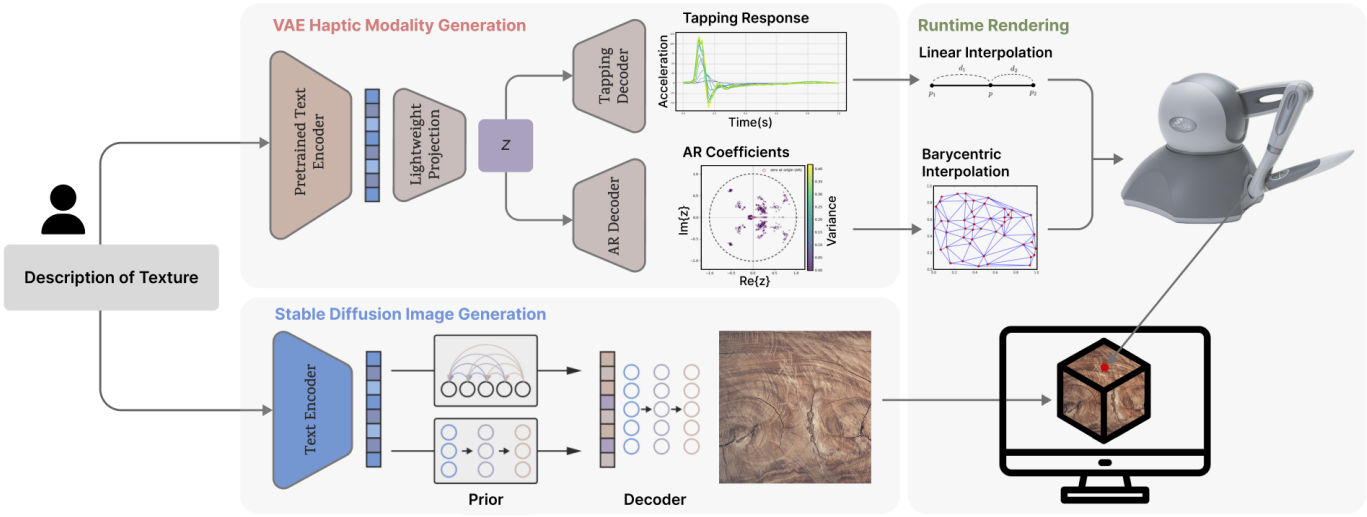

Fig. 1: **System overview.** A text prompt is encoded into a haptic latent z. Modality-specific decoders then output (i) a tap bank for hardness rendering and (ii) an AR matrix for sliding-produced vibrations. The same prompt also drives a decoupled text-to-image diffusion model for the visual preview.

Here, $y[n]$ is the vibration signal at n^{th} time step and $a_k(f, v)$ are the AR coefficients corresponding to the current force f and speed v. $\varepsilon[n]$ is a zero-mean random excitation, of which variance $\sigma^2(f, v)$ controls the overall energy. Intuitively, $a_k(\cdot)$ sets the "timbre" (e.g., smoother vs. scratchier) and $\sigma^2(\cdot)$ sets the "loudness" (how strong the vibration feels). Thus, holding the AR coefficients $a_k(f, v)$ fixed, increasing σ^2 makes the texture feel more energetic or "bursty" (as with gritty abrasives), whereas decreasing it yields a quieter glide (as with polished or compliant surfaces). During inference, we continuously interpolate both a_k and σ^2 from the recorded grid to match the user's current (f, v), so the produced vibrations using AR processes adapt to the user's actions [7], [73].

b) Tapping transients: Hardness is conveyed by short, high-frequency transient forces at impact [40], [73]. The spectral centroid of these transients is roughly constant for a specific material, with high-frequency transients corresponding to stiff materials and vice versa for soft ones. The amplitude of the transients scales linearly with impact speed. For each material, we store 13 recorded acceleration traces covering low to high impact speeds and linearly interpolate between them at runtime to match the measured impact speed.

Sliding and tapping are rendered together so that users perceive both surface micro-geometry and substrate firmness. We discuss the force rendering of these components, including surface stiffness, in Sec. III-F.

B. Data, Representation, and Normalization

a) Corpus: We use haptic models from the Penn Haptic Texture Toolkit (HaTT) [72]: for each of 100 materials, we load (i) an XML with AR processes (stored as *line spectral frequencies* (LSFs)) over a unified force–speed conditions (number of conditions $=$ 18), and (ii) a tapping file with acceleration traces at 13 different impact speed (each T=100 samples at 10 kHz) collected per [73]. We note that the textures in HaTT are **isotropic** and **rigid**: (A) *isotropy*, meaning that the rendered high-frequency vibrations do not depend on in-plane sliding direction; (B) *rigidity*, meaning that the substrate does not undergo large-scale deformation and the feel is dominated by surface asperities and contact dynamics. These assumptions allow us to focus on high-frequency vibration physics rather than large-scale geometric or structural effects.

b) Augmentation: Because the HaTT corpus contains only 100 distinct materials, the resulting feature space is naturally sparse. To enable the model to learn a continuous manifold capable of smooth interpolation, we must augment the models to densify the training distribution. To increase variability of the haptic models while preserving identity of the underlying materials, we implement (a) *tap mixing*: linearly mix each tap set with traces from 19 distinct classes using high target weights (0.95:0.05), so small cross-material perturbations are added without erasing hardness cues; (b) *AR resampling:* Following [16], we first cluster all recorded force–speed samples in HaTT into 18 bins to define a unified $f-v$ grid across textures. For each texture, the AR model is augmented 20 times by sampling one entry from each bin to construct a new instance. Each sampled entry is labeled with the bin's centroid $f-v$ and its parameters are computed via interpolation. This yields 2,000 (AR, tap) training pairs.

c) Tensors and normalization: Each item is stored as an **AR tensor** of size $18\times(21+1)$ (21 LSFs $+$ one variance for each of the 18 force-speed conditions) and a **tap tensor** of size 13×100 (100-datapoint stream for each of the 13 impact speeds). We apply channel-wise z-normalization which keeps each channel on a comparable numerical scale across all materials, preventing high-magnitude features from dominating learning. We utilize the full dataset of 2,000 pairs for training without a held-out zero-shot set. Given the limited class count,

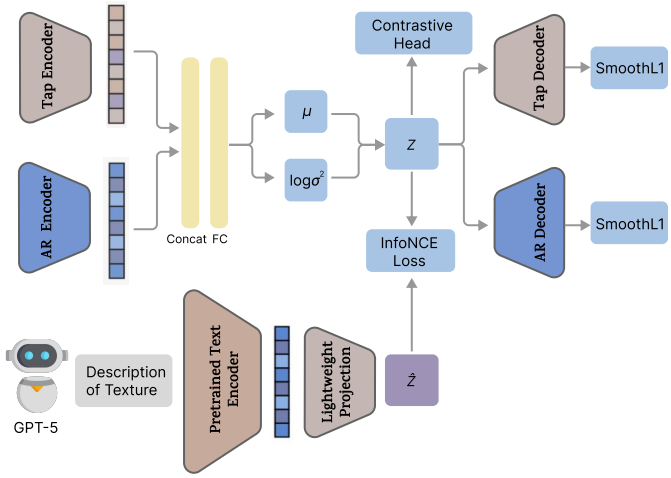

Fig. 2: **Learning latent.** A bimodal VAE reconstructs tap bank and AR matrix from haptic latent z; a contrastive head aligns z with CLIP text, with KL regularization keeping the latent compact and reconstruction losses preserving haptic fidelity.

withholding distinct textures would create large semantic gaps in the latent space, severely impairing the model's ability to extrapolate or interpolate between material categories.

C. System Architecture

Fig. 1 showcases the runtime pipeline. A user provides a short description (e.g., "rough abrasive sheet with gritty particles"), which is encoded into a feature vector by a frozen `CLIP ViT-B/32` encoder. A lightweight projection head transforms this vector into a haptic latent $z \in \mathbb{R}^{64}$. Two decoders then generate:

- an **AR matrix** of size $18\times(21+1)$ (LSFs + variance at each force–speed condition), and
- a **tap bank** of size 13×100 (100 post-contact samples at 13 impact speeds).

These outputs are streamed to the haptic renderer to produce kinesthetic and vibrotactile feedback via a customized haptic device. In parallel, the same text drives a diffusion model to render a visual preview. The visual and haptic branches are *decoupled* at inference; cross-modal coherence arises from the *shared language prompt*, as discussed in the following subsection.

D. Learning Haptic Latent

We aim to learn a unified, *action-aware* latent that co-generates tapping transients and sliding vibrations, and can be *steered by language*. We describe inputs, network, and objectives as follows (Fig. 2).

a) Inputs and text supervision: Each training sample offers an AR tensor and a tap tensor (Sec. III-B). For language conditioning, we generate five short captions per material using `GPT-5` [74], guided by *visual image grounding* and *few-shot expert prompting*. Specifically, the prompt for generating caption uses a template that includes all of the following:

TABLE I: Specification Summary

Latent:	d_z=64 Gaussian $(\mu, \log\sigma^2)$.
AR encoder:	$\mathbb{R}^{18\times22}\to396\to256\to256\to128\to64$ (ReLU).
Tap encoder:	$\mathbb{R}^{13\times100}\to1300\to256\to64$ (ReLU).
Decoders:	AR: $64\to256\to512\to$[3×Res(512, ReLU, Dropout 0.1)] with two heads (LSF 18×21 and variance 18×1). Tap: $64\to256\to512\to$[2×Res(512)] $\to1300$.
Text side:	CLIP ViT-B/32 (frozen) with MLP proj: $512\to256\to64$; latent proj: $64\to128\to64$ (BN, Dropout 0.1), both L2-normalized.
Note:	To guarantee stability, we output LSFs via softmax→cumsum, scale to $(0,\pi)$, and clamp to $\pi-10^{-4}$ before LSF→AR conversion.

(i) the corresponding texture image from the HaTT as visual context, (ii) three expert-crafted examples containing tactile adjectives and multimodal phrasing (iii) explicit guidance to vary linguistic focus across trials—covering visual appearance, tactile sensation, and material composition—to encourage descriptive diversity.

This setup ensures that generated captions are both semantically rich and perceptually relevant, capturing haptic qualities grounded in visual texture appearance rather than generic language priors. During training, one caption is sampled per step, and all captions are embedded using a frozen `CLIP ViT-B/32` text encoder [18].

b) Network: As shown in Fig. 2, a *bimodal VAE* encodes taps and AR tensors into a shared Gaussian posterior $q_\theta(z|x)$ over $z\in\mathbb{R}^{64}$; modality-specific decoders map z back to a tap bank and an AR matrix. Two small projection heads (one from z, one from $\hat{z}$) are L2-normalized so that a contrastive loss can align matched (text, z) pairs. Details are listed in Table I.

c) Objectives: Our training optimizes three terms: (i) **Reconstruction**: a Smooth-L_1 loss on taps and AR tensors to ensure decoded signals match the dataset; (ii) **Latent regularization**: a KL term that keeps z compact and well-behaved; and (iii) **Language alignment**: a InfoNCE with in-batch negatives to align text embeddings with the corresponding material latents. Concretely, with reconstructions $\hat{x}_{\text{tap}}, \hat{x}_{\text{AR}}$, and L2-normalized projections e_x (latent) and e_t (text), we minimize

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{rec}}(\lambda_{\text{tap}}\|x_{\text{tap}}-\hat{x}_{\text{tap}}\|_1+\lambda_{\text{ar}}\|x_{\text{AR}}-\hat{x}_{\text{AR}}\|_1)\\ & +\beta(t)\,D_{\text{KL}}\big[q_\theta(z|x)\,\|\,\mathcal{N}(0,I)\big]\\ & +\lambda_{\text{text}}\,\mathcal{L}_{\text{InfoNCE}}(e_x,e_t;\tau)+\lambda_{\text{align}}\|\hat{\mu}-\mu\|_2^2, \end{aligned}\tag{2}$$

with τ=0.1, batch size 32, λ_{tap}=λ_{ar}=2.0, λ_{text}=0.1, and $\beta(t)$ linearly annealed from 0 to 0.001 over 20–120 epochs [75]. We use Adam optimizer (learning rate 10^{-4}) with gradient clipping at 1.0.

E. Image Generation (decoupled visual preview)

For the preview image, we use an off-the-shelf `Stable Diffusion 2`-base pipeline [76], optionally fine-tuned with PolyHaven textures [77]. The image is not conditioned on the haptic latent z; cross-modal coherence emerges because both haptic modalities and image are driven by the same language description.

F. Runtime Haptic Rendering

We deploy on a 3D Systems Touch device at a 1 kHz servo loop. At each tick, the device state (pose, penetration depth δ, tangential velocity $\mathbf{v}_t$ with speed $v_t = \|\mathbf{v}_t\|$) is measured and used to compute the output force $\mathbf{F}$. The total force applied at the stylus is

$$\mathbf{F} \;=\; F_n\,\hat{\mathbf{n}} \;-\; F_t\,\hat{\mathbf{t}} \;+\; F_{\text{vib}}\,\hat{\mathbf{t}} \tag{3}$$

where $\hat{\mathbf{n}}$ and $\hat{\mathbf{t}}$ are the surface normal and a unit tangential direction of the velocity at the contact point, respectively. F_n is normal contact force, F_t is friction force, and F_{vib} is vibration-induced force. We now define each term.

1) Normal contact (surface stiffness): The normal term models surface stiffness via a virtual spring:

$$F_n \;=\; k_n\,\delta \tag{4}$$

with virtual stiffness $k_n > 0$ and penetration depth $\delta \geq 0$. During impacts, a transient normal impulse is added (Sec. III-F-*Tapping transients*).

2) Sliding vibrations (texture component): When $\delta > 0$, the decoded AR matrix provides force/speed–conditioned contact dynamics. We barycentrically interpolate the AR coefficients and excitation variance at the current contact condition (F_n, v_t):

$$\{a_k(F_n, v_t),\, \sigma^2(F_n, v_t)\}$$

which is used to synthesize the vibration signal $y[n]$ by the AR recursion in Eq. 1. This signal is scaled by the device gain g_{vib} and the maximum continuous output $F_{\max}$ to output vibration-induced force F_{vib}:

$$F_{\text{vib}} \;=\; g_{\text{vib}}\,y[n]\,F_{\max} \tag{5}$$

3) Friction rendering (tangential damping): To render slipperiness, we estimate a friction coefficient from nearby materials in the latent space. With an anchor set $\mathcal{A} = \{(\mathbf{z}_i, \mu_i)\}$ of paired latent z_i and friction coefficient μ_i, the coefficient for a query latent $\mathbf{z}_q$ is

$$\mu(\mathbf{z}_q) = \frac{\sum_{i\in\text{top-}k} \exp(\cos(\mathbf{z}_q, \mathbf{z}_i)/\tau)\,\mu_i}{\sum_{i\in\text{top-}k} \exp(\cos(\mathbf{z}_q, \mathbf{z}_i)/\tau)}, \tag{6}$$

where $\tau > 0$ controls weighting softness. The friction force in Eq. 3 is then

$$F_t \;=\; \mu(\mathbf{z}_q)F_n \tag{7}$$

4) Tapping transients (hardness cue): For impacts, each texture includes 13 precomputed acceleration traces across contact speeds. The impact speed v_{tap} is estimated from normal motion; the transient acceleration $a_{\text{tap}}(t)$ is obtained by linear interpolation between the two nearest traces in v_{tap}. The corresponding normal impulse for effective device mass m_{eff} is

$$F_{\text{tap}}(t) \;=\; m_{\text{eff}}\,a_{\text{tap}}(t) \tag{8}$$

which is added to the normal force during the impact window:

$$F_n \;\leftarrow\; F_n + F_{\text{tap}}(t) \tag{9}$$

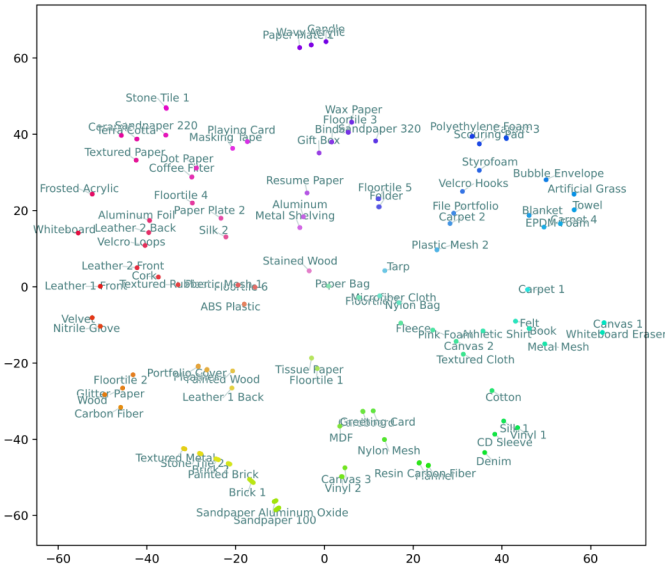

Fig. 3: **VAE latent space embedding.** Each point represents the posterior mean $\boldsymbol{\mu}$ for one of the 100 texture classes after dimensionality reduction (PCA + t-SNE). Nearby points correspond to materials that the model internally considers perceptually similar.

IV. Experimental Results

We assess whether the learned haptic latent (trained on AR matrix + tap bank only) captures perceptually consistent relationships among materials; whether text prompts generate plausible tri-modal outputs (two haptic channels, and one visual); and whether interpolations within the latent space produce smooth and interpretable transitions aligned with human perception.

A. Haptic Latent Assessment

1) Qualitative visualization: For each of the 100 textures in HaTT, we extract the posterior mean vector $\boldsymbol{\mu}$ from the learned latent representation, which compactly encodes each material's vibration and impact dynamics. To visualize these high-dimensional features, we first apply Principal Component Analysis (PCA) to remove minor noise and emphasize dominant variance directions, followed by t-distributed Stochastic Neighbor Embedding (t-SNE) for two-dimensional projection. t-SNE preserves local similarity, such that materials positioned close together share similar internal representations, while distant points indicate strong perceptual differences.

Fig. 3 reveals clear structure in the model's internal representation. For instance, soft foams cluster near other compliant materials, abrasive papers group tightly together, and metals or hard plastics occupy a distinct region—showing that the model organizes textures according to underlying perceptual attributes and physical properties. In other words, proximity in latent space corresponds to similarity in how materials feel through touch.

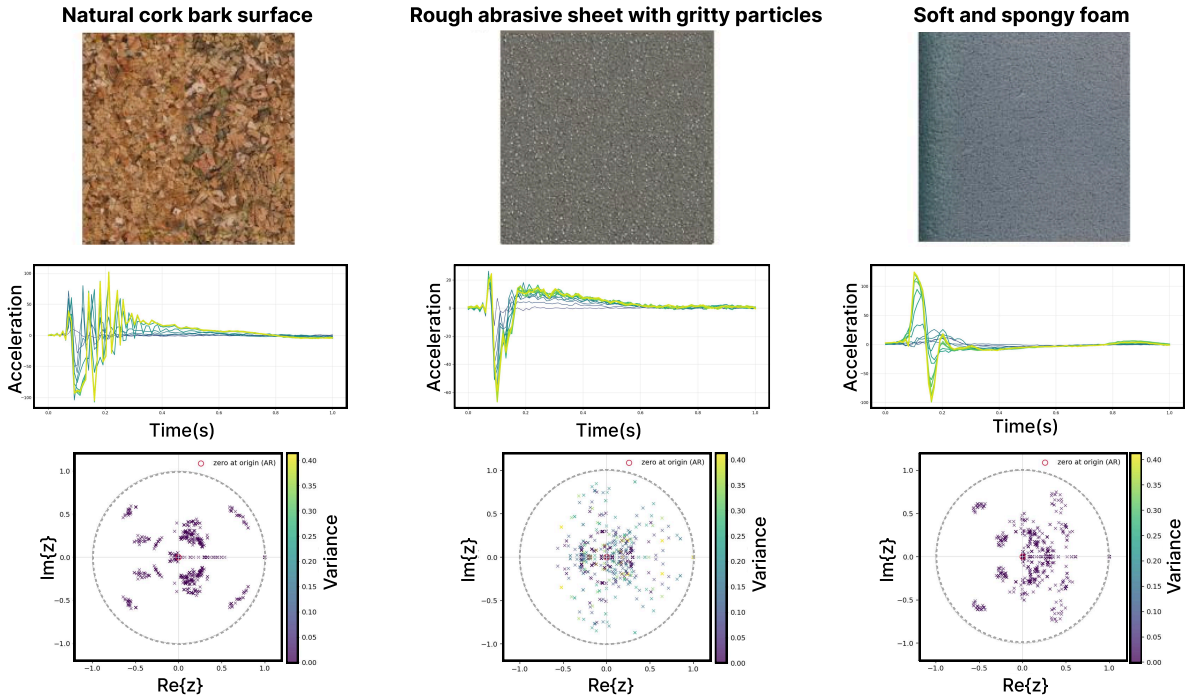

Fig. 4: Tri-modal texture generation from language. **Top**: image generated by a diffusion model. **Middle**: 13 synthesized tapping responses, where color saturation indicates higher impact velocity. **Bottom**: pole distributions of the generated AR models plotted on the unit circle, with marker color denoting excitation variance.

2) Quantitative validation: To confirm that the observed patterns reflect genuine structure rather than visual artifacts, we evaluate cluster compactness and alignment using standard metrics from representation learning [78]–[80]:

- **Cluster compactness.** High internal consistency (*Silhouette* $\approx$ 0.96, *Calinski–Harabasz* $\approx 2.3\times10^4$) indicates that materials grouped together in latent space also share highly similar haptic behaviors.
- **Category separation.** A low *Davies–Bouldin* index ($\approx$ 0.14) confirms minimal overlap between clusters. Materials that feel different remain well separated in representation.
- **Agreement with ground truth.** The model's clusters align strongly with the 100 labeled texture classes (*Adjusted Rand Index* $\approx$ 0.97, *Normalized Mutual Information* $\approx$ 0.99), implying that the latent organization closely parallels human-defined categories.

This structural organization underpins subsequent interpolations and user evaluations. The smooth, semantically ordered latent geometry enables coherent generation paths (Sec. IV-C) and supports human evaluation (Sec. V-A) to probe how users perceive transitions between nearby regions in this space.

B. End-to-End Generation

To validate the full text-to-tri-modal pipeline, we show three representative results generated end-to-end from prompts (*"natural cork bark surface," "rough abrasive sheet coated with gritty particles,"* and *"soft and spongy foam"*) by visualizing the Images, Tapping Transient, AR models in Fig. 4. The generated images are semantically aligned and directionally neutral, consistent with the isotropy assumption in rendering. The haptic signals exhibit material-appropriate signatures:

- **Cork:** Taps show broader, quieter onsets with rapid damping across all speeds. AR poles cluster further from the unit circle with moderate variance, consistent with the compliant, cellular structure of the material.
- **Abrasive sheet:** Taps produce the sharpest onsets and shortest decays, particularly at higher speeds. AR poles spread toward the unit circle with elevated variance, reflecting collision-dominated dynamics at micro-asperity contact points.
- **Foam:** Taps display the softest, most spread-out onsets with gradual decays characteristics. AR poles concentrate deep within the unit circle with minimal variance, consistent with the material's strong energy-absorbing properties.

Across prompts, these patterns yield the expected ordering in spectral sharpness and excitation strength (Abrasive sheet $>$ Cork $>$ Foam) while preserving within-sample speed progressions in the tap stack. The variance coloring in the bottom row distinguishes materials dominated by stochastic micro-collisions (Abrasive sheet) from those exhibiting more damped responses (Cork, Foam).

C. Latent Space Interpolation

We visualize how the decoded signals evolve when interpolating around the average latent between two real textures.

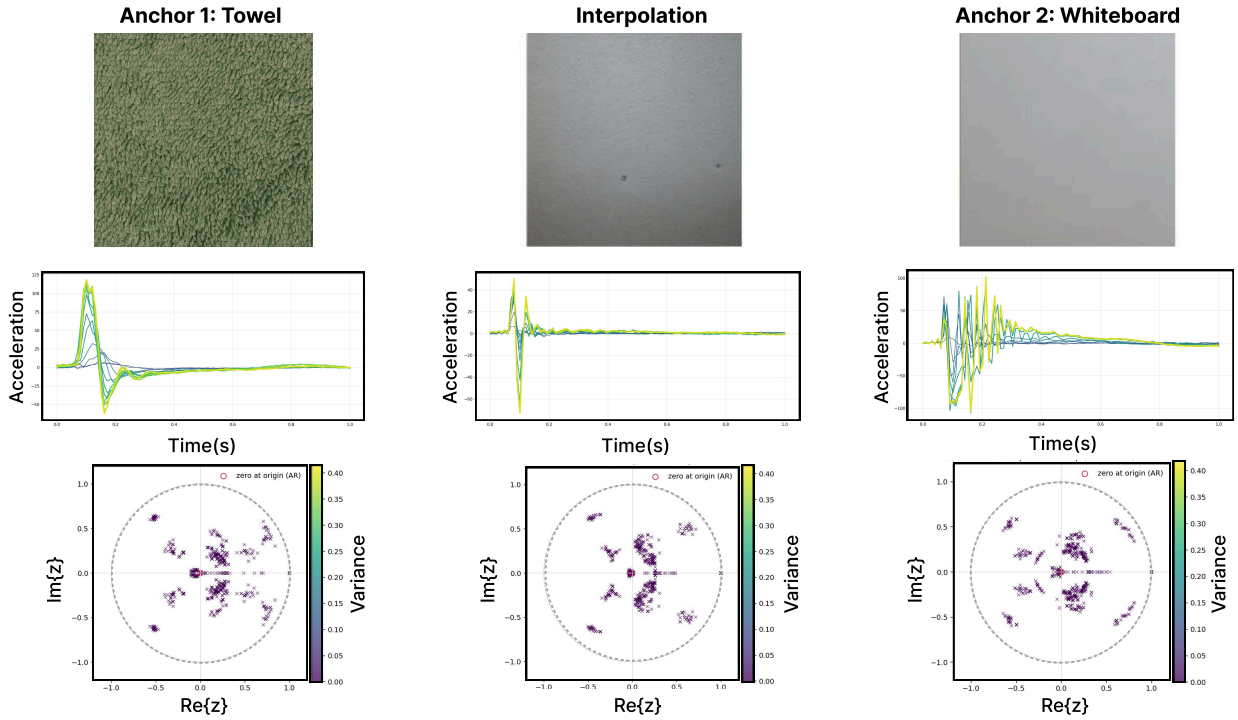

Fig. 5: **Latent interpolation exemplar.** Columns show TOWEL→latent midpoint→WHITEBOARD. *Top*: diffusion image (semantics only, no imposed orientation). *Middle*: synthesized tap responses across 13 impact speeds (higher saturation = higher speed). *Bottom*: AR pole distributions on the unit circle, colored by excitation variance.

Given the posterior means $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ for two anchors, we compute their average

$$\bar{\mathbf{z}} = \frac{1}{2}(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B), \tag{10}$$

and sample nearby latent around $\bar{\mathbf{z}}$ to generate the corresponding haptic reconstructions from the VAE decoder. Visuals are generated by diffusion using the corresponding text prompt (e.g., "texture between *towel* and *whiteboard*").

Fig. 5 presents TOWEL→WHITEBOARD as a representative path; we observe similar behaviors on other pairs. We note that the user study in Sec. V-A analyzes how participants perceptually judge such generated latent interpolations; here we only visualize the signals produced by the model.

From towel to the latent midpoint, the AR spectra evolve in a structured but non-affine manner. Low-band energy associated with fibrous drag gradually recedes, while high-band content linked to polished sliding becomes increasingly prominent toward the whiteboard end. This transition follows a curved trajectory rather than a straight interpolation: mid-frequency regions adjust earlier than the highest bands, consistent with human roughness sensitivity peaking at mid frequencies and with regime mixing in the decoder, where the latent midpoint does not simply average AR coefficients across all force–speed settings.

Tapping transients at the midpoint exhibit a crisper onset and slightly longer ringing than towel, but are far less peaky than whiteboard, reflecting increased contact stiffness with residual compliance.

Visually, the diffusion-generated midpoint softens the towel weave and reduces the whiteboard sheen while maintaining isotropy. This demonstrates that the model preserves coherent, semantically aligned structure across modalities.

Across latent paths, both AR spectra and tapping transients move in material-plausible ways, while the corresponding images remain consistent with the underlying language semantics. However, these observations are signal-level; in the next section (Sec. V-A), we present users' perceptual ratings for these latent interpolations.

V. USER EVALUATION

We conducted a two-phase user study (N=17; 10 male, 6 female, 1 non-binary). Twelve participants reported no prior experience with haptic devices (e.g., 3D Systems Touch, Novint Falcon). The study was approved by the University of Southern California Institutional Review Board under protocol UP-20-01131; all participants gave informed consent and participated voluntarily.

Our goals were to: (i) test whether the learned latent supports *directional, semantically interpretable* changes between reconstructed anchors, and (ii) evaluate whether latent interpolations produce *coherent, novel* textures that inherit recognizable properties from both anchors without collapsing to trivial averages.

A. Anchor-Referenced Perceptual Interpolation

Setup. We formed all $\binom{5}{2}$=10 anchor pairs (A, B) from *Sandpaper*, *Velvet*, *Styrofoam*, *Aluminum Foil*, and *Velcro*

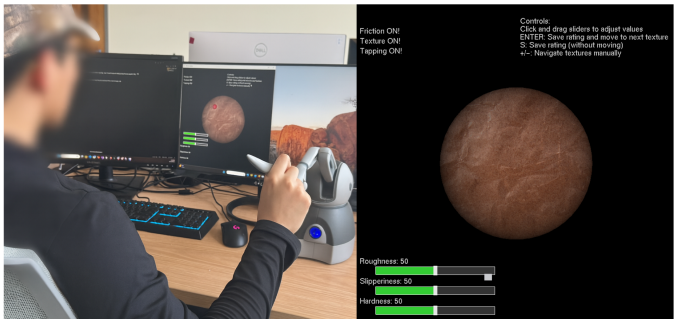

Fig. 6: **User study setup.** *Left:* Participant interacting with the 3D Systems Touch device while viewing the texture and rating UI. *Right:* On-screen interface showing the rendered texture, controls, and three sliders for rating Roughness, Slipperiness, and Hardness.

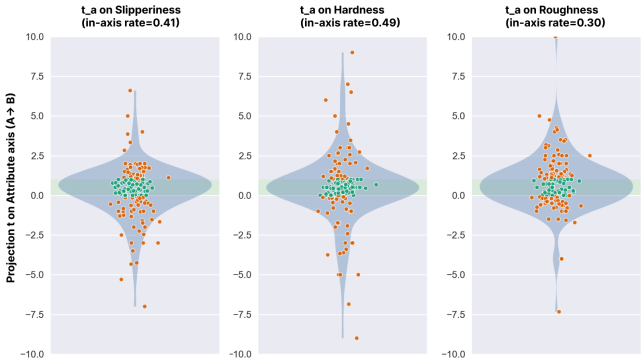

Fig. 7: **Attribute-wise axis projections.** Each dot is a participant–pair instance. Green band denotes "between anchors" ($t_a \in [0,1]$). Inside-axis rates: **Slipperiness** 0.41, **Hardness** 0.49, **Roughness** 0.30.

Hooks (selected from HaTT Corpus for distinct perceptual features). For each pair, we generated the haptic intermediate by decoding the arithmetic mean of the two anchors' encoder posterior means, and an image produced by diffusion model prompted as "texture between *anchor1* and *anchor2*." as described in Sec. IV-C. Participants explored each sample by sliding and tapping with a 3D Systems Touch device while viewing the corresponding texture image. They were asked to use three 0–100 sliders for the ratings of **R**oughness, **S**lipperiness, and **H**ardness for the given texture (Fig. 6). These dimensions were selected as they represent the principal axes of tactile perception of textures through a tool [10], whereas attributes like temperature or macro-geometry were excluded due to hardware and modeling constraints.

We summarize observations with two visualizations:

(1) Attribute-wise projection scatter (Fig. 7)

For each participant, pair, and attribute $a \in \{R, S, H\}$, we compute the anchor–axis projection

$$t_a = \frac{(X_a - A_a)(B_a - A_a)}{(B_a - A_a)^2}, \tag{11}$$

which places the rating X_a on a 1-D line from anchor A_a to B_a: t_a=0 at A, t_a=1 at B, t_a<0 beyond A, and t_a>1 beyond B. Thus, $t_a \in [0,1]$ ("green band" in Fig. 7) means the generated blend is rated between the two anchors along attribute a. A violin plot is fitted for each attribute subplot to show the distribution of the projected ratings.

(2) Per-pair median grid (Fig. 8)

For each pair, we take *participant-wise medians* of the ratings for anchor A, the generated blend X, and anchor B on each attribute and plot them together.

Result: **Hardness (H): strongest "between-anchors" placement.** With an inside-axis rate of 0.49, participants most often judged the generated hardness to sit between the two anchors (Fig. 7, middle). The violin distribution shows the tightest density within the green band, with relatively shallow tails, indicating frequent "between-anchors" placement and smaller excursions outside [0,1]. Pairs containing a compliant anchor (Velvet or Styrofoam) pulled H *below* an arithmetic midpoint, consistent with tap transients conveying onset stiffness while soft backings damp ringing (sub-additive firmness). See *Foil↔Styrofoam* and *Sandpaper↔Velvet* in Fig. 8.

Slipperiness (S): moderate "between-anchors," biased by smooth films. Inside-axis rate is 0.41 (Fig. 7, left). The violin plot shows a broader density with a slight skew toward higher t_a values, corresponding to shifts toward smoother-film anchors (Foil or Velvet). This indicates that participants tended to rate blends as *more slippery* when a smooth surface film was present, largely independent of hardness. This pattern aligns with our runtime friction estimate tied to latent neighbors and expressed as tangential damping (Eq. 7).

Roughness (R): least "between-anchors," rough anchor dominates. Inside-axis rate drops to 0.30 (Fig. 7, right). The violin distribution exhibits the widest spread and heavy tails, consistent with the dominance of asperity cues—blends often inherit the rougher anchor's grain strongly enough to exit the [0,1] interval. Users frequently reported the blend retaining the properties of the rough anchor (Sandpaper or Hooks) even when paired with a smoother surface. It reflects that AR models emphasize high-band energy from micro-collisions. See *Foil↔Hooks* and *Sandpaper↔Velvet* in Fig. 8.

Putting *R/S/H* together. Across pairs in Fig. 8, users perceived the generated textures as *plausible hybrids*: (i) R is shaped by asperity statistics (rough anchor dominance), (ii) H is governed by compliance (soft-backings reduce felt firmness), and (iii) S follows surface films (smoothness over stiffness). These user ratings explain why combined analyses show "between-anchor" trends along the principal direction, yet exhibit bounded deviations when attributes pull differently, indicating coherent but *non-collinear* cue mixing.

B. Phase I: Qualitative Analysis (HXI)

To complement anchor-referenced projections and per-pair median analyzes, we also assess interaction usefulness, engagement, perceived realism, cross-modal coherence, and mismatch, to provide a validity check on the overall experience.

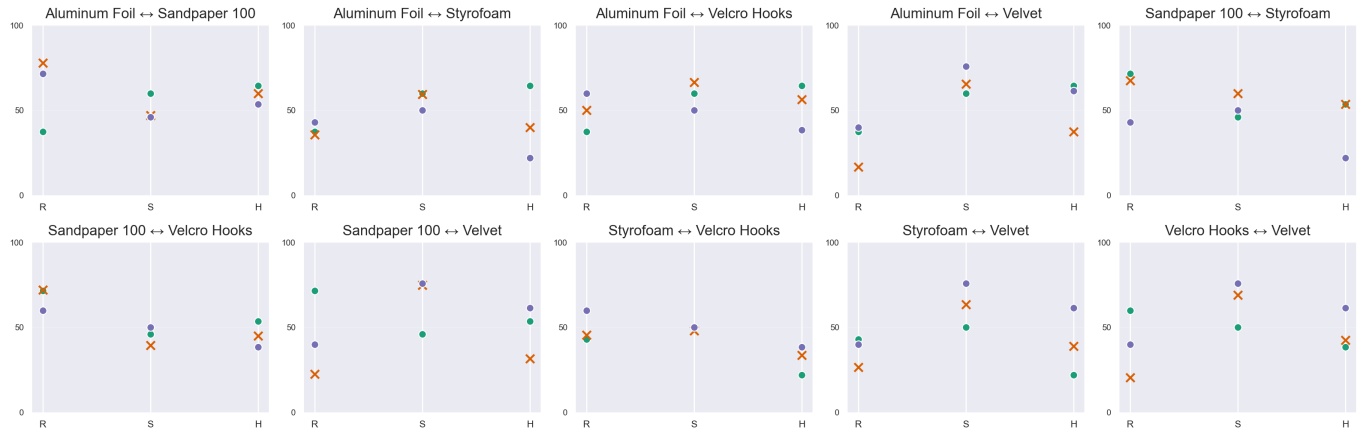

Fig. 8: **Attribute-wise axis projections with distribution.** Violin plots (density) overlaid with per-trial scatter. The green band marks "between anchors" ($t_a \in [0, 1]$).

We grounded our questionnaire in the Haptic Experience Inventory (HXI) factors—**Autotelics**, **Involvement**, **Realism**, **Harmony**, **Discord** (4 items each)—and adapted items to our usecase following [81]. Items used a 7-point Likert scale.

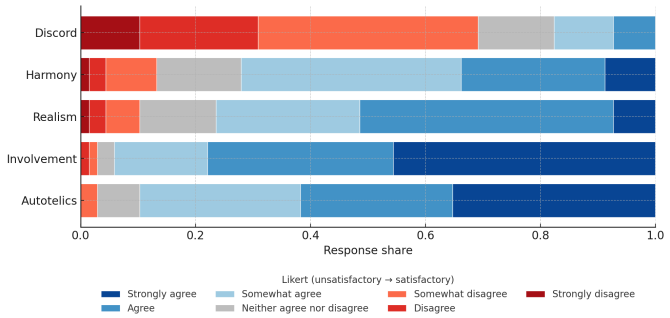

Fig. 9: **HXI user ratings.** 100% stacked distributions per factor (Discord reversed for scoring). Color encodes unsatisfactory→satisfactory from red to blue

HXI user ratings in Fig. 9 show three clear trends. First, **Autotelics** and **Involvement** are strongly right-skewed: participants *enjoyed* the sensations and felt that the haptics *helped them engage*. Second, **Realism** and **Harmony** lean to agreement: users experienced the outputs as *plausible* and *coordinated* with the concurrent image even though, at inference, image and haptics are only coupled via the *shared prompt*. Third, reverse-scored **Discord** stays low (i.e., explicit reports of mismatch are rare).

Beyond HXI, we asked participants whether each interaction mode was useful and which attribute was easiest to judge. We visualize both as 100% stacked bars with the same red→blue ramp.

For ratings of interaction usefulness in Fig. 10, *Sliding* concentrates responses on the satisfactory side, indicating that it gives clear texture information during exploration. *Tapping* is also positive but with greater spread, matching its role in conveying short, high-contrast events (impact and compliance)

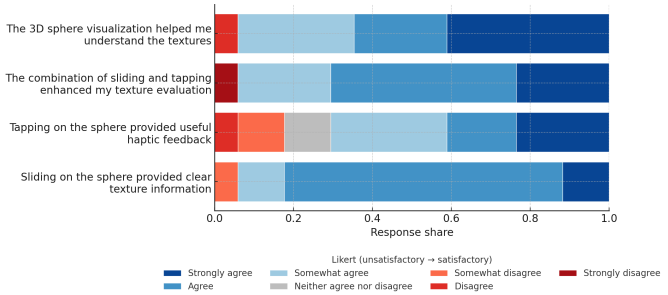

Fig. 10: **Interaction usefulness** for Sliding, Tapping, their combination, and 3D visualization. Distributions concentrate toward the satisfactory (blue) end; the combination shifts furthest right, indicating complementarity rather than redundancy in haptic channels.

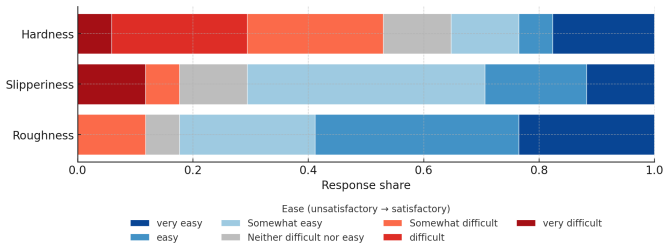

Fig. 11: **Attribute-rating ease.** Roughness is easiest, Hardness next, Slipperiness hardest—consistent with our signal pathways (roughness salient while sliding; hardness from tap transients; slipperiness depends on friction/adhesion cues)

that some users find subtler to interpret. Notably, the *Sliding+Tapping* condition shifts the entire distribution further toward satisfaction, showing the two actions are complementary rather than redundant. The *3D visualization* similarly skews toward satisfactory, supporting our choice to keep a visual reference even though the image is generated independently from the haptic latent.

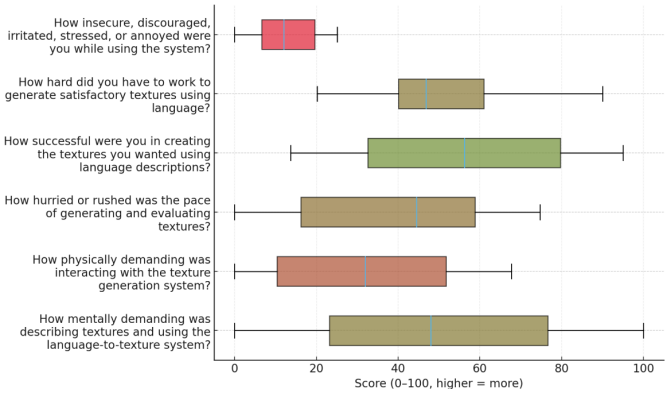

Fig. 12: **NASA–TLX** Mental, physical, temporal demand, effort, frustration, and perceived success. Scores are shown on their native scales; lower is better for demand/effort/frustration, higher is better for success.

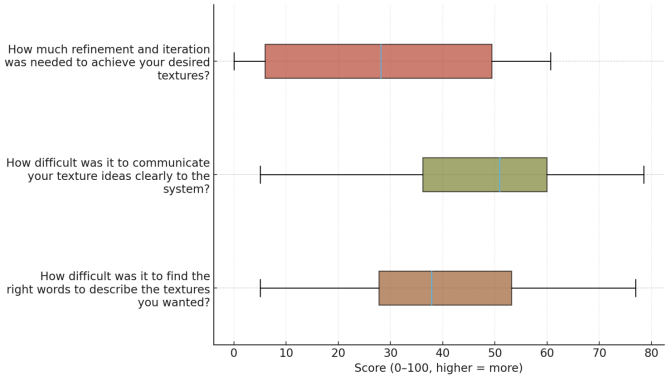

Fig. 13: **Language-specific outcomes.** Agreement that the output matched the description and felt realistic, plus self-rated ability to create and refine using language prompts.

In Fig. 11, **Roughness** is rated the easiest attribute to judge, followed by **Hardness**, while **Slipperiness** is perceived as the most difficult. This ordering aligns with nature of the signals produced by our renderer: roughness is strongly conveyed in the sliding AR band, producing large and easily perceived variations; hardness is primarily carried by tap transients, which are distinct but brief; and slipperiness depends on friction that varies with contact force, speed, and micro-geometry, making it inherently noisier and more difficult to perceive reliably.

C. Phase II: Language as Pragmatic Navigation

Each participant completed three *prompt→texture* trials. In each trial, they entered a short natural-language description of a texture they wanted to create. Participants explored the texture by tapping and sliding using the haptic device while viewing the image. After exploration, they rated their experience using the NASA–TLX [82] workload items and language-specific questions on realism, description-match, and satisfaction.

a) Findings.: As shown in Fig. 12, satisfaction levels are consistently high across NASA–TLX dimensions. Mental and temporal demands convert to mid-to-high satisfaction, while physical effort remains lowest—expected for stylus-based interaction. Frustration scores are low, and NASA–TLX success is notably high, indicating that participants felt confident and effective during texture creation. Language-specific workload shows a similar trend, suggesting that phrasing prompts and making one single refinement felt straightforward and manageable.

Agreement ratings (Fig. 13) also skew strongly toward satisfaction. Participants agreed that generated textures matched their intended descriptions and appeared realistic. Ratings for *creation ability* were positive, while *refinement ability* showed a small neutral portion but remained agreement-dominant overall. Together, these findings indicate that participants could use language as a *pragmatic navigation tool* through the latent manifold; most users reached the texture they envisioned with one prompt and, at most, a light adjustment, all at moderate cognitive cost and minimal physical effort.

Open-ended feedback corroborates these findings. Participants frequently noted that the workflow "got me close" on the first try, and that a single adjective change (e.g., "less gritty," "more waxy") often produced the desired feel. Roughness and hardness were intuitive to control through adjectives such as "coarse," "fine," "soft," or "rigid," while slipperiness required more nuanced descriptors ("waxy," "matte," "oily"). Participants also valued the strong cross-modal coherence between visuals and haptics—"the image matched the feel"—and suggested lightweight refinements for future versions, including small attribute sliders for roughness, slipperiness, and hardness, a visible prompt history, A/B comparison thumbnails, and short hover hints linking adjectives to their expected haptic effects.

VI. Discussion

A. Language as a Control Modality for Texture Authoring

Our findings validate natural language as an interpretable control for texture generation. Linguistic prompts reliably steered haptic and visual signals along perceptual gradients, with user ratings confirming that generated intermediates represent meaningful interpolations rather than arbitrary blends. High HXI scores (Autotelics, Involvement, Realism) further demonstrate that users found the workflow engaging and the results believable. Collectively, these outcomes suggest text is a practical, intuitive driver for creative design, paving the way for hybrid interfaces that combine high-level descriptive prompting with fine-grained parametric refinement.

We also examined the model's behavior on materials not present in the HaTT dataset. Because CLIP aligns unseen concepts (e.g., "gelatin," "skin") with semantically related training examples in the latent space, the model generates haptics based on the closest learned physics. In our tests, "skin" consistently yielded sensations reported as realistic, and "gel" was described as "soft and sticky"—likely interpolating from compliant materials like rubber. However, this extrapolation is

bounded by the dataset's physical scope; "sand," for instance, rendered as a rigid "sandpaper-like" texture, as the underlying training data lacks the granular mechanics of loose particles.

B. Limitations and threats to validity

- **Data coverage.** Current training relies on synthetic augmentation of the HaTT dataset. We are actively addressing this by compiling a corpus of $\sim$300 textures with sufficient intra-class variation to eliminate the need for augmentation. Additionally, the reliance on GPT-5 descriptions was necessitated by the unavailability of physical source samples for human evaluators; future work will utilize physically accessible libraries to enable human-verified haptic grounding.
- **Device dependence.** Playback was on a 3-DOF kinesthetic device at 1 kHz. Absolute feel (gain, bandwidth, friction rendering) may differ on other hardware; replication on alternative stylus/actuator stacks is a priority.
- **Decoupled visuals.** Images are generated by a separate diffusion model; while users rarely reported mismatch, we do not guarantee physical consistency (e.g., SVBRD-F/roughness) beyond semantic agreement.
- **Baseline Comparison.** Our evaluation focused on the proposed system in isolation. Since no other language-guided multimodal texture authoring tool currently exists, and manual parameter tuning represents a fundamentally different interaction paradigm, a direct baseline comparison was not feasible. Consequently, the reported metrics (e.g., HXI) should be interpreted as validation of the system's absolute usability and workflow feasibility, rather than evidence of quantitative improvement over existing methods.
- **Linguistic Robustness.** We did not systematically evaluate the model's sensitivity to paraphrasing. Future work will quantify latent space stability across synonymous descriptors (e.g., comparing "smooth" vs. "polished") to ensure that minor textual variations result in predictable, consistent haptic outputs.

C. Engineering lessons

We observe three issues during system construction and user study: (i) participants report occasional prompt$\rightarrow$render latency spikes, (ii) default friction/roughness gains were sometimes too strong for some participants, and (iii) abrupt behavior occurred outside the recorded force–speed hull. Mitigations for (i) and (ii) include streaming decode and per-material gain auto-scaling. (iii) has since been resolved by incorporating *edge models* that smoothly extend dynamics beyond the measured range, eliminating the discontinuities observed during early testing.

D. Future work

Cross-modal alignment. Our immediate goal is to make the diffusion image respond to, and constrain, the haptic latent rather than relying on text alone. We will (i) learn a shared, bidirectionally predictive space in which images and haptics encode to the same z and can cross-reconstruct, and (ii) cross-condition the diffusion UNet on z via a lightweight adapter (e.g., cross-attention) so that image microgeometry (grain/pores/finish) is consistent with AR band energy and tap rise/decay. To stabilize alignment, we will add weak physics-aware priors that correlate visual cues with haptic statistics and train on text–image–haptics triplets (paired or pseudo-paired) with consistency losses.

Human Generated Texture. Our text-conditioning currently relies on AI-generated descriptions to represent material semantics. A valuable next step is to incorporate **human-annotated descriptions** of textures—collected through crowdsourcing or expert labeling—to establish stronger grounding between language and human haptic perception. Such human supervision could refine semantic alignment from human tactile understanding.

Beyond isotropy/rigidity. We plan to relax our assumptions by modeling *anisotropy* (directional grains, woven fabrics) and *rate dependence* (viscoelastic media). Practically, this means collecting directional AR grids and variable-rate tap banks and extending the latent to $z(f, v, \theta, \dot{v})$ so synthesis reflects exploration heading and speed changes. Additionally, modeling soft materials can be approached either through data driven or model-based methods.

Authoring and portability. For authoring, we will expose semantic sliders (rough$\leftrightarrow$smooth, hard$\leftrightarrow$soft, slippery$\leftrightarrow$sticky) that move z along calibrated axes, add token-level attribution to show which words influenced each attribute, and provide certainty cues and auto-gain to avoid over/under-emphasis. For portability, we will release calibration recipes and learn device adapters so the same latent produces comparable feel across different end-effectors (e.g., voice-coil, piezo, ultrasonic friction).

VII. Conclusion

We presented a language-guided authoring system that synthesizes coordinated haptic (vibration, tapping) and visual textures from natural language. By aligning a multimodal VAE with CLIP, we established a semantic latent space where prompts like "gritty stone" yield physically grounded, cross-modally consistent signals. User evaluations confirmed that this latent structure correlates with human perception across roughness, hardness, and slipperiness, enabling an intuitive, prompt-driven workflow. Future work will address current data constraints by expanding the system to support bare-finger interactions and open-set material generalization.

REFERENCES

[1] K. Theivendran, A. Wu, W. Frier, and O. Schneider, "Rechap: an interactive recommender system for navigating a large number of mid-air haptic designs," *IEEE Transactions on Haptics*, vol. 17, no. 2, pp. 165–176, 2023.

[2] Z. Wang, A. Li, Z. Li, and X. Liu, "Genartist: Multimodal llm as an agent for unified image generation and editing," *Advances in Neural Information Processing Systems*, vol. 37, pp. 128 374–128 395, 2024.

[3] J. M. Romano and K. J. Kuchenbecker, "Creating realistic virtual textures from contact acceleration data," *IEEE Transactions on Haptics*, vol. 5, no. 2, pp. 109–119, 2011.

[4] C. Basdogan, S. De, J. Kim, M. Muniyandi, H. Kim, and M. A. Srinivasan, "Haptics in minimally invasive surgical simulation and training," *IEEE Computer Graphics and Applications*, vol. 24, no. 2, pp. 56–64, 2004.

[5] P. Xia, A. M. Lopes, and M. T. Restivo, "A review of virtual reality and haptics for product assembly (part 1): rigid parts," *Assembly Automation*, vol. 33, no. 1, pp. 68–77, 2013.

[6] S. Choi and K. J. Kuchenbecker, "Vibrotactile display: Perception, technology, and applications," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2093–2104, 2012.

[7] H. Culbertson, J. Unwin, and K. J. Kuchenbecker, "Modeling and rendering realistic textures from unconstrained tool-surface interactions," *IEEE Transactions on Haptics*, vol. 7, no. 3, pp. 381–393, 2014.

[8] H. Seifi, K. Zhang, and K. E. MacLean, "Vibviz: Organizing, visualizing and navigating vibration libraries," in *IEEE World Haptics Conference (WHC)*, 2015, pp. 254–259.

[9] H. Culbertson, J. J. L. Delgado, and K. J. Kuchenbecker, "One hundred data-driven haptic texture models and open-source methods for rendering on 3d objects," in *IEEE Haptics Symposium (HAPTICS)*, 2014, pp. 319–325.

[10] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical dimensions of tactile perception of textures," *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 81–93, 2012.

[11] Y. Ujitoko and Y. Ban, "Vibrotactile signal generation from texture images or attributes using generative adversarial network," in *International conference on human haptic sensing and touch enabled computer applications*. Springer, 2018, pp. 25–36.

[12] Y. Sung, K. John, S. H. Yoon, and H. Seifi, "Hapticgen: Generative text-to-vibration model for streamlining haptic design," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–24.

[13] W. M. B. Tiest, "Tactual perception of material properties," *Vision research*, vol. 50, no. 24, pp. 2775–2782, 2010.

[14] W. Hassan, A. Abdulali, and S. Jeon, "Authoring new haptic textures based on interpolation of real textures in affective space," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 1, pp. 667–676, 2020.

[15] K. Tozuka, B. Poitrimol, G. Sasaki, K. Kobayashi, and H. Igarashi, "Integrating texture models through regression of vibration and texture characteristics," *ROBOMECH Journal*, vol. 12, no. 1, p. 25, 2025.

[16] S. Lu, M. Zheng, M. C. Fontaine, S. Nikolaidis, and H. Culbertson, "Preference-driven texture modeling through interactive generation and search," *IEEE Transactions on Haptics*, vol. 15, no. 3, pp. 508–520, 2022.

[17] N. Heravi, W. Yuan, A. M. Okamura, and J. Bohg, "Learning an action-conditional model for haptic texture generation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 11 088–11 095.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[19] H. Culbertson, S. B. Schorr, and A. M. Okamura, "Haptics: The present and future of artificial touch sensation," *Annual review of control, robotics, and autonomous systems*, vol. 1, no. 1, pp. 385–409, 2018.

[20] J. E. Colgate, L. A. Jones, and H. Z. Tan, "Twenty years of world haptics: Retrospective and future directions," *IEEE Transactions on Haptics*, vol. 18, no. 3, pp. 452–455, 2025.

[21] J. J. Gibson, "Observations on active touch." *Psychological review*, vol. 69, no. 6, p. 477, 1962.

[22] M. A. Heller, "Visual and tactual texture perception: Intersensory cooperation," *Perception & psychophysics*, vol. 31, no. 4, pp. 339–344, 1982.

[23] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *Trends in Cognitive Sciences*, vol. 8, no. 4, pp. 162–169, 2004.

[24] R. L. Klatzky and S. J. Lederman, "Multisensory texture perception," in *Multisensory object perception in the primate brain*. Springer, 2010, pp. 211–230.

[25] B. Xiao, W. Bi, X. Jia, H. Wei, and E. H. Adelson, "Can you see what you feel? color and folding properties affect visual–tactile material discrimination of fabrics," *Journal of Vision*, vol. 16, no. 3, pp. 34–34, 2016.

[26] C. Koniaris, D. Cosker, X. Yang, and K. Mitchell, "Survey of texture mapping techniques for representing and rendering volumetric mesostructure," *Journal of Computer Graphics Techniques*, 2014.

[27] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real-world surfaces," *ACM Trans. Graph.*, vol. 18, no. 1, p. 1–34, Jan. 1999.

[28] K. Park, K. Rematas, A. Farhadi, and S. M. Seitz, "Photoshape: photorealistic materials for large-scale shape collections," *ACM Trans. Graph.*, vol. 37, no. 6, Dec. 2018.

[29] Y. Fang, X. Zhang, W. Xu, G. Liu, and J. Zhao, "Bidirectional visual-tactile cross-modal generation using latent feature space flow model," *Neural Networks*, vol. 172, p. 106088, 2024.

[30] L. R. Manfredi, H. P. Saal, K. J. Brown, M. C. Zielinski, J. F. Dammann III, V. S. Polashock, and S. J. Bensmaia, "Natural scenes in tactile texture," *Journal of neurophysiology*, vol. 111, no. 9, pp. 1792–1802, 2014.

[31] A. L. Stefani, N. Bisagno, A. Rosani, N. Conci, and F. De Natale, "Signal processing for haptic surface modeling: A review," *Signal Processing: Image Communication*, p. 117338, 2025.

[32] S. Chen, T. Yuan, L. Xu, W. Ru, and D. Wang, "A systematic review of haptic texture reproduction technology," *Intelligence & Robotics*, vol. 5, no. 3, pp. 607–30, 2025.

[33] A. Okamura, M. Cutkosky, and J. Dennerlein, "Reality-based models for vibration feedback in virtual environments," *IEEE/ASME Transactions on Mechatronics*, vol. 6, no. 3, pp. 245–252, 2001.

[34] C. G. McDonald and K. J. Kuchenbecker, "Dynamic simulation of tool-mediated texture interaction," in *2013 World Haptics Conference (WHC)*, 2013, pp. 307–312.

[35] H. Culbertson, J. Unwin, B. E. Goodman, and K. J. Kuchenbecker, "Generating haptic texture models from unconstrained tool-surface interactions," in *World Haptics Conference (WHC)*, 2013, pp. 295–300.

[36] S. Shin, R. H. Osgouei, K.-D. Kim, and S. Choi, "Data-driven modeling of isotropic haptic textures using frequency-decomposed neural networks," in *IEEE World Haptics Conference (WHC)*, 2015, pp. 131–138.

[37] J. B. Joolee and S. Jeon, "Data-driven haptic texture modeling and rendering based on deep spatio-temporal networks," *IEEE Transactions on Haptics*, vol. 15, no. 1, pp. 62–67, 2022.

[38] N. Heravi, H. Culbertson, A. M. Okamura, and J. Bohg, "Development and evaluation of a learning-based model for real-time haptic texture rendering," *IEEE Transactions on Haptics*, vol. 17, no. 4, pp. 705–716, 2024.

[39] X. Yi, J. Wang, and P. Sun, "Time series diffusion method: A denoising diffusion probabilistic model for vibration signal generation," *arXiv preprint arXiv:2303.01234*, 2023.

[40] K. J. Kuchenbecker, J. Fiene, and G. Niemeyer, "Improving contact realism through event-based haptic feedback," *IEEE transactions on visualization and computer graphics*, vol. 12, no. 2, pp. 219–230, 2006.

[41] L. Winfield, J. Glassmire, J. E. Colgate, and M. Peshkin, "T-pad: Tactile pattern display through variable friction reduction," in *Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'07)*, 2007, pp. 421–426.

[42] D. J. Meyer, M. A. Peshkin, and J. E. Colgate, "Fingertip friction modulation due to electrostatic attraction," in *2013 World Haptics Conference (WHC)*, 2013, pp. 43–48.

[43] C. Schwarz, "The slip hypothesis: tactile perception and its neuronal bases," *Trends in neurosciences*, vol. 39, no. 7, pp. 449–462, 2016.

[44] S. Lu, Y. Chen, and H. Culbertson, "Towards multisensory perception: Modeling and rendering sounds of tool-surface interactions," *IEEE Transactions on Haptics*, vol. 13, no. 1, pp. 94–101, 2020.

[45] B. Zhong, S. Je, and M. Z. Sagesser, "Auditory materiality: Exploring auditory effects on material perception in virtual reality," in *2025 IEEE*

Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2025, pp. 659–662.

[46] A. Niijima, T. Takeda, T. Mukouchi, and T. Satou, "Thermalbitdisplay: Haptic display providing thermal feedback perceived differently depending on body parts," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–8.

[47] E.-H. Lee, S.-H. Kim, and K.-S. Yun, "Three-axis pneumatic haptic display for the mechanical and thermal stimulation of a human finger pad," in *Actuators*, vol. 10, no. 3. MDPI, 2021, p. 60.

[48] V. Shen, T. Rae-Grant, J. Mullenbach, C. Harrison, and C. Shultz, "Fluid reality: High-resolution, untethered haptic gloves using electroosmotic pump arrays," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–20.

[49] W. Qian, C. Gao, A. Sathya, R. Suzuki, and K. Nakagaki, "Shape-it: Exploring text-to-shape-display for generative shape-changing behaviors with llms," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, pp. 1–29.

[50] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "Texturegan: Controlling deep image synthesis with texture patches," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8456–8465.

[51] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM computing surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[52] Y. Koyama, I. Sato, and M. Goto, "Sequential gallery for interactive visual design optimization," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 88–1, 2020.

[53] D. Z. Chen, Y. Siddiqui, H.-Y. Lee, S. Tulyakov, and M. Nießner, "Text2tex: Text-driven texture synthesis via diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 558–18 568.

[54] Y. Ujitoko and Y. Ban, "Vibrotactile signal generation from texture images or attributes using generative adversarial network," in *Haptics: Science, Technology, and Applications*, D. Prattichizzo, H. Shinoda, H. Z. Tan, E. Ruffaldi, and A. Frisoli, Eds. Cham: Springer International Publishing, 2018, pp. 25–36.

[55] M. Zhang, S. Terui, Y. Makino, and H. Shinoda, "TEXasGAN: Tactile texture exploration and synthesis system using generative adversarial network," *arXiv preprint arXiv:2407.11467*, 2024.

[56] F. Faruqi, M. Perroni-Scharf, J. S. Walia, Y. Zhu, S. Feng, D. Degraen, and S. Mueller, "Tactstyle: Generating tactile textures with generative ai for digital fabrication," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–16.

[57] S. Cai, L. Zhao, Y. Ban, T. Narumi, Y. Liu, and K. Zhu, "Gan-based image-to-friction generation for tactile simulation of fabric material," *Computers & Graphics*, vol. 104, pp. 219–228, 2022.

[58] G. Cao, J. Jiang, N. Mao, D. Bollegala, M. Li, and S. Luo, "Vis2hap: Vision-based haptic rendering by cross-modal generation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12 443–12 449.

[59] Q. Xi, F. Wang, L. Tao, H. Zhang, X. Jiang, and J. Wu, "Cm-avae: Cross-modal adversarial variational autoencoder for visual-to-tactile data generation," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5214–5221, 2024.

[60] H. Zhan, J. Chen, and L. Huang, "Method for audio-to-tactile cross-modality generation based on residual u-net," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

[61] R. Song, X. Sun, and G. Liu, "Cross-modal generation of tactile friction coefficient from audio and visual measurements by transformer," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[62] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[63] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[64] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[65] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 300–309.

[66] J. Xu, X. Wang, W. Cheng, Y.-P. Cao, Y. Shan, X. Qie, and S. Gao, "Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 908–20 918.

[67] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[68] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[69] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[70] J. Tu, H. Fu, F. Yang, H. Zhao, C. Zhang, and H. Qian, "Texttoucher: Fine-grained text-to-touch generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 7455–7463.

[71] M. Naeem, M. I. Awan, and S. Jeon, "Text-driven generative framework for multimodal visual and haptic texture synthesis," in *2025 IEEE World Haptics Conference (WHC)*, 2025, pp. 140–146.

[72] H. Culbertson and K. J. Kuchenbecker, "Penn haptic texture toolkit: A collection of textures for haptic rendering," in *IEEE Haptics Symposium (HAPTICS)*, 2014, pp. 99–106.

[73] ——, "Importance of matching physical friction, hardness, and texture in creating realistic haptic virtual surfaces," *IEEE Transactions on Haptics*, vol. 10, no. 1, pp. 63–74, 2016.

[74] OpenAI, "Chatgpt (gpt-5)," https://chat.openai.com, 2025, large language model developed by OpenAI. URL: https://chat.openai.com.

[75] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.

[76] "Stable diffusion v2-base — stabilityai / hugging face," https://huggingface.co/stabilityai/stable-diffusion-2-base, 2025, accessed: 2025-09-30.

[77] "Textures — poly haven," https://polyhaven.com/textures, 2025, accessed: 2025-09-30.

[78] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "Clustergan: Latent space clustering in generative adversarial networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 4610–4617.

[79] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*. PMLR, 2016, pp. 478–487.

[80] M. Ben-Yosef and D. Weinshall, "Gaussian mixture generative adversarial networks for diverse datasets, and the unsupervised clustering of images," *arXiv preprint arXiv:1808.10356*, 2018.

[81] T. Shi and O. Schneider, "Development and initial validation of the haptic experience inventory (hxi)," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–21.

[82] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in Psychology*. Elsevier, 1988, vol. 52, pp. 139–183.