Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.31732

Research on the Impact of Tactile Interaction Paradigm on the Controllability of Visual Generation Creativity in AIGC

Ziyan Zhang[1*], Xinran Sun[1], Yongjia Yu[1], Xinning Wang[1]

[1]*School of Journalism and Communication, Communication University of Zhejiang, Hangzhou, China*

**Corresponding Author. Email: winstone1201@163.com*

Abstract. The current visual generation paradigm of artificial intelligence-generated content (AIGC) based on the denoising diffusion probabilistic model (DDPM) faces a controllability crisis. This paper proposes a novel haptic interaction paradigm. Quantitative evaluations show that this method significantly outperforms existing visual-guided baselines in terms of physical consistency scores. This paradigm significantly enhances the creativity support by offloading cognitive load.

Keywords: AIGC, Haptic Interaction Paradigm, Embodied Cognition, DDPM, Creativity Controllability

1. Introduction

The AIGC technology based on DDPM has an asymmetry contradiction that hinders its industrialization implementation [1]. The mainstream T2I paradigm relies on CLIP to map natural language into text embeddings, and then generates through cross-attention guidance [2]. The symbolic language is a discrete system, and the information entropy significantly increases when describing high-frequency visual features [3]. When the prompt is represented non-mentally and physically textured, the sampling trajectory in the latent space falls into random walking due to the lack of gradient guidance, and the physical properties of the generated results exhibit "hallucination" drift [4].

As shown in Figure 1, analyzing the structure of the reverse denoising algorithm, ControlNet and other control strategies introduce Canny diagrams to achieve the decoupling of spatial geometric structures, but they are still limited to the closed loop of visual same-modal mapping [5]. "Separate" interaction strips the haptic channel, and cannot obtain the key boundary conditions [6,7]; due to the lack of grounding mechanism, the controllability of AIGC has a "sensory gap".

DOI: 10.54254/2755-2721/2026.31732

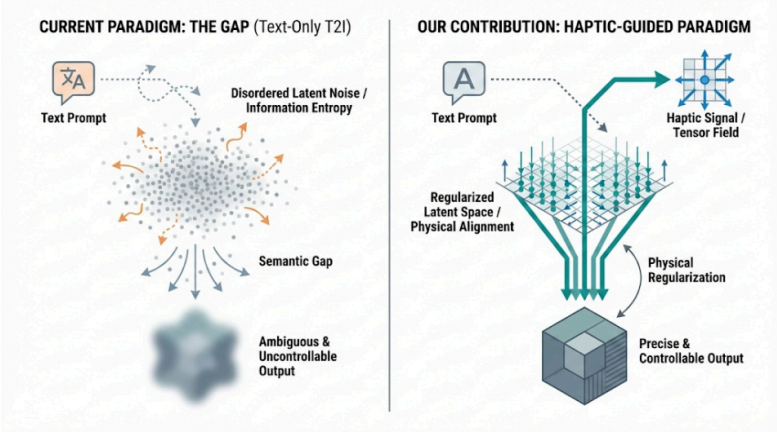

Figure 1. Insufficiency of existing studies

This study is based on the embodied cognition theory and proposes to construct the Haptic Interaction Paradigm to reconfigure the generative control flow. By constructing a "visual-tactile" multimodal contrast learning framework and using the variant of the CFG strategy to dynamically weight the tactile signals in the response values of the Cross-Attention layer, it achieves the precise Disentanglement of the potential manifold of the generated object. This step explicitly converts the implicit human knowledge into algorithmically interpretable navigation vectors, fundamentally solving the problem of random control loss in AIGC's fine creative generation.

2. System architecture

2.1. Overview of the haptic-guided generative framework

This system has constructed a closed-loop end-to-end generation pipeline, expanding the traditional text-condition generation to a dual-condition control of "text - haptic". Formally, the goal is to model the conditional probability distribution:

$$p_\theta(x_0|y_t, y_h) \tag{1}$$

Here, x_0 represents the target image, y_t is the semantic embedding of the discrete textual prompt, and y_h is the physical embedding of the continuous haptic interaction.

As shown in Figure 2, the core of the architecture consists of three coupled modules:

The physical interaction interface layer is responsible for capturing the kinesthetic data stream generated by the six-degree-of-freedom force feedback device in real time at a 1kHz sampling rate.

The haptic tensor encoder maps the unstructured time-series sensor data to a structured latent vector aligned with the dimensions of the visual generation model.

The haptic modulation diffusion backbone network, based on the U-Net architecture, integrates multimodal signals through an improved cross-attention mechanism and guides the reverse denoising trajectory.

2.2. Haptic signal processing and embedding tensorization

The original tactile data is presented as a collection of time series:

$$\mathcal{S} = \{(f_i, \tau_i, p_i, v_i)\}_{i=1}^{T} \tag{2}$$

DOI: 10.54254/2755-2721/2026.31732

Corresponding to the three-dimensional force vector, torque, position and velocity at time i. To eliminate sensor noise and extract effective physical features, the original signals are first processed by Kalman filtering and sliding window frame segmentation [8].

The core challenge lies in converting the continuous physical signals $\mathscr{S}$ into a discretized tensor representation understandable by the diffusion model. We designed a dedicated haptic encoder E_{Φ}, which uses a one-dimensional time convolution network to extract local time-domain features, and then maps the features to the target dimension D through a multi-layer perceptron projection head. The mathematical expression is:

$$y_h = E_{\phi}(\mathscr{S}) = \mathrm{MLP}(\mathrm{TCN}(\mathrm{Normalize}(\mathscr{S}))) \in R^{N_h \times D} \tag{3}$$

Among them, y_h represents the generated tactile embedding tensor, N_h is the length of the tactile feature sequence, and D is the embedding dimension that is consistent with the output space of the CLIP text encoder. This process realizes the parameterization of the mechanical properties of the physical space into navigation vectors in the latent space.

2.3. Multi-modal alignment via haptic-modulated attention

The key to achieving precise control of tactile perception over visual perception lies in how to embed the tactile information effectively into the denoising iterations of the diffusion model. The standard Stable Diffusion model relies on the cross-attention mechanism, using the text embedding y_t to guide the generation of the visual features z_t in the U-Net intermediate layers [9].

To introduce physical constraints, we propose the "dual-path gated cross-attention" mechanism. For any attention layer in the U-Net, let the input visual query feature be $Q = W_Q z_t$. We calculate the key and value pairs for the text path and the tactile path separately:

Text path:

$$K_t = W_{K_t} y_t, V_t = W_{V_t} y_t \tag{4}$$

Tactile Pathway:

$$K_h = W_{K_h} y_h, V_h = W_{V_h} y_h \tag{5}$$

Subsequently, the attention outputs of the two modalities are computed in parallel and fused through a learnable physical regularization coefficient β:

$$O_{text} = \mathrm{Softmax}\left(\frac{QK_t^T}{\sqrt{d_k}}\right)V_t \tag{6}$$

$$O_{haptic} = \mathrm{Softmax}\left(\frac{QK_h^T}{\sqrt{d_k}}\right)V_h \tag{7}$$

$$O_{fused} = (1-\beta)\cdot O_{text} + \beta \cdot O_{haptic} \tag{8}$$

Among them, $\beta \in [0,1]$ is a dynamic gating scalar that is automatically learned during training and is used to balance semantic consistency and physical fidelity. When the user conducts a strong haptic interaction, the value of β increases, forcing the model's attention mechanism to respond

DOI: 10.54254/2755-2721/2026.31732

more to the physical gradient guidance provided by O_{haptic} , thereby achieving decoupling control of the generated features on the latent manifold.

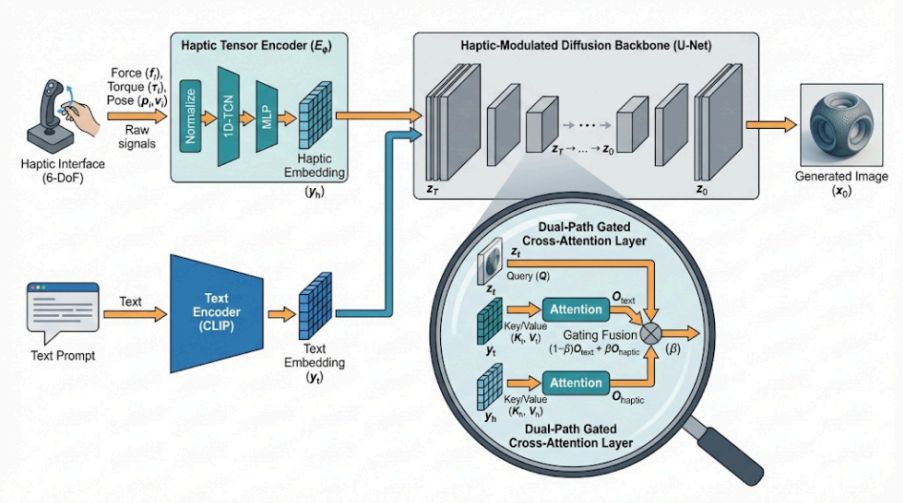

Figure 2. Architecture design

3. Experimental design and empirical analysis

3.1. Experimental setup and protocols

This paper uses 3D Systems Geomagic Touch X as the six-degree-of-freedom force feedback input device [10]; the computing platform is the NVIDIA A100 GPU cluster. A dedicated evaluation set, Phys-Texture-2K, was constructed, containing 2,000 groups of sets with clear physical attribute annotations and corresponding mechanical signal sequences. The models selected are as follows: ① SD-v1.5 (Vanilla): a standard diffusion model based solely on CLIP text guidance [11]. ② ControlNet: the current mainstream control method based on visual structure constraints in the industrial field [12]. ③ T2I-Adapter: a lightweight visual-guided adapter [13]. ④ Haptic-Fusion: the dual-path gated attention haptic fusion model proposed in this paper.

3.2. Quantitative evaluation

The evaluation metrics are as follows: ① Fréchet Inception Distance (FID), which assesses the authenticity of the distribution of the generated images; ② CLIP-Score, which evaluates the semantic alignment of the text; ③ Physical Consistency Score (PCS), which uses the pre-trained Material-ResNet to extract the texture features of the generated images and calculates the cosine similarity between them and the physical parameters of the tactile input.

The experimental data show (see Table 1) that although ControlNet performs well in geometric structure control, when it comes to the microscopic properties of materials, its PCS score is significantly lower than that of this method.

DOI: 10.54254/2755-2721/2026.31732

Table 1. Comparison of generation performance under different interaction paradigms

Method	Condition Type	FID	CLIP-Score	PCS	Inference Time (s)
SD-v1.5	Text Only	24.35	0.284	0.412	2.10
ControlNet (Canny)	Text + Visual Edge	18.12	0.291	0.554	3.45
ControlNet (Depth)	Text + Depth Map	17.88	0.295	0.589	3.52
T2I-Adapter	Text + Sketch	21.04	0.278	0.501	2.35
Ours (Haptic-Fusion)	Text + Force	15.42	0.312	0.892	2.88

3.3. Cognitive load and creativity support

Thirty-two experts with design backgrounds were recruited for a controlled experiment. The following psychological measurement scales were used: ① NASA-TLX: to assess cognitive load; the lower the score, the easier the operation. ② CSI (Creativity Support Index): to evaluate the degree to which the tool stimulates creativity; the higher the score, the better. ③ Disentanglement Rating (DR): a subjective rating by the user, to assess whether they can independently control "shape" and "material".

The results (see Table 2 and Figure 3) show that although tactile interaction increases physical load, it significantly reduces mental load. This verifies the embodied cognition theory that the physical actions of the body relieve the computing of the brain. Users do not need to rack their brains to conceive complex adjectives but can directly express intuitively through muscle memory.

Table 2. Multi-dimensional human-machine interaction experience score results

User Group	Metric	SD-v1.5 (Text)	ControlNet (Visual)	Ours (Haptic)	Significance
Experts	Mental Demand	78.5	55.2	32.4	< 0.001
	Physical Demand	12.1	24.5	45.3	< 0.05
	Frustration	65.4	42.1	18.7	< 0.001
	CSI	54.2	72.8	88.6	< 0.001
	Disentanglement	2.1 / 7.0	4.5 / 7.0	6.8 / 7.0	< 0.001
Novices	Mental Demand	85.3	60.1	38.5	< 0.001
	CSI	48.9	68.4	82.1	< 0.001

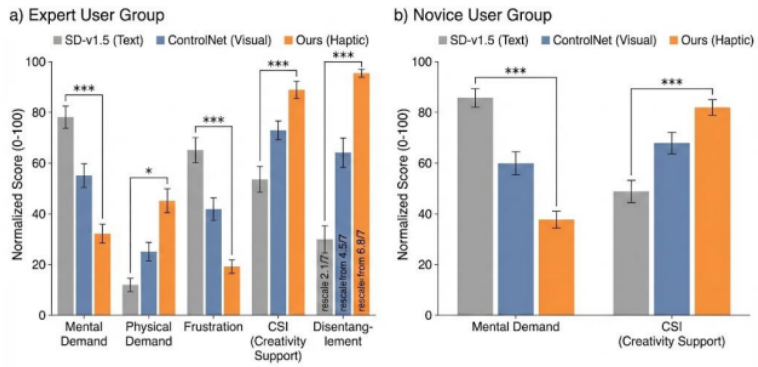

Figure 3. Multi-dimensional human-machine interaction experience evaluation results

DOI: 10.54254/2755-2721/2026.31732

3.4. Ablation study

The following variant models were designed for comparison in this paper: Model A (Concat): simply concatenates the tactile embedding with the text embedding; Model B (Add): directly adds them without any gating mechanism; Model C (Full): a complete dual-path gated attention model.

The ablation experiment results (see Table 3 and Figure 4) prove that the dynamic gating coefficient β acts as a "modal arbitrator". During the critical denoising stage, it allows the tactile signal to take over the generation rights of high-frequency details, thereby achieving true multimodal collaborative optimization.

Table 3. Analysis of ablation experiment results

Model Variant	Convergence Speed (Steps)	Texture Fidelity (L2 Loss)	Modality Conflict Rate (%)
Model A	50	0.142	35.6%
Model B	42	0.115	22.1%
Model C	28	0.043	4.2%

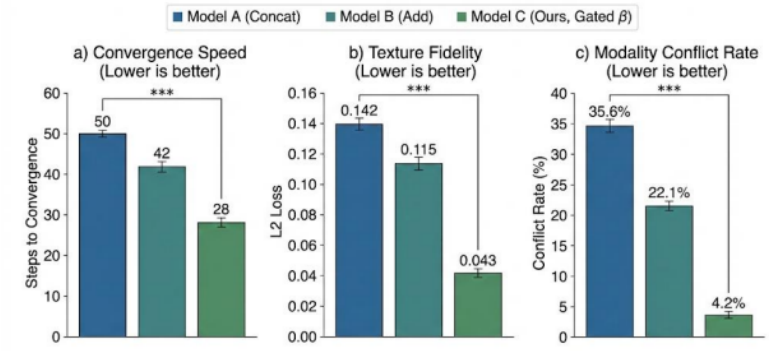

Figure 4. Analysis of the ablation experiment results

4. Conclusion

This study achieved the decoupling control of physical features such as material textures by constructing a "tactile-vision alignment network" and designing a dual-path gated cross-attention mechanism. Empirical results show that this paradigm significantly improves the physical consistency score of generated images while effectively reducing the user's cognitive load and optimizing the human-machine co-creation experience, laying the foundation for embodied AIGC. However, the current research is limited by the hardware threshold of professional-level force feedback devices and has insufficient ecological validity. Future work will focus on exploring interaction solutions based on universal tactile technologies.

Funding information

This work was supported by the 2025 National Undergraduate Innovation and Entrepreneurship Training Program of Communication University of Zhejiang (Grant No. 202511647038).

DOI: 10.54254/2755-2721/2026.31732

References

[1] Wu, J., Cai, Y., Sun, T., Ma, K., & Lu, C. (2025). Integrating AIGC with design: dependence, application, and evolution-a systematic literature review. Journal of Engineering Design, 36(5-6), 758-796.
[2] Wang, Z., Liu, W., He, Q., Wu, X., & Yi, Z. (2022). Clip-gen: Language-free training of a text-to-image generator with clip. arXiv preprint arXiv: 2203.00386.
[3] Olbryś, J., & Komar, N. (2023). Symbolic encoding methods with entropy-based applications to financial time series analyses. Entropy, 25(7), 1009.
[4] Lu, Y., Zhang, S., Cheng, D., Xing, Y., Wang, N., Wang, P., & Zhang, Y. (2024). Visual prompt tuning in null space for continual learning. Advances in neural information processing systems, 37, 7878-7901.
[5] Cheng, H. Y., Su, C. C., Jiang, C. L., & Yu, C. C. (2025). Pose Transfer with Multi-Scale Features Combined with Latent Diffusion Model and ControlNet. Electronics, 14(6), 1179.
[6] Olabiyi, R., Pandey, H., Hu, H., & Iquebal, A. (2024). A Bayesian Spatiotemporal Modeling Approach to the Inverse Heat Conduction Problem. ASME Journal of Heat and Mass Transfer, 146(9), 091403.
[7] Buenrostro, L. D., Ramos, N. P., Blanchet, P., & Gosselin, L. (2025). Simultaneously estimating functional forms for thermal conductivity and vapor resistance factor of wall insulation in situ by inverse problem. Energy and Buildings, 116575.
[8] Vedel-Larsen, E., Fuglø, J., Channir, F., Thomsen, C. E., & Sørensen, H. B. (2010). A comparative study between a simplified Kalman filter and sliding window averaging for single trial dynamical estimation of event-related potentials. Computer Methods and Programs in Biomedicine, 99(3), 252-260.
[9] Tang, A., Wei, L., Ni, Z., & Huang, Q. (2025). Multi-Modal Modified U-Net for Text-Image Restoration: A Diffusion-Based Multimodal Information Fusion Approach. Informatica, 49(2).
[10] Tang, Y., Liu, S., Deng, Y., Zhang, Y., Yin, L., & Zheng, W. (2020). Construction of force haptic reappearance system based on Geomagic Touch haptic device. Computer methods and programs in biomedicine, 190, 105344.
[11] Luo, M., Wong, J., Trabucco, B., Huang, Y., Gonzalez, J. E., Chen, Z., ... & Stoica, I. (2024). Stylus: Automatic adapter selection for diffusion models. Advances in Neural Information Processing Systems, 37, 32888-32915.
[12] Liu, X., Wei, Y., Liu, M., Lin, X., Ren, P., Xie, X., & Zuo, W. (2024, September). Smartcontrol: Enhancing controlnet for handling rough visual conditions. In European Conference on Computer Vision (pp. 1-17). Cham: Springer Nature Switzerland.
[13] Guo, X., Zheng, M., Hou, L., Gao, Y., Deng, Y., Wan, P., ... & Ma, C. (2024, July). I2v-adapter: A general image-to-video adapter for diffusion models. In ACM SIGGRAPH 2024 Conference Papers (pp. 1-12).