Human Factors, Architecture, Sustainable Urban Planning and Infrastructure, Vol. 229, 2026, 202–208
https://doi.org/10.54941/ahfe1007968

PEFT Strategies for Human–AI Co-Creation in Architectural Morphogenesis

Chie Fuyuki[1], Shuichi Nagao[2], Vadim Grigorev[1], Rim El Filali[1], and Jian Du[3]

[1]Tsinghua University, Beijing, China
[2]ZEALS Inc, Tokyo, Japan
[3]Tsinghua Shenzhen International Graduate School, China

ABSTRACT

Generative AI is reshaping early-stage architectural design, yet the cognitive demands placed on designers navigating highly parameterised workflows remain underexamined. This study investigates how four parameter-efficient fine-tuning (PEFT) strategies in Stable Diffusion influence human–AI co-creation, interpretability, and designer control in biomimetic architectural morphogenesis. Four configurations—Baseline, Shared LoRA, IP-Adapter (w = 0.8), and a custom fine-tuned LoRA—were evaluated using MorphEvalBench, a newly developed evaluation toolkit. Fidelity and diversity metrics were operationalised as proxies for designer cognition: biomimetic fidelity (DSS), distributional fidelity (FID/KID), and diversity (DIV). A visual benchmarking grid across representative token levels complemented quantitative analysis. Results reveal that IP-Adapter achieved the strongest quantitative profile across all dimensions, yet its visual conditioning mechanism introduces patterns independently of prompt specification at $w \geq 0.6$, reducing designer control and requiring management of multiple interacting parameters. Fine-tuned LoRA demonstrated a balanced profile with greater semantic stability, enabling prompt-driven control and emergent diversity driven by designer input. Notably, the fine-tuning process itself constitutes a design act, embedding the designer's intent into the model prior to generation. These findings demonstrate that quantitative superiority does not guarantee cognitive supportiveness, and propose four implications for designing cognitively supportive human–AI workflows in architectural practice.

Keywords: Generative AI, Human–AI co-creation, Computational design, Morphogenesis, Cognitive load, Parameter-efficient fine-tuning

INTRODUCTION

The integration of generative AI into architectural design practice has accelerated dramatically, offering designers new capacities for morphological exploration. However, as generative workflows grow more complex, the cognitive demands placed on designers have received comparatively little attention. In particular, when designers interact with parameter-efficient fine-tuning (PEFT) strategies in diffusion-based models, they face challenges

Received March 20, 2026; Revised April 24, 2026; Accepted May 12, 2026; Available online July 20, 2026

of interpretability, predictability, and control that go beyond conventional usability concerns.

Cognitive load in human–AI co-creation is shaped not only by interface design but by the underlying generative behaviour of the model itself. A system that produces unpredictable or semantically inconsistent outputs forces designers to invest additional cognitive resources in interpreting and correcting results, reducing the capacity available for higher-order tasks such as creative reasoning. Despite growing interest in AI-assisted design tools, the human-factors implications of different model configurations remain underexplored.

This study examines how four PEFT configurations—Baseline, Shared LoRA, IP-Adapter, and a custom fine-tuned LoRA— shape the interpretive demands placed on designers in a biomimetic architectural morphogenesis task. Using MorphEvalBench, a newly developed perceptual evaluation toolkit, fidelity and diversity metrics are operationalised as proxies for interpretability and designer control. The findings contribute four implications for the design of cognitively supportive human–AI workflows in architectural practice.

RELATED WORK

Research on human–AI co-creative systems has identified cognitive overload, role ambiguity, and control conflicts as critical challenges that shape the quality of collaborative design outcomes (Salma, Hijón-Neira and Pizarro, 2025). In generative design contexts, these challenges are amplified by the stochastic nature of diffusion-based models, where small changes in configuration can produce dramatically different outputs, imposing unpredictable interpretive demands on designers. According to Cognitive Load Theory, when processing demands are high, fewer cognitive resources remain for other demanding cognitive activities, including those involved in creative reasoning (Sweller, 1988). While statistical metrics such as FID have been widely adopted to evaluate generative outputs, they may overlook domain-specific and functional aspects of design that are critical for architectural and urban evaluation (Brama et al., 2024). By extension, this limitation suggests that evaluation in AI-assisted design should also consider how outputs support designer interpretation, decision-making, and creative reasoning.

PEFT in diffusion models. Parameter-efficient fine-tuning methods enable targeted adaptation of large pretrained diffusion models without full retraining. LoRA (Hu et al., 2021) introduces low-rank weight adaptations for domain-specific tuning with minimal parameters. IP-Adapter (Ye et al., 2023) conditions generation on reference images via decoupled cross-attention, enabling strong visual transfer. While these methods have been evaluated primarily for visual quality, their implications for designer cognition remain unaddressed.

Perceptual evaluation as a human-factors proxy. DreamSim-based perceptual similarity measures have emerged as tools sensitive to human perceptual judgements beyond pixel-level fidelity. This study reframes such metrics as proxies for the interpretive effort designers must expend when evaluating and iterating on AI-generated outputs.

METHOD

Generative framework. Four PEFT configurations were evaluated within a unified workflow built on Stable Diffusion XL (Podell et al., 2023) and ControlNet (Zhang, Rao and Agrawala, 2023). All configurations were implemented using ArchMorphSD (Fuyuki, 2025), a custom ComfyUI workflow designed for architectural morphogenesis. The four configurations are: (A) Baseline SDXL with no adaptation; (B) Shared LoRA trained on landscape imagery; (C) IP-Adapter conditioned on Tafoni stone reference images (w = 0.8); and (D) a custom fine-tuned LoRA trained exclusively on Tafoni textures. Unlike the other configurations, Fine-tuned LoRA involves a dedicated training phase in which the curation of training data itself constitutes a design decision, embedding the designer's intent into the model prior to generation. All configurations used identical ControlNet outlines, prompts, and seeds to ensure comparability. A consistent prompt structure ('A stunning white organic building...') was used across token levels, with the pattern-specifying key term varied to define each token level (Token 1: key term not specified; Token 5: 'tafoni rock pattern'; Token 7: 'tafonihoney rock formation pattern'). Fifty samples per configuration were generated across multiple token levels.

Evaluation toolkit. To assess generative outputs across three fidelity dimensions relevant to designer cognition, MorphEvalBench (Nagao, 2025) was used as the evaluation infrastructure for this study: biomimetic fidelity (alignment to natural referents, measured via DreamSim-based DSS), distributional fidelity (consistency within a generative family, measured via FID and KID variants), and diversity (variation across outputs, measured via DreamSim-derived DIV score). These metrics were operationalised as proxies for interpretability and designer control: low DSS indicates closer biomimetic alignment to the target Tafoni referents; low FID/KID indicates distributional consistency; high DIV indicates exploratory range. In addition, a visual benchmarking grid was conducted across three representative token levels under a fixed prompt structure (Prompt #1): Token 1 (key term not specified; 'A stunning white organic building'), Token 5 (with the key term 'tafoni rock pattern'), and Token 7 (with the key term 'tafonihoney rock formation pattern'). This setup isolates the effect of pattern token specification on each PEFT configuration.

Human-factors framing. Cognitive load implications were inferred from metric profiles across configurations. DSS is operationalised as a DreamSim distance measure, where lower values indicate stronger biomimetic alignment: configurations with low DSS demonstrate closer correspondence to the target domain, reducing the interpretive effort required to recognise intent alignment in outputs. However, DSS alone does not capture designer control; a configuration may achieve low DSS through autonomous visual conditioning rather than prompt-driven specification—a distinction with direct human-factors consequences. Configurations with high FID/KID indicate distributional unpredictability, potentially increasing corrective iteration cycles. Configurations with low DIV offer insufficient exploratory variation, reducing the utility of the generative system for early-stage design.

RESULTS

Quantitative results are summarised in Table 1 (FID, KID, DIV, and DSS). Group means were computed across all token levels per configuration. Bold values indicate best performance per metric.

Baseline (A) produced moderate scores across all metrics. The absence of domain-specific adaptation resulted in shallow pattern expression, offering limited biomimetic alignment and insufficient exploratory range. From a human-factors perspective, outputs were neither sufficiently faithful nor sufficiently diverse to support productive co-creative workflows.

Table 1: Mean evaluation metrics per PEFT configuration (bold = best value per metric; DSS ↓: lower = stronger biomimetic fidelity).

Configuration	FID ↓	KID ↓	DIV ↑	DSS ↓
A: Baseline	454.6	0.504	0.163	0.676
B: Shared LoRA	459.9	0.534	0.150	0.682
C: IP-Adapter	**393.4**	**0.382**	**0.165**	**0.546**
D: Fine-tuned LoRA	420.6	0.450	0.164	0.627

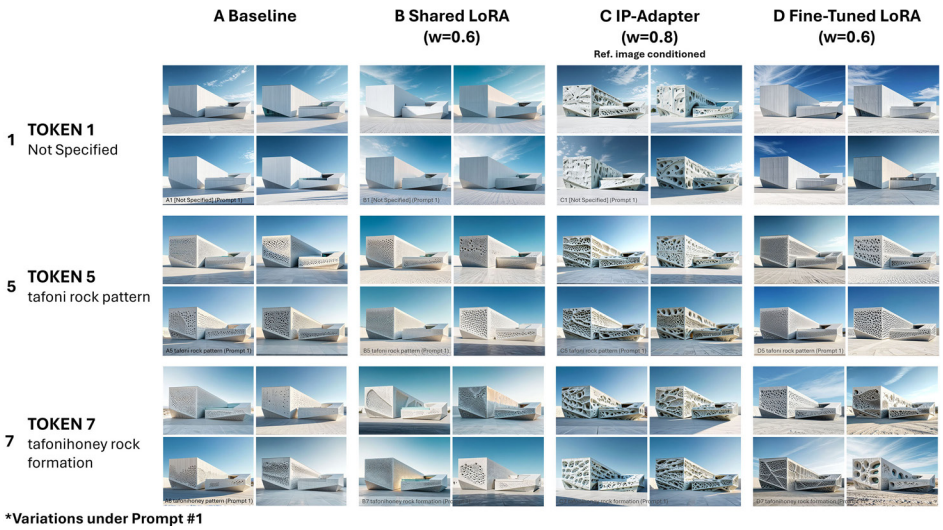

Figure 1: Visual benchmarking grid across four PEFT configurations (A–D) at representative token levels (1, 5, and 7). Token levels reflect increasing semantic complexity under a consistent prompt structure, from no semantic guidance (Token 1) to material grounding (Token 5) to distributionally complex pattern formation (Token 7).

Shared LoRA (B) showed the weakest overall profile, with the highest FID and KID values and the lowest diversity score. Landscape-trained weights partially overwrote SDXL's intrinsic micro-texture capacity without introducing Tafoni semantics. This configuration may represent the least cognitively supportive option.

IP-Adapter (C) achieved the strongest metric profile across all four dimensions: lowest FID (393.4), lowest KID (0.382), highest DIV (0.165), and strongest biomimetic fidelity (DSS: 0.546). However, visual benchmarking

revealed that at w = 0.8, IP-Adapter introduced Tafoni patterns even at Token 1, where no semantic guidance was specified. While IP-Adapter supports prompt-responsive modes through parameter adjustment, achieving consistent prompt alignment in practice requires managing multiple interacting variables—weight, prompt importance, and reference image selection— imposing a learning curve and trial-and-error burden that may itself constitute a significant source of interpretive demand.

To further investigate this behaviour, Figure 2 presents IP-Adapter outputs across weight values (w=0.0–1.0) under the Token 1 condition (key term not specified; see Methods). Results confirm that at practical adaptation strengths (w≥0.6), reference image features are introduced regardless of prompt specification. This suggests that the observed reduction in designer control is an inherent characteristic of visual conditioning at operationally relevant weight settings.

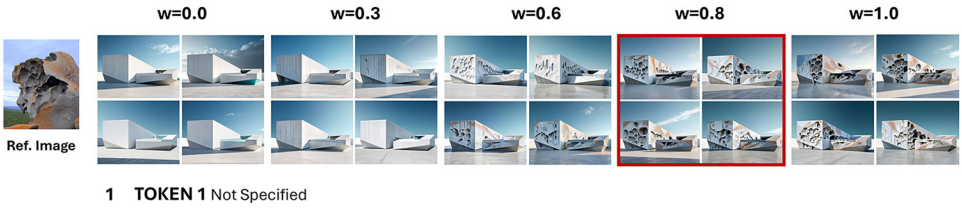

Figure 2: IP-Adapter outputs under the Token 1 condition (Prompt #1, key term not specified; cf. Figure 1, top row) across weight values (w=0.0–1.0). At w≥0.6, reference image features emerge despite the absence of any pattern-related token in the prompt. Red box indicates w=0.8, the setting used in the primary experiment.

Fine-tuned LoRA (D) demonstrated a balanced metric profile: distributional fidelity second only to IP-Adapter, competitive diversity (0.164), and stronger biomimetic alignment (DSS: 0.627). Visual benchmarking suggested that outputs responded to prompt guidance, with Tafoni patterns emerging progressively as semantic complexity increased, indicating that designers may more readily anticipate and control outputs.

DISCUSSION

The metric profiles across configurations reveal distinctly different cognitive affordances for designers engaging in human–AI co-creation.

Quantitative performance vs. cognitive supportiveness. The results reveal a fundamental tension between metric-based performance and human-factors suitability. IP-Adapter achieved the strongest quantitative profile, yet its visual conditioning mechanism at w=0.8 operates independently of prompt intent—an inherent architectural characteristic confirmed by Figure 2. This demonstrates that quantitative superiority does not necessarily translate into cognitive supportiveness, and that distributional metrics alone are insufficient indicators of human-centred suitability.

Visual conditioning vs. semantic internalisation. IP-Adapter operates through direct visual conditioning, transferring reference image features regardless of prompt guidance at practical adaptation strengths. Fine-tuned

LoRA internalises domain-specific logics and responds progressively to prompt specification. This distinction has critical implications: Fine-tuned LoRA enables designers to anticipate and control outputs through prompt intent, reducing interpretive burden. Furthermore, prompt responsiveness creates the conditions for emergent diversity driven by designer input rather than model-imposed variation—arguably the more desirable mode of human-AI co-creation, in which the designer remains the primary creative agent. This aligns with Cognitive Load Theory, which suggests that reduced processing demands free cognitive resources for other demanding cognitive activities, including those involved in creative reasoning (Sweller, 1988).

IP-Adapter requires designers to simultaneously manage multiple interacting parameters—weight, prompt importance, and reference image selection—each influencing the balance between visual conditioning and prompt responsiveness. While prompt-responsive modes exist, achieving consistent prompt alignment in practice imposes a learning curve that Fine-tuned LoRA, having embedded domain intent at the training stage, does not require. Furthermore, the training phase itself introduces a new dimension of human-AI co-creation: the selection and curation of training data becomes a design act, embedding the designer's domain knowledge before any generation occurs.

Practical implications for workflow design. Fine-tuned LoRA is better suited to workflows where predictability, control, and iterative refinement are prioritised. IP-Adapter may be more appropriate for exploratory phases where visual richness and morphological variation are prioritised over semantic control. Neither configuration is universally superior; their cognitive profiles reflect different creative affordances that designers should consciously select based on workflow stage and intent.

Baseline and Shared LoRA demonstrated insufficient adaptation depth to support productive workflows. These configurations are likely to increase designer effort: without sufficient adaptation depth, designers must invest corrective iteration without receiving meaningful creative return.

Limitations and future directions. This study infers cognitive load implications from metric profiles and visual observation rather than direct user measurement. Future work should incorporate user studies with practising architects employing standardised cognitive load measures such as NASA-TLX to validate and extend the framework. The present study also examined a single biomimetic reference domain; broader evaluation across diverse design contexts would strengthen generalisability.

CONCLUSION

This study investigated how four PEFT configurations shape the interpretive demands and designer effort in a biomimetic architectural morphogenesis task, using MorphEvalBench metrics as proxies for interpretability and designer control. Four implications emerge:

First, PEFT strategies should be evaluated not only for visual performance but for their cognitive affordances. Strong metric performance does not guarantee cognitive supportiveness, suggesting that distributional metrics alone are insufficient indicators of human-centred suitability.

Second, the distinction between visual conditioning and semantic internalisation has direct human-factors consequences. Achieving consistent prompt alignment with IP-Adapter in practice requires managing multiple interacting variables, imposing a significant learning curve. Fine-tuned LoRA enables designers to anticipate and direct outputs through prompt specification alone, supporting lower cognitive load and more efficient iterative workflows.

Third, prompt responsiveness enables emergent diversity driven by designer input rather than model-imposed variation, positioning Fine-tuned LoRA as supporting a more desirable mode of human-AI co-creation in which the designer remains the primary creative agent.

Fourth, fine-tuning introduces a new temporal dimension of designer agency: the curation of training data itself becomes a design act, embedding intent into the model before generation begins. Designing the model and designing with the model are inseparable processes.

Together, these findings propose a pathway toward generative AI workflows that are not only visually powerful but cognitively supportive, interpretable, and aligned with the situated creative agency of the designer.

ACKNOWLEDGMENT

The authors would like to thank Professor Xu Weiguo for his supervision. Special thanks to Trapoom Ukarapol at Tsinghua University for developing the initial version of the evaluation framework that informed MorphEvalBench.

REFERENCES

Brama, H., Dalach, A., Grinshpoun, T. and Dortheimer, J. (2024) 'Towards a Robust Evaluation Framework for Generative Urban Design', eCAADe 2024, pp. 529–538.

Fuyuki, C. (2025) ArchMorphSD: A custom ComfyUI workflow for architectural morphogenesis [Computer software]. GitHub. Available at: https://github.com/cfuyuki/ArchMorphSD

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2021) 'LoRA: Low-Rank Adaptation of Large Language Models', arXiv:2106.09685.

Nagao, S. (2025) MorphEvalBench: Evaluation code for ArchMorphSD [Computer software]. GitHub. Available at: https://github.com/cfuyuki/MorphEvalBench

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J. and Rombach, R. (2023) 'SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis', arXiv:2307.01952.

Salma, Z., Hijón-Neira, R. and Pizarro, C. (2025) 'Designing Co-Creative Systems: Five Paradoxes in Human–AI Collaboration', Information, 16, 909.

Sweller, J. (1988) 'Cognitive load during problem solving: Effects on learning', Cognitive Science, 12(2), pp. 257–285.

Ye, H., Zhang, J., Liu, S., Han, X. and Yang, W. (2023) 'IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models', arXiv:2308.06721.

Zhang, L., Rao, A. and Agrawala, M. (2023) 'Adding Conditional Control to Text-to-Image Diffusion Models', arXiv:2302.05543.