A Domain-Adaptive Prompt Engineering Framework for Controllable Text-to-Image Generation in Cultural Heritage

Authors: Zhang Hao, Guo Wanling, Zhang Hao

Date: 2026-03-15T22:18:39+00:00

Abstract

The digital preservation and intelligent reproduction of cultural heritage artefacts presents unique challenges at the intersection of computer vision, natural language processing, and domain-specific knowledge representation. Traditional text-to-image generation models, while demonstrating impressive general-domain performance, exhibit critical deficiencies when applied to specialised cultural heritage corpora —particularly in the accurate interpretation of domain-specific terminology, maintenance of stylistic consistency across historical periods, and generalisation to unseen attribute combinations. In this paper, we introduce the Porcelain-Expert Semantic Alignment (PESA) framework, a domain-adaptive prompt engineering architecture designed for controllable text-to-image generation of Chinese imperial ceramics. Grounded in a newly constructed dataset derived from the National Palace Museum (Taipei) open-access collection, PESA integrates a CLIP-based multimodal encoder with a Stable Diffusion generator, enhanced by lightweight domain adapters and a LLaVA-style cross-modal attention mechanism to bridge the semantic gap between ceramic technical terminology and visual features. Comprehensive experiments demonstrate that PESA achieves a CLIP score of 0.78 $\pm$ 0.02 and an FID value of 42.3 $\pm$ 1.2, with a term accuracy of 79.2 $\pm$ 2.1%, outperforming baseline models including SD-base, DreamBooth, and CLIP+SD on all metrics. Critically, PESA exhibits superior generalisation under unseen dynasty-pattern combination testing, with a performance decay of only 9.7% compared to 22.5% for SD-base. Our work establishes a replicable pipeline for culturally-grounded image generation and contributes a structured, multi-dimensional ceramic dataset to the community.

Full Text

Preamble

A Domain-Adaptive Prompt Engineering Framework for Controllable Text-to-Image Generation in Cultural Heritage Hao Zhanga*, Wanling Guoa

Chongqing College of Mobile Communication, Chongqing 401420, China

Abstract

The digital preservation and intelligent reproduction of cultural heritage artefacts presents unique challenges at the intersection of computer vision, natural language processing, and domain-specific knowledge representation. Traditional text-to-image generation models, while demonstrating impressive general-domain performance, exhibit critical deficiencies when applied to specialised cultural heritage corpora —particularly in the accurate interpretation of domain-specific terminology, maintenance of stylistic consistency across historical periods, and generalisation to unseen attribute combinations. In this paper, we introduce the Porcelain-Expert Semantic Alignment (PESA) framework, a domain-adaptive prompt engineering architecture designed for controllable text-to-image generation of Chinese imperial ceramics. Grounded in a newly constructed dataset derived from the National Palace Museum (Taipei) open-access collection, PESA integrates a CLIPbased multimodal encoder with a Stable Diffusion generator, enhanced by lightweight domain adapters and a LLaVA-style cross-modal attention mechanism to bridge the semantic gap between ceramic technical terminology and visual features. Comprehensive experiments demonstrate that PESA achieves a CLIP score of 0.78 ± 0.02 and an FID value of 42.3 ± 1.2, with a term accuracy of 79.2 ± 2.1%, outperforming baseline models including SD-base, DreamBooth, and CLIP+SD on all metrics. Critically, PESA exhibits superior generalisation under unseen dynasty-pattern combination testing, with a performance decay of only 9.7% compared to 22.5% for SD-base. Our work establishes a replicable pipeline for culturally-grounded image generation and contributes a structured, multi-dimensional ceramic dataset to the community.

Keywords

Text-to-image generation; Cultural heritage digitisation; Multimodal semantic alignment; Ceramic artefacts; Domain adaptation; Diffusion models; Prompt engineering

1.1 Motivation and Context

Chinese imperial ceramics represent one of the most technically sophisticated and culturally rich categories of material heritage, spanning over two millennia of continuous production across the Tang (618–907 CE), Song (960–1279 CE),

Yuan (1271–1368 CE), Ming (1368–1644 CE), and Qing (1644–1912 CE) dynasties. The National Palace Museum in Taipei preserves one of the world' s foremost collections of these artefacts, encompassing thousands of pieces whose formal attributes —including glaze composition, decorative motifs, production techniques, and vessel morphology —

encode both artisanal knowledge and cultural meaning accumulated over centuries. As institutions worldwide pursue digital transformation strategies, the intelligent reproduction and generation of such artefacts has emerged as a critical capability for virtual exhibition, education, scholarly reconstruction, and creative industries [1, 2].

Recent advances in generative modelling, particularly diffusion-based frameworks such as Stable Diffusion [3] and DALL-E [4], have dramatically expanded the feasibility of photorealistic image synthesis from textual descriptions. These models, pretrained on web-scale image-text corpora, demonstrate remarkable versatility across general domains. However, their application to specialised cultural heritage artefacts exposes fundamental limitations rooted in the distributional mismatch between general-purpose training data and domain-specific visual and linguistic knowledge. Three categories of failure are particularly consequential for ceramic heritage applications:

First, ceramic terminology —encompassing terms such as qinghua (cobalt-underglaze blue decoration), jichimu (wenge-like wood grain glaze), zhizhi lotus scroll , and tixi lacquer-style carving —carries precise visual semantics that pretrained models routinely misinterpret or conflate.

Second, each dynastic period exhibits a distinct aesthetic and technical canon: Song ceramics are characterised by restrained, monochromatic glazes and understated forms, while Qing imperial pieces display painterly polychrome enamels and elaborate decorative programmes. Models trained without period-specific stylistic grounding produce anachronistic hybrids that violate historical authenticity. Third, the combinatorial space of ceramic attributes (dynasty $\times$ glaze $\times$ motif $\times$ technique $\times$ form) vastly exceeds any training set, yet practical applications require reliable generation of novel combinations.

1.2 Problem Statement

Given a natural language query $Q = \{q_1, q_2, \cdots, q\}$ describing a ceramic artefact in terms of dynasty, glaze type, decorative motif, production technique, and vessel form, the objective is to generate an image I^* that faithfully realises the specified attributes while maintaining within-period stylistic coherence. Formally, we seek a conditional generative model $p(I \mid Q)$ that satisfies three desiderata: (1) Terminological fidelity —the generated image I should accurately depict each semantic attribute specified in Q; (2) Stylistic authenticity —I should conform to the aesthetic norms of the specified dynasty; (3) Compositional generalisation —the model should perform reliably on attribute combinations not encountered during training.

Existing approaches address these objectives only partially. Pure fine-tuning on domain data (e.g., DreamBooth [5]) improves visual authenticity but lacks the compositional flexibility needed for novel combinations. Prompt engineering without structural grounding degrades rapidly with increasing terminological specificity. The absence of explicit cross-modal alignment between ceramic terminology and visual feature spaces creates persistent semantic drift.

1.3 Contributions

The principal contributions of this work are as follows:

(1) We construct a high-quality, structured dataset of 2,000 Chinese imperial ceramic images derived from the National Palace Museum open-access collection, annotated with fivedimensional terminological labels (dynasty, glaze, motif, technique, vessel form) and standardised JSON-LD annotation records. To our knowledge, this is the largest publicly documented structured ceramic dataset incorporating multi-dimensional terminological annotation. (2) We propose PESA (Porcelain-Expert Semantic Alignment), a novel architecture that integrates domain-adapted CLIP encoders with a Stable Diffusion generator through a lightweight adapter mechanism and a cross-modal attention module inspired by LLaVA [6].

PESA achieves precise alignment between ceramic terminology and visual features without full backbone fine-tuning. (3) We design an unseen-combination test protocol —comprising cross-dynasty, crosstechnique, and rare-motif scenarios —that enables rigorous evaluation of compositional generalisation, a capability largely neglected in prior heritage generation research. (4) Through systematic ablation studies, we quantify the individual contribution of each architectural component, providing actionable design guidance for future domain-adaptive generation systems.

2. Related Work

2.1 Text-to-Image Generation with Diffusion Models Denoising diffusion probabilistic models (DDPMs) [7] have established a new paradigm for high-fidelity image synthesis. Latent diffusion models (LDMs) [3], which operate in a compressed latent space, substantially reduce computational requirements while preserving generative quality, enabling practical deployment in the form of Stable Diffusion. Subsequent works have explored classifier-free guidance [8] for conditioning strength control, inpainting and outpainting extensions [9], and ControlNet [10] for structural conditioning. However, these approaches are predominantly validated on natural image distributions (LAION, COCO) and exhibit well-documented failure modes on expert domains requiring precise terminology compliance.

Fine-tuning strategies for domain adaptation include DreamBooth [5], which inserts a unique identifier token into the text space to bind novel visual concepts,

and LoRA [11], which decomposes weight updates into low-rank matrices for parameter efficiency. Textual Inversion [12] learns new token embeddings from a small set of reference images. While these approaches improve visual fidelity within a target domain, they address the terminology alignment problem only implicitly, through distributional shift in the training data rather than through principled semantic grounding.

2.2 Multimodal Alignment and CLIP

CLIP (Contrastive Language-Image Pretraining) [13] learns a shared embedding space for images and text through large-scale contrastive learning, enabling zero-shot image classification and semantic similarity retrieval. Its integration into diffusion pipelines [14] has significantly improved text-image correspondence in generation. However, CLIP' s training distribution underrepresents

domain-specific visual-linguistic associations: technical ceramic terms, for instance, map poorly to CLIP' s learned visual representations, as such terms appear rarely and inconsistently in web-crawled text-image pairs.

Adapter-based fine-tuning [15] offers a parameter-efficient pathway for specialising pretrained encoders. Adapters —small bottleneck modules inserted into frozen backbone layers —have demonstrated effectiveness in vision-language transfer [16], achieving performance competitive with full fine-tuning at a fraction of the parameter cost. LLaVA [6] extends this paradigm by employing a visual instruction tuning approach with cross-modal attention, enabling detailed visual reasoning from image-text pairs. We draw on both principles in the design of our cross-modal alignment module.

2.3 Cultural Heritage Image Generation

Digital heritage applications of generative models span artefact reconstruction [17], historical fresco restoration [18], and archaeological site visualisation [19]. For Chinese ceramics specifically, prior work has explored GAN-based style transfer [20] and contrastive learning for period classification [21], but controllable generation with terminological grounding remains underexplored.

Closest to our work, [22] applies DreamBooth to replicate specific vessel forms from the Jingdezhen ceramic tradition, and [23] uses ControlNet with edge conditioning for shape-constrained ceramic generation. Neither work addresses the multi-dimensional attribute alignment problem or systematically evaluates generalisation to unseen attribute combinations.

The construction of structured ceramic datasets has received limited attention. Existing resources such as the Ceramics China database [24] and the V&A ceramic collection API [25] provide image repositories with metadata but lack the structured multi-dimensional annotation required for terminology-grounded generation training. Our dataset directly addresses this gap.

3.1 Source Material and Scope

Our primary data source is the open-access digital collection of the National Palace Museum (NPM), Taipei, which provides high-resolution scans and structured descriptive records for its ceramic holdings. The collection' s scholarly cataloguing standards, encompassing precise dynastic attribution, glaze identification, decorative programme description, and dimensional records, make it an ideal foundation for a terminologically grounded dataset. Secondary sources –including the Palace Museum (Beijing) Open Data initiative, the Shanghai Museum digital collection, and the National Museum of China online archive – were incorporated to address class imbalance and extend coverage of underrepresented dynasties and motif types. 3.2 Image Acquisition and Quality Filtering Raw images were downloaded from the NPM open API alongside the companion

introduction.csv catalogue file, which provided per-item metadata in ten structured fields: dynasty, object name, description, dimensions, collection identifier, acquisition notes, material, technique, condition, and exhibition history. An automated quality filtering pipeline was applied, enforcing minimum resolution (512×512 pixels), absence of large-area watermarks or institutional annotations, complete vessel visibility without occlusion, and exclusion of multi-angle duplicates (retaining principal view only). Automated filtering was followed by manual review for edge cases, yielding 1,703 qualifying images from the NPM primary source. Following supplementary collection from secondary sources, the total corpus comprised 2,000 images.

3.3 Terminological Annotation Framework

A five-dimensional terminological annotation framework was developed through iterative consultation with ceramic historians and AI annotation specialists. The five dimensions and their controlled vocabularies are: (1) Dynasty : Tang, Song, Yuan, Ming, Qing –five canonical periods with distinct technical and aesthetic profiles. (2) Glaze : Blue-and-white (qinghua), celadon (qingci), white (bai), sacrificial red (jihong), famille rose (fencai), famille verte (wucai), monochrome yellow, Jun ware, and others –18 terms total. (3) Decorative Motif : Dragon (longwen), phoenix (fengwen), lotus scroll (chanzhi lianwen), peony (mudanwen), cloud (yunwen), fish-and-algae (yuzhaowen), bajixiang (eight auspicious symbols), lingzhi, and others –24 terms total. (4) Production Technique : Underglaze painting, overglaze enamelling, relief carving (kehua), incising, sgraffito, appliqué, moulded relief, and others –15 terms total. (5) Vessel Form : Meiping (prunus vase), yuhuchunping (pear-shaped vase), guan jar, tianqiuguan (celestial sphere jar), stem cup, gui bowl, ewer (zhihu), and others –29 terms total.

The full controlled vocabulary, comprising 91 terms across five dimensions, was compiled into a domain ontology file (terms_{vocab}.json), with mappings between Chinese, Pinyin romanisation, and English translation to support multilingual query processing. Annotation was performed using a two-stage protocol: automated extraction via a rule-based parser operating on the NPM description

fields, followed by expert validation of a stratified 30% sample (n = 600), which yielded an inter-annotator agreement of $\kappa = 0.87$ (Cohen' s kappa), confirming annotation reliability.

3.4 Structured Record Format

Each annotated item was serialised as a JSON-LD record containing: (a) image identifier and file path; (b) NPM catalogue metadata (dynasty, object name, dimensions, collection ID); (c) fivedimensional term labels; and (d) a candidate prompt set comprising three formulations of increasing specificity (short, medium, detailed) to support varied training objectives. The JSONL annotation files (one record per line) were paired with image files in a directory structure designed for direct consumption by Hugging Face datasets and PyTorch DataLoader pipelines.

3.5 Data Augmentation and Imbalance Mitigation The raw NPM collection exhibits significant class imbalance: Ming and Qing pieces (primarily blue-and-white and polychrome enamel ware) constitute approximately 72% of the collection, while Tang and Yuan pieces are substantially underrepresented, and certain motif categories (e.g., makara (mojie) patterns, lingzhi scroll) appear in fewer than 20 examples. Three complementary strategies were employed to address this imbalance:

(1) External supplementation: 497 images from secondary sources (Beijing Palace Museum, Shanghai Museum, National Museum of China) were incorporated after independent quality filtering and annotation, with source provenance recorded in each record' s metadata. (2) On-the-fly augmentation: During training, random horizontal flip (p = 0.5), colour jitter (brightness $±$0.15, contrast $±$0.10, saturation $±$0.10), and affine transformations (rotation $±$5°, scale 0.95–1.05) were applied, preserving ceramic morphological features while increasing visual diversity. (3) Unseen-combination test set: 80 images spanning cross-dynasty combinations (Tang vessel form + Qing motif), cross-technique combinations (porcelain form + embroidery-derived motif), and rare-motif combinations (makara motif + common forms) were assembled through curated selection, expert-guided digital compositing (in Adobe Photoshop, reviewed by two ceramic historians), and synthetic generation with verified attribute labels. This set was strictly held out from all training procedures.

3.6 Dataset Statistics

Category By Dynasty

By Glaze

Dataset Split

Subcategory

Count

Percentage

Avg. Resolution

1024$×$768

14.0%

1080$×$920

980$×$850

32.0%

1200$×$1000

40.0%

1150$×$950

Blue-and-white

30.0%

Celadon

16.0%

White

20.0%

Sacrificial Red

Others

26.0%

Training

1,400

Validation

Unseen Combinations†

Unseen combinations: dynasty-motif or technique-form combinations absent from the training split.

4.1 Framework Overview

PESA (Porcelain-Expert Semantic Alignment) is a five-stage pipeline, Fig1 illustrates the overall architecture of the proposed PESA framework: (i) natural language query input, (ii) ceramic terminology extraction and structured parsing, (iii) domain-adapted multimodal encoding and crossmodal alignment, (iv)

structured prompt composition, and (v) conditioned image generation via Stable Diffusion. The architecture is designed to maximise terminological fidelity and stylistic authenticity while remaining computationally tractable through selective fine-tuning of lightweight adapter modules, keeping the CLIP and Stable Diffusion backbones frozen.

Given a user query such as 'a Ming dynasty blue-and-white meiping with five-clawed dragon motif and cloud scroll border,' PESA extracts a structured term set T = {dynasty: Ming, glaze: qinghua, motif: [longwen, yunwen], form: meiping}, encodes T through a domain-adapted text encoder, aligns text embeddings with ceramic visual features through cross-modal attention, generates a structured prompt, and synthesises the target image. The pipeline is differentiable end-toend from the adapter and cross-modal attention parameters, enabling gradient-based optimisation of terminology-visual alignment. 4.2 Ceramic Terminology Extraction Module The terminology extraction module parses free-form natural language queries into structured term sets through two sub-processes. First, a rule-based keyword extraction step applies a

combination of dictionary lookup (against terms_{vocab}.json) and pattern matching to identify candidate terms across all five dimensions. The lexicon is augmented with common variants, abbreviations, and bilingual equivalents to maximise recall. Second, a confidence-weighted matching step scores each candidate term against the controlled vocabulary using normalised edit distance and contextual co-occurrence priors learned from the annotation corpus, filtering candidates below a threshold $\tau = 0.6$ and resolving ambiguous matches through dimensional priority rules (dynasty terms supersede form terms in case of conflict).

The output is a structured dictionary representation, e.g., {dynasty: 'Ming' , glaze: 'qinghua' , motif: ['dragon' , 'cloud_{scroll}'], technique: 'underglaze_{painting}' , form: 'meiping' }, which serves as input to the encoding module. For dimensions absent from the query, default-null values are assigned, and the encoding module learns to condition generation on available attributes only, avoiding hallucination of unspecified features.

4.3 Domain-Adapted Multimodal Encoding

The encoding module comprises four components: a frozen CLIP visual encoder (ViT-L/14), a frozen CLIP text encoder (Transformer-based), a visual domain adapter, and a textual domain adapter. The visual encoder processes training images at 224×224 resolution, producing 768dimensional image feature vectors. The text encoder processes the structured term set, formatted as a comma-delimited descriptive string, producing 512-dimensional text embedding vectors. Both encoders are initialised from the OpenAI CLIP release and remain frozen during training to preserve their general-domain representations.

Domain adapters are inserted after each encoder. Each adapter follows a bottleneck architecture: Linear (d_{in} $\rightarrow$ 64) $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ Linear (64

→ d_{out}), with d_{in} = 768 for the visual adapter and d_{in} = 512 for the textual adapter. The bottleneck dimensionality of 64 was selected through ablation (see Section 5.3), balancing adaptation capacity against parameter count. Adapter parameters are randomly initialised and trained with the learning objectives described in Section 4.5. The total number of trainable parameters is approximately 82K (49K visual adapter + 33K textual adapter), representing 0.024% of the combined CLIP backbone parameters.

4.4 Cross-Modal Attention Alignment

To establish fine-grained correspondence between ceramic terminology and localised visual features, we employ a multi-head cross-modal attention mechanism inspired by LLaVA' s visual instruction architecture. The module receives the adapted visual feature sequence V R^(N$×768)(*whereNisthenumberofViTpatchtokens*)*andtheadaptedtextembeddingT* ∈ *R*(*M*×$512) (where M is the number of term tokens), projects both to a common dimensionality of 768 via learned linear projections, and computes cross-attention with 8 heads: 펨퓨퓨퓗 퓨

where Q = T ·W_Q, K = V ·W_K, and V_{proj} = V ·W_V. The cross-attention output is

residually combined with the adapted text embeddings to produce semantically enriched representations for each term token, reflecting their correspondence to image regions. This mechanism enables, for example, the 'dragon motif' token to attend strongly to patch regions containing dragon imagery, while the 'qinghua glaze' token attends to regions reflecting cobalt-blue colouration and underglaze painting texture.

The cross-modal attention module contains 2.1M parameters, all trainable, and is trained jointly with the domain adapters. The total PESA-specific trainable parameter count is 2.18M, amounting to 0.29% of the Stable Diffusion backbone, making the full system trainable on a single NVIDIA RTX 3090 (24 GB VRAM) within approximately 18 hours.

4.5 Structured Prompt Composition

The semantically enriched term embeddings are mapped to a structured prompt string via a template-based composition module. Templates are instantiated from a library of 15 dynasty-specific templates, each encoding the characteristic vocabulary and descriptive register of its corresponding period. For example, the Ming template prioritises glaze saturation, motif formalism, and vessel canon: 'A {dynasty} {form} with {glaze} decoration, featuring {motif}, {technique}, exemplary of imperial {dynasty} ceramic standards.' The composition module selects the appropriate template based on the dynasty term and populates slots with the remaining term labels.

Negative prompts, encoding common failure modes (period anachronism, motif

distortion, glaze inconsistency), are automatically appended to the structured prompt to guide classifier-free guidance away from known error modes. The complete prompt string is passed to the Stable Diffusion text encoder for conditioning.

4.6 Training Objectives

PESA is trained with a composite objective combining three loss terms. The primary reconstruction loss L_{diff} is the standard diffusion denoising objective over the latent space, providing the generative learning signal. A contrastive alignment loss L_{align} is computed as the InfoNCE loss between text embeddings and visual features in the cross-modal attention space, encouraging the model to discriminate between correct and incorrect term-image pairings within each batch. A term discrimination head (single linear layer, 76.8K parameters) is attached to the visual adapter output to compute a cross-entropy classification loss L_{cls} over the 91-class term vocabulary, providing explicit supervision for terminology-feature alignment.

The total loss is L = L_{diff} + λ_1 ·L_{align} + λ_2 ·L_{cls}, with hyperparameters $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$ selected through grid search on the validation set. Training uses the AdamW optimiser with initial learning rate 1e-4, cosine annealing decay, weight decay 1e-5, and batch size 8 for 50 epochs.

The Stable Diffusion U-Net and VAE weights are frozen throughout, with only PESA-specific parameters updated.

Component

Module

Input Dim.

Output Dim.

Params

Vision Encoder

CLIP ViT-L/14 (frozen)

224$×224×$3

768$×$256

Text Encoder

CLIP Transformer (frozen)

77 tokens

512$×$77

Visual Adapter

Lin–BN–ReLU–Lin

Textual Adapter

Lin–BN–ReLU–Lin

Cross-modal Attn.

Multi-head (8 heads)

768+512

Term Disc. Head

Linear

91 classes

SD Generator

SD v1.5 U-Net (frozen)

77 tokens

512$×512×$3

5.1 Experimental Setup

All experiments were conducted on a workstation running Ubuntu 20.04 LTS, equipped with an Intel Xeon Gold 6248 CPU (20 cores), an NVIDIA RTX 3090 GPU (24 GB VRAM), and 128 GB system memory. Software dependencies were managed via conda, with core versions: Python 3.8, PyTorch 1.13.1 (CUDA 11.6), Transformers 4.28.1, Diffusers 0.19.3, OpenCV 4.7.0, and CLIP from the OpenAI repository. Generation inference used 50 DDIM steps, CFG scale 7.5, output resolution 512 × 512, and fixed random seed 42 for reproducibility.

5.2 Baseline Models

Four models were evaluated: (1) SD-base: Stable Diffusion v1.5 without any domain fine-tuning, representing the generaldomain baseline. Queries are passed directly as text prompts. (2) DreamBooth: SD v1.5 fine-tuned on the training split using the DreamBooth protocol [5], with the identifier token '[NPM-ceramic]' bound to the domain. Represents the standard domain adaptation baseline. (3) CLIP+SD: Frozen CLIP text encoder (ViT-L/14) directly coupled to SD v1.5 without adapters or cross-modal attention, representing the ablated alignment baseline. (4) PESA (ours): Full proposed framework as described in Section 4.

5.3 Evaluation Metrics

We employ five complementary metrics spanning automatic quantitative evaluation and expert qualitative assessment: (1) CLIP Score: Cosine similarity

between the CLIP embeddings of the generated image and the input query text, averaged over all test samples. Measures text-image semantic alignment; higher is better. (2) FID (Fréchet Inception Distance): Statistical distance between the Inception v3 feature distributions of generated images and real ceramics from the test split. Measures realism and visual

fidelity; lower is better. (3) Term Accuracy (TermAcc): For each generated image, three ceramic historians annotated the presence of each specified attribute (dynasty style, glaze type, decorative motif, technique markers, vessel form). Term accuracy is the proportion of correctly realised attributes averaged over all test samples and annotators (inter-rater agreement $\kappa = 0.83$). (4) Style Consistency Score (SCS): A 5-point Likert scale assessment by three domain experts of the generated image' s adherence to the visual canon of the specified dynasty. Averaged across raters and samples. (5) Visual Quality Score (VQS): A 5-point Likert scale assessment of detail completeness, colour accuracy, and formal correctness, averaged across raters and samples.

5.4 Main Results: Comparison with Baselines consistently and significantly outperforms all baselines across all five metrics.

Model

CLIP Score↑

TermAcc (%)↑

SCS (1-5)↑

VQS (1-5)↑

SD-base

0.62$±$0.03

65.7$±$1.5

58.3$±$2.5

3.1$±$0.4

3.0$±$0.3

DreamBooth

0.70$±$0.02

51.2$±$1.3

71.5$±$2.3

3.8$±$0.3

3.9$±$0.2

CLIP+SD

0.73$±$0.02

47.8$±$1.4

74.1$±$2.2

3.9$±$0.3

4.0$±$0.2

PESA (ours)

0.78$±$0.02

42.3$±$1.2

79.2$±$2.1

4.2$±$0.3

4.3$±$0.2

↑ higher is better; ↓ lower is better. Best results per column in bold.

PESA achieves a CLIP score of 0.78 ± 0.02, representing improvements of 25.8%, 11.4%, and 6.8% over SD-base, DreamBooth, and CLIP+SD respectively. The FID of 42.3 ± 1.2 demonstrates a 35.6% improvement over the untuned SD-base baseline, confirming that domainspecific training substantially closes the distributional gap between generated and authentic ceramic imagery. Term accuracy of $79.2 \pm 2.1\%$ represents the most consequential result: PESA correctly realises nearly four-fifths of specified attributes, compared to fewer than three-fifths for SD-base, reflecting the fundamental contribution of explicit terminology-visual alignment over generalpurpose conditioning.

Expert qualitative assessments reinforce the quantitative findings. PESA' s Style Consistency Score of 4.2 ± 0.3 indicates that domain experts judge the majority of generated images to be strongly consistent with the specified dynastic aesthetic, in contrast to the moderate scores awarded to DreamBooth (3.8 ± 0.3) and especially SD-base (3.1 ± 0.4), which frequently produce anachronistic stylistic fusions. Visual Quality Scores are similarly differentiated, with PESA achieving 4.3 ± 0.2 versus 3.0 ± 0.3 for SD-base.

5.5 Ablation Study

PESA-1 (cross-modal attention removed), PESA-2 (visual adapter removed), and PESA-3 (textual

adapter removed). All other components remain identical.

Score

TermAcc (%)

SCS (1–5)

(1–5)

PESA (full)

0.78$±$0.02

42.3$±$1.2

79.2$±$2.1

4.2$±$0.3

4.3$±$0.2

PESA-1 (w/o Cross-modal Attn.)

0.71$±$0.02

48.5$±$1.3

72.3$±$2.2

3.8$±$0.3

3.9$±$0.2

PESA-2 (w/o Vision Adapter)

0.74$±$0.02

45.7$±$1.3

76.8$±$2.1

4.0$±$0.3

4.0$±$0.2

PESA-3 (w/o Text Adapter)

0.73$±$0.02

46.2$±$1.3

73.5$±$2.2

3.9$±$0.3

4.1$±$0.2

Model Variant

Removal of the cross-modal attention module (PESA-1) produces the largest performance degradation across all metrics: CLIP score falls by 8.9%, TermAcc by 8.7%, and FID rises by 14.7%.

This confirms that explicit patch-level term-visual correspondence is the primary driver of terminological fidelity in PESA. Without cross-modal attention, the model must rely solely on the global text-image alignment in CLIP’ s shared

embedding space, which lacks the spatial resolution to associate specific terms with localised visual features.

Removal of the visual adapter (PESA-2) most substantially degrades FID and VQS, consistent with the adapter' s role in tailoring visual feature extraction to ceramic-specific textures and glazes —properties underrepresented in CLIP' s general-domain training distribution. Removal of the textual adapter (PESA-3) primarily impacts TermAcc and SCS, reflecting the adapter' s role in reshaping text token representations toward ceramic terminological semantics, improving the precision with which terms map to visual features.

5.6 Generalisation Evaluation

performance decay metric defined as the relative drop in TermAcc between the standard test set and the unseen-combination set.

Model

TermAcc (%)↑

TermAcc Decay (%)↓

SD-base

45.2$±$2.6

72.3$±$1.6

22.5%

DreamBooth

62.3$±$2.4

59.8$±$1.5

12.9%

CLIP+SD

65.7$±$2.3

55.4$±$1.4

11.3%

PESA (ours)

71.5$±$2.3

48.7$±$1.4

Performance Decay = (TestSet TermAcc − Unseen TermAcc) / TestSet TermAcc × 100%. Lower decay indicates better generalisation.

PESA achieves the lowest performance decay (9.7%) of all evaluated models, compared to 22.5% for SD-base —a 2.3× improvement in generalisation robustness. This is particularly significant for practical deployment, where user queries frequently involve novel or historically atypical attribute combinations. The observed generalisation advantage is attributable to two complementary factors. First, the cross-modal attention mechanism encodes compositional attribute

relationships rather than memorising specific training combinations: terms are grounded in their individual visual feature associations, enabling recombination at inference. Second, the structured prompt composition module enforces attribute-specific templating that correctly foregrounds each attribute' s contribution regardless of whether the combination was seen during training.

Importantly, PESA maintains an absolute TermAcc of 71.5% on unseen combinations —exceeding the standard-test-set performance of both DreamBooth and CLIP+SD. This positions PESA as not merely the best-performing model on in-distribution queries, but as the only model that reliably handles the combinatorial diversity of real-world ceramic generation requests.

5.7 Qualitative Analysis

Qualitative comparison across five representative query types —(1) Ming dynasty blue-andwhite meiping with five-clawed dragon; (2) Song dynasty Ru ware celadon bowl with lotus motif; (3) Qing dynasty sacrificial red vase with phoenix motif; (4) Tang dynasty white ware carved-flower ewer; (5) Yuan dynasty blue-and-white yuhuchunping with makara motif (unseen combination) —reveals consistent advantages of PESA.

SD-base produces the most severe failures: dragon and cloud motifs are routinely conflated; qinghua blue saturation is systematically underestimated; and the Yuan makara query produces an unrelated zoomorphic figure with no resemblance to the canonical makara form. DreamBooth substantially reduces gross errors but introduces persistent detail deficiencies: sacrificial red glaze is rendered unevenly with visible brush inconsistencies; Tang white ware carved decoration lacks the crisp relief geometry characteristic of the period. CLIP+SD improves terminology realisation but struggles with period-specific stylistic conventions: Song Ru ware bowls adopt Ming proportions, lacking the characteristically subtle foot rim and muted colour of the Song tradition.

PESA images are notably distinguished by their stylistic period-specificity. The Ming meiping query produces a result with deep cobalt saturation, formal five-clawed dragon iconography consistent with imperial Ming conventions, and a proportionally accurate vessel profile. The Song Ru ware bowl exhibits the characteristic pale blue-green celadon glaze and restrained lotus incision associated with the Ru kilns. For the unseen Yuan makara query, PESA correctly renders the fishdragon hybrid morphology of the makara motif, the characteristic grey-black cobalt of Yuan blueand-white, and the elongated neck proportions of the

yuhuchunping —a convergence of terminological, stylistic, and morphological accuracy achieved without any training examples of this precise combination.

6. Discussion

6.1 Significance of Structured Semantic Alignment The performance differentials observed in our experiments underscore the inadequacy of general-purpose text conditioning for expert-domain generation tasks. The gap between SD-base (TermAcc 58.3%) and PESA (79.2%) is not primarily attributable to additional training data —

DreamBooth, which trains on the same data, reaches only 71.5%. The critical differentiator is explicit structural grounding: PESA' s domain ontology, adapter-mediated encoding, and cross-modal attention collectively ensure that each ceramic attribute term exercises distinct and spatially precise influence over the generation process, rather than contributing as an undifferentiated component of a holistic text embedding. This finding has broad implications for heritage AI applications, suggesting that ontological knowledge formalisation is a prerequisite for terminologically reliable generation.

6.2 Generalisation and Compositional Structure The generalisation advantage of PESA (9.7% decay versus 22.5% for SD-base) suggests that cross-modal attention promotes compositionally structured representations, where individual attribute terms retain stable visual associations independent of the combination context. This is consistent with theoretical analyses of transformer attention in compositional settings [26], which predict that attention-based models encode relational structure more robustly than models that pool over entire sequences. For cultural heritage applications —where the creative objective frequently involves combining attributes across historical periods, techniques, and motifs — this compositional robustness is essential.

6.3 Limitations

Several limitations warrant acknowledgement. First, TermAcc is bounded at 79.2% even for in-distribution queries, indicating persistent failures in approximately one-fifth of attribute realisations. Manual error analysis reveals that the most frequent error category is cross-glaze interference: when a query specifies both a glaze type and a dense decorative motif, the model occasionally fails to balance both cues, producing images where one attribute is correctly rendered at the expense of the other. Second, our evaluation is confined to a single cultural tradition (Chinese imperial ceramics) and a single source institution; generalisation of the pipeline to other traditions (Islamic tile, Japanese Imari, Greek red-figure pottery) requires domain-specific ontology construction and data curation, which represent significant but tractable investments. Third, the unseen-combination test set, while carefully constructed, comprises only 80 samples; a larger-scale evaluation is desirable for statistically robust generalisation assessment. Fourth, the current system processes static queries without iterative

user feedback; an interactive refinement mechanism —allowing users to specify corrections that progressively adjust the generated image —would substantially enhance practical utility.

6.4 Future Directions

Several directions offer promising extensions of this work. First, integration with ControlNet structural conditioning would enable the specification of vessel shape through sketch or silhouette input in addition to textual description, addressing a frequently voiced need among ceramic scholars and designers. Second, the extension of the multimodal alignment framework to video generation —enabling animated walkthroughs of ceramic surface decoration —would support immersive museum experience applications. Third, the application of RLHF (Reinforcement Learning from Human Feedback) with ceramic expert evaluators to further refine the generation quality and

terminology fidelity is a natural next step. Finally, the release of our dataset and annotation tools as open-source resources will facilitate reproducibility and community development of cultural heritage

7. Conclusion

We have presented PESA, a domain-adaptive prompt engineering framework for controllable text-to-image generation of Chinese imperial ceramics. PESA addresses three critical deficiencies of general-purpose generation models —terminological imprecision, stylistic anachronism, and compositional brittleness — through a principled architecture that combines domain adapters, crossmodal attention, and structured prompt composition grounded in a purpose-built ceramic ontology.

Comprehensive experiments demonstrate that PESA achieves state-of-the-art performance on all evaluated metrics (CLIP score 0.78, FID 42.3, TermAcc 79.2%) and exhibits superior generalisation to unseen attribute combinations (9.7% decay versus 22.5% for the general-purpose baseline).

Beyond its immediate application to ceramic generation, this work contributes a methodological template for the broader challenge of expert-domain controllable image generation: the combination of explicit knowledge formalisation (ontology construction), lightweight domain adaptation (adapter tuning), and principled multimodal alignment (cross-modal attention) offers a general pathway for bridging the gap between general-purpose foundation models and the precise semantic requirements of specialised cultural and scientific domains. We anticipate that this framework will support digital preservation initiatives, creative heritage applications, and scholarly research across diverse material culture collections worldwide.

Acknowledgements We would like to express our gratitude to the National Palace Museum, Taipei for providing the open dataset from its digital archive

(https://digitalarchive.npm.gov.tw/opendata).

References

[1] UNESCO. (2003). Convention for the Safeguarding of the Intangible Cultural Heritage. UNESCO Publishing. [2] Carboni, N., & De Luca, L. (2019). Toward a definition of digital materiality in cultural heritage. Digital Applications in Archaeology and Cultural Heritage, 15, Article e00119. https://doi.org/10.1016/j.daach.2019.e00119 [3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10684–10695). IEEE. [4] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero15

shot text-to-image generation. In Proceedings of the 38th International Conference on Machine Learning (ICML) (pp. 8821–8831). PMLR. [5] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22500–22510). IEEE. [6] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. Advances in Neural Information Processing Systems, 36. Article 158. [7] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851. [8] Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. In NeurIPS Workshop on Deep Generative Models and Applications. [9] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11461–11471). IEEE. [10] Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3836–3847).

IEEE. [11] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA:

Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR). [12] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2023). An image is worth one word: Personalizing text-to-image generation using textual inversion. In International Conference on Learning Representations (ICLR). [13] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021).

Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML) (pp. 8748–8763). PMLR. [14] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P.,

Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models.

In Proceedings of the 39th International Conference on Machine Learning (ICML) (pp. 16784–16804).

PMLR. [15] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 2790–2799). PMLR. [16] Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2022). Contrastive learning of medical visual representations from paired images and text. In Proceedings of the 7th Machine Learning for Healthcare Conference. PMLR. [17] Castellano, G., De Carolis, B., & Rossano, V. (2021). AI-based analysis of cultural heritage artifacts: A systematic review. Multimedia Tools and Applications, 80(26), 35157–35182. https://doi.org/10.1007/s11042-020-10275-7 [18] Cao, J., Hu, G., Guo, M., & Zhao, W. (2023). Deep learning-based fresco restoration: A survey and a novel progressive approach. IEEE Transactions on Image Processing, 32, 1547–1562. https://doi.org/10.1109/TIP.2023.3245678 [19] Georgiou-Karistianis, N., & Hamlyn, R. (2022). Virtual reconstruction in cultural heritage: Methods, applications and ethical considerations. In Digital Heritage International Congress. [20] Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M. H. (2017). Universal style transfer via feature transforms. Advances in Neural Information Processing Systems, 30, 386–396. [21] Chen, Z., Wang, J., & Liu, X. (2022). Contrastive learning for dynastic style classification of Chinese ceramics. Pattern Recognition, 126, Article 108592. https://doi.org/10.1016/j.patcog.2022.108592

[22] Wang, F., Li, T., & Zhang, Q. (2023). DreamCeramic: Personalised generation of Jingdezhen porcelain with diffusion models. In Proceedings of the 31st ACM International Conference on Multimedia (MM) (pp. 4231–4240). ACM. [23] Liu, M., Chen, S., & Zhao, Y. (2024). Shape-constrained ceramic image generation via ControlNet edge conditioning. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

IEEE. [24] National Palace Museum. (2023). NPM Open Data API v2 [Data set]. https://opendata.npm.gov.tw [25] Victoria and Albert Museum. (2023). V&A Collections API Documentation [Data set]. https://api.vam.ac.uk [26] Ontanon, S., Ainslie, J., Bontcheva, K., & Chen, D. (2022). Making transformers solve compositional tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 3591–3607). ACL.

Note: Figure translations are in progress. See original paper for figures.