Knowledge-Aware Artifact Image Synthesis with LLM-Enhanced Prompting and Multi-Source Supervision

Shengguang Wu[†], **Zhenglun Chen**[‡*], **Qi Su**[†]

[†]Peking University, [‡]Cornell University

Abstract

Ancient artifacts are an important medium for cultural preservation and restoration. However, many physical copies of artifacts are either damaged or lost, leaving a blank space in archaeological and historical studies that calls for artifact image generation techniques. Despite the significant advancements in open-domain text-to-image synthesis, existing approaches fail to capture the important domain knowledge presented in the textual description, resulting in errors in recreated images such as incorrect shapes and patterns. In this paper, we propose a novel knowledge-aware artifact image synthesis approach that brings lost historical objects accurately into their visual forms. We use a pretrained diffusion model as backbone and introduce three key techniques to enhance the text-to-image generation framework: 1) we construct prompt with explicit archaeological knowledge elicited from large language models (LLMs); 2) we incorporate additional textual guidance to correlated historical expertise in a contrastive manner; 3) we introduce further visual-semantic constraints on edge and perceptual features that enable our model to learn more intricate visual details of the artifacts. Compared to existing approaches, our proposed model produces higher-quality artifact images that align better with the implicit details and historical knowledge contained within written documents, thus achieving significant improvements across automatic metrics and in human evaluation.

1 Introduction

Ancient artifacts are crucial for cultural preservation, as they represent tangible evidence of the past, offering insights into history. In recent years, innovative artifact-related projects have emerged, including the restoration of degraded character images (Shi et al. 2022), the generation of captions for ancient artwork (Sheng and Moens 2019), and the deciphering of oracle bone inscriptions (Chang et al. 2022). These works have opened up new avenues for researchers to study artifacts and gain insights into the past. Despite these advancements, there is still much to be explored in artifact-related tasks, one of which is to recreate visual images of artifacts from text descriptions as many physical copies of artifacts are often damaged or lost, leaving only textual records behind. This task could prove immensely invaluable to historical studies and cultural preservation, as it provides historians with new visual angles to study the past and enables people to connect with their cultural heritage.

One area that has shown potential to aid in the recreation of visual images of ancient artifacts is text-to-image synthesis. This task has been a popular area of research, especially in recent years with the introduction of diffusion models (Yang et al. 2022; Ho, Jain, and Abbeel 2020; Rombach et al. 2021; Song et al. 2021; Nichol and Dhariwal 2021; Song, Meng, and Ermon 2022) that have demonstrated significant capabilities in generating photorealistic images based on a given text prompt in open-domain problems (Nichol et al. 2022; Rombach et al. 2021; Saharia et al. 2022; Gu et al. 2022; Ramesh et al. 2022). However, in the specialized area of archaeological studies, where data is often limited and domain knowledge is required, vanilla diffusion models struggle to produce promising results even with finetuning, as shown in Figure 1. The generated images often display errors in shape, patterns, and details that fail to match the implicit knowledge in the textual information and the underlying historical context of the target artifact.

We have identified a key cause for this problem to be the lack of knowledge supervision during the generating process, which can be attributed to two main aspects. 1). Current text prompts may not be infused with domain-specific knowledge from the archaeological and historical fields, leading to noisiness and lack of well-presented knowledge information in the text prompt. 2). The text and visual modules in the vanilla diffusion models (Rombach et al. 2021; Sohl-Dickstein et al. 2015; Song et al. 2020; Ho, Jain, and Abbeel 2020; Song and Ermon 2020) may be unable to capture domain-specific knowledge under the standard training pipeline, resulting in the lack of detailed textual and visual signals of ancient artifacts in the generation process.

To address these challenges, we propose our knowledge-aware artifact image synthesis approach with a pretrained Chinese Stable Diffusion model (Rombach et al. 2021; Zhang et al. 2022a) as our backbone. Our method can generate visualizations of lost artifacts that well align with the underlying domain-knowledge presented in their textual records. Specifically: 1). To address the issue of noisiness and lack of well-presented knowledge information in the text prompt, we propose to use Large Language Models (LLMs) to enhance our text prompts in two ways: for one, we use LLMs to extract the core and meaningful information in the

[*]Work done during internship at Peking University.

given text prompt and reorganize them in a more structured way to explicitly present the current knowledge information; for another, we use LLMs as an external knowledge base to retrieve relevant archaeological knowledge information and augment them in the restructured text prompt. 2). To address the lack of both textual and visual knowledge supervision in the generation process, we introduce additional supervisions in both modalities. Firstly, we introduce a constrastive training paradigm that enables the text encoder to make the textual representation of the artifact more in line with their archaeological knowledge. Secondly, we apply stricter visual constraints using edge loss (Seif and Androutsos 2018) and perceptual loss (Johnson, Alahi, and Fei-Fei 2016) to make the final visual output align with the visual domain knowledge of ancient artifacts. Both quantitative experiments and user study demonstrate that our knowledge-aware artifact image synthesis approach significantly outperforms existing text-to-image models and greatly improves the generation quality of historical artifacts.

Our main contributions can be summarized as follows:

- We propose to use LLMs as both information extractor and external knowledge base to aid better prompt construction in a specialized domain, which in our case is archaeological studies.
- We introduce additional multimodal supervisions to enable our model to learn textual representations and visual features that better align with archaeological knowledge and historical context, thus improving the current fine-tuning paradigm of diffusion models.
- To our best knowledge, we are the first to explore text-to-image synthesis task in the archaeological domain as an attempt to recreate lost artifacts, and thus aiding archaeologists to gain deeper insights into history.

2 Related Work

Text-to-image Synthesis. Text-to-image synthesis tasks have long been a vital task at the intersection between computer vision and natural language processing, of which models are given a plain text description to generate the corresponding image. One major architecture in this area is GAN (Goodfellow et al. 2014), whose variations (Zhang et al. 2016; Xu et al. 2017; Kang et al. 2023) have resulted in the state-of-art performance of text-to-image synthesis tasks. Recently, diffusion models (Rombach et al. 2021; Sohl-Dickstein et al. 2015; Song et al. 2020; Ho, Jain, and Abbeel 2020; Song and Ermon 2020) also have demonstrated their ability to achieve new state-of-the-art results (Dhariwal and Nichol 2021). Diffusion models make use of two Markov chain: forward and reverse. The forward chain gradually add noise to the data with Gaussian prior. The reverse chain aims to denoise the data gradually. The transition probability at each timestep is learned by a deep neural network, which in the case of text-to-image synthesis is usually a U-Net (Ronneberger, Fischer, and Brox 2015) model.

Large Language Models. Language models are a family of probabilistic models that predicts the probability of the next word, given a sequence previous words within a context. The introduction of GPT-3 (Brown et al. 2020), which contains 175B parameters, has led to the emergence of Large Language Models (LLMs), referring to language models with a large number of parameters. These LLMs have demonstrated never-seen-before abilities, expanding the frontiers of what is possible with language models. One emerging ability of LLMs is in-context learning (Brown et al. 2020), where LLMs are able to perform downstream tasks after being prompted with just a few examples without further parameter update. Thus, by providing carefully designed examples, we can make use of LLMs as a information extractor given a noisy and unstructured text. LLMs have also shown their ability to acquire world knowledge from the massive training corpus (Petroni et al. 2019; Wang, Liu, and Zhang 2021; Liu et al. 2022). An efficient way to extract the implicit knowledge from LLMs is to ask questions with proper prompt engineering as LLMs are highly sensitive to the prompt input (Liang et al. 2022).

Figure 1: Images of artifacts generated by a vanilla diffusion model, the shape, color, pattern and material differ greatly from the ground truth.

3 Background

Problem Statement. As mentioned in Section 1, for many artifacts, text documents are the only available source of information. Hence our task is to recreate a visual image I_i' given artifact text information T_i. The synthesized image I_i' needs not only to align with the textual meanings conveyed in T_i but also to be in line with the implicit historical knowledge about the artifact. Only then will the generated image be historically correct and thus valuable to archaeological studies. Correspondingly, the training dataset D comes in the form of pairs as $D = \{(T_i, I_i)\}_i^n$, where $T_i \in T$ is the

available text information and $I_i \in I$ is the corresponding artifacts image. The raw text information available for an artifact - as often catalogued in museums - contains roughly four parts: the ***name*** or title of the artifact; the ***time period*** of origin; a raw ***description*** of the artifact (often presented in a messy way); the physical ***size*** of the artifact. Formulated from accessible resources of such kind, the task of our work is then to generate accurate artifact images based on these textual descriptions about historical objects.

Diffusion Preliminaries. We build our model upon the text-conditioned Stable Diffusion pipeline (Rombach et al. 2021) and we present a detailed mathematical introduction of diffusion models and Stable Diffusion in Appendix G. Here, we briefly summarize the standard training objective of a Stable Diffusion model as follows:

$$L_{SD}(\theta) := \mathbb{E}_{t,\mathcal{E}(x_0),\epsilon}||\epsilon - \epsilon_\theta(z_t, t, w)||^2 \quad (1)$$

where ϵ is the applied random noise. $z_t \in \mathcal{Z}$ is the representation of a noised image in the latent space at time step t and w is the encoded text representation. ϵ_θ aims to denoise the latent space. x_0 is the real image at timestep 0 and $\mathcal{E}$ is the latent space encoder.

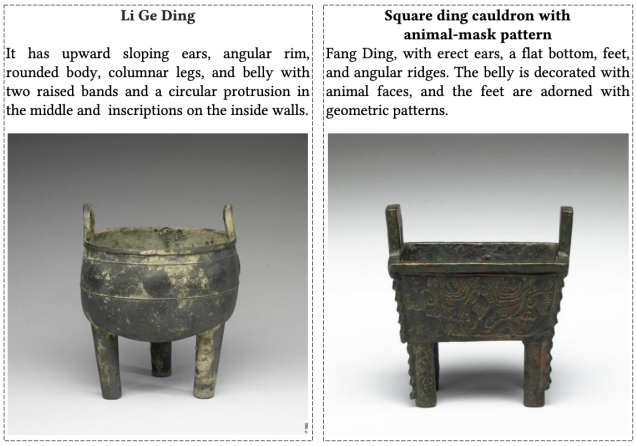

Figure 2: Raw artifact descriptions fail to depict the artifact with sufficient archaeological information, such as their artifact-"type" which determines important aspects of their visual appearance.

4 Our Method

Our proposed approach for knowledge-aware artifact image synthesis is built upon a pretrained Stable Diffusion model, which retains its powerful generative capability of common domains and is further finetuned to align with the specific characteristics of ancient artifacts. The generic Stable Diffusion model, even with finetuning, however, struggles to generate visually and historically accurate artifact images and shows multifaceted errors, as is demonstrated in Figure 1.

To address these issues, we propose specific modifications at three steps in the Stable Diffusion system: **1).** Given source text information T_i of an artifact, we pass it into a LLM (in our case, *GPT-3.5-TURBO*) with carefully designed querying message and in-context examples to obtain a clean and augmented prompt input of our diffusion model T_i'. **2).** During training, when T_i' is passed into the text encoder, we apply an additional constrastive learning module on the text encoder to align the description of an artifact with its name, which is essentially an expert-summary of its description. **3).** After the added noise is predicted in the training phase, we reconstruct the model-predicted image I_i' and apply additional visual-semantic supervision with edge loss (Seif and Androutsos 2018) and perceptual loss (Johnson, Alahi, and Fei-Fei 2016) to steer the generation of our model closer towards the ground-truth appearance of artifacts.

The overall framework for our approach is illustrated in Figure 3 and explained in detail in the following subsections.

4.1 Prompt-Construction Enhanced by LLM

We have noticed that the raw description of an artifact accessible in museum resources (as mentioned in Section 3) is far from ideal for prompting a text-to-image model. It is often incomplete and filled with noisy messages and fails to sufficiently depict an historical object. Other than the messiness problem, these off-the-shelf descriptions may well lack specific information of an artifact that is essential to its visual form, such as its fundamental classification (or: artifact-"type"). An example[1] of this is given in Figure 2. That the artifact on the left side is classified as a "Round Ding" rather than a "Square Ding" confines its shape to a round body having only three legs as opposed to four legs. However, key archaeological information of this kind is often missing in the raw description of the artifact, prohibiting an text-to-image model from sufficiently understanding the association between the visual appearance and the textual prompt.

To alleviate the problems of noisiness and knowledge deficiency in the original text information, we propose to utilize a LLM as both an information extractor to retrieve the most useful information, and as an external knowledge base to complete any missing important attributes of the artifact.

Based on archaeological expertise, we have compiled a list of key attributes that are vital for effectively describing artifacts and defining their physical forms, see Table 1. Examples of these attributes are given in Table 5 in Appendix A[2]. While the "name", "time period" and "size" of an artifact are usually available in museum resources, the specific "material", "shape" and "pattern" need to be extracted or derived from the raw description of the object. Further, as explained above, the classified "type" of an artifact determines certain fundamental aspects of its looks, which are specified by the generic definition of this artifact-type (*i.e.*, "type definition"). It requires a general knowledge of archaeology to be able to categorize an ancient object into a certain artifact-type and to define the basic appearance of this type.

A LLM is well-suited for fulfilling these two tasks with

[1]To maintain a consistent language usage throughout the paper, we translate all Chinese text (*e.g.*, the textual descriptions of artifacts) into English via ChatGPT.

[2]All technical appendices mentioned in the main text are submitted as supplementary material.

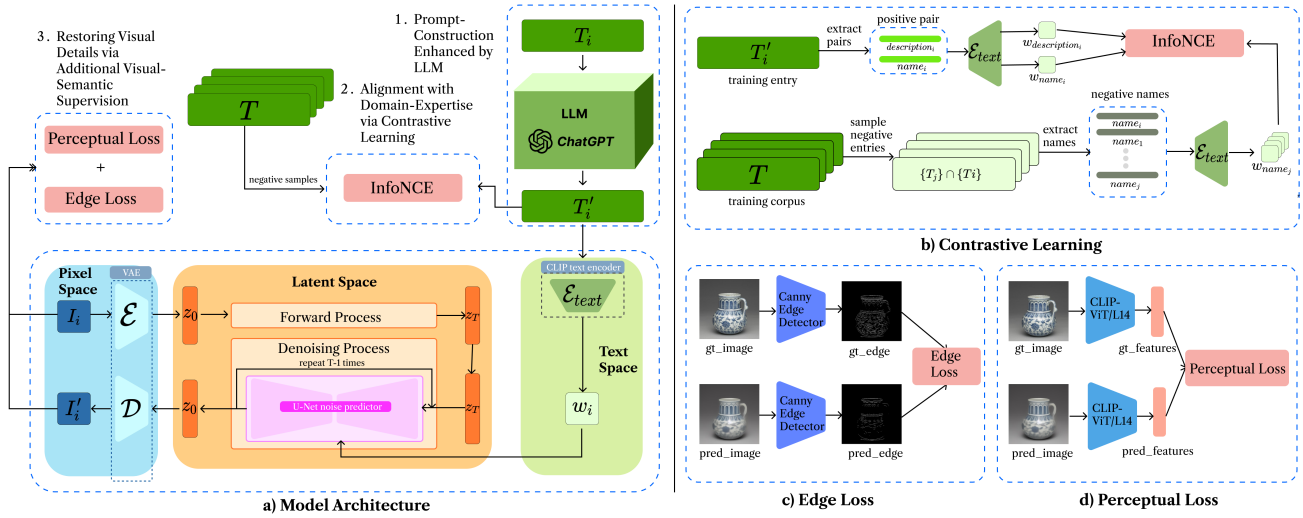

Figure 3: Our proposed knowledge-aware approach is illustrated in a). It features a Chinese Stable Diffusion model as backbone and our proposed three key techniques labeled as follows: b) illustrates the way of performing textual contrastive learning, which is discussed in Section 4.2; c) is edge loss and d) is perceptual loss, both of which are part of the additional visual-semantic supervision, as discussed in Section 4.3.

Expert Attribute	Definition
Name	name or title of an artifact
Material	the material an artifact is made of
Time Period	time period of origin
Type	classified type of an artifact
Type Definition	general definition of artifact type
Shape	shape and structure of an artifact
Pattern	patterns/motifs on an artifact
Size	physical dimensions of an artifact

Table 1: Expert attributes of artifacts that are vital to their visual appearance according to archaeological expertise.

its ability of obtaining a certain extent of world knowledge from the massive pretraining corpus (Liang et al. 2022) and to learn to perform specialized downstream tasks using the in-context learning paradigm (Brown et al. 2020). Specifically, we use *GPT-3.5-TURBO* as our knowledge-base LLM, and the prompt for querying GPT-3.5 is designed with a similar format following self-instruct (Wang et al. 2022). Our prompt template consists of three parts: 1). A task statement that describes to GPT-3.5 the task to be done; 2). Two in-context examples of high quality sampled from our labeled pool of 54 artifacts written by archaeology experts; 3). The target artifact whose "material", "shape", "pattern", "type" and "type defintion" are left as blank and need to be answered by GPT-3.5. The former 3 attributes can be retrieved from the given "description" and the latter 2 artifact-type related features need to be fulfilled via the world knowledge of GPT-3.5 and its in-context learning from the given human-labeled examples. An example of our prompt for querying artifact information is illustrated in Figure 5 in Appendix B.

By leveraging the power of LLM as both an information extractor and external knowledge provider, we are able to collect all the key attributes of a given artifact, which are then rearranged into prompt to our diffusion model with a $[SEP]$ (implemented as a Chinese comma in our work) splitting each key feature, as shown by the example in Table 5 in Appendix A. Such input prompt thus contains enriched text information that provides well-defined archaeology-knowledge guidance. It assists the text-to-image diffusion model in synthesizing a more realistic result that better corresponds to the ground-truth artifacts.

4.2 Alignment with Domain-Expertise via Contrastive Learning

Another issue we identify is that the text encoder might not encode the text into a representation that reflects the underlying archaeological knowledge and thus needs further fine-tuning. We observe that the names of ancient artifacts are often accurate and concise summarization of the artifact's key attributes, while the descriptions provide an extended version of the artifact's features. Given that both the names and descriptions are provided by domain experts - as written in museum sources, they reflect a high level of expertise in the field. Thus, we believe that closely aligning the names and descriptions is essential to reflect this domain knowledge.

To achieve this goal, we propose the use of contrastive learning, which aims to minimize the distance between positive pairs, consisting of matching $([description]_i, [name]_i)$ pairs extracted from T_i', and to maximize the distance between negative ones with mismatching names. However, we have also observed that artifacts with similar attributes (*i.e.*, similar description contents) and origins share similar names, making it unintuitive to finetune the text encoder to differentiate between the similar pairs. We believe that such

Figure 4: Comparison between the finetuned *Taiyi-SD* baseline model and OUR approach trained with additional edge loss and perceptual loss against the ground truth.

pairs should be close to each other in the semantic space. Therefore, we readjust our sampling strategy for negatives pairs. From historical studies perspective, time period is one of the most determining factors on the style and appearance of an artifact, where different artifact from different eras can be vastly different. Therefore, aiming to separate hard negatives rather than slightly different ones, we sample our negative samples from artifact names in different eras.

In our approach, we use **InfoNCE** (van den Oord, Li, and Vinyals 2018) to penalize the misalignment in the representation encoded by the text encoder. The formula for text constrastive learning can be written as:

$$L_{\text{text}} := -\mathbb{E}_X \log \frac{EXP(x_i)}{\sum_{x_j \in X} EXP(x_j)} \tag{2}$$

where $x_i = \mathcal{E}([description]_i) \cdot \mathcal{E}([name]_i)$ denotes the similarity between a pair positive sample in the text encoder's embedding space. And X is the set of N similarities between sampled pairs from the entire dataset, containing one positive sample x_i and $N-1$ negative samples $x_j \in X$ where $i \neq j$ and $PeriodOf(T_i) \neq PeriodOf(T_j)$.

4.3 Restoring Visual Details via Additional Visual-Semantic Supervision

Generated artifact images by vanilla Stable Diffusion model suffer from blurry edges and false color and patterns under current setting (see Figure 1), implying that stricter visual constraints need to be enforced to address these issues. Therefore, we propose to use **edge loss** (Seif and Androutsos 2018) and **perceptual loss** (Johnson, Alahi, and Fei-Fei 2016) that apply additional visual-semantic supervision on images generated by our Stable Diffusion model.

Edge Loss. Building upon the insights from (Seif and Androutsos 2018), we penalize the differences in contours between two images, by aiming to minimize the L_2 distance between their edge maps, as shown in part c) of Figure 3. Since the vanilla Stable Diffusion model often produces images that suffer from the problem of incorrect and blurry shape compared to the ground-truth artifact, it is necessary to penalize such errors as defined here in the edge loss:

$$L_{\text{edge}} := ||EDGE(I_i) - EDGE(I_i')||^2 \tag{3}$$

where $EDGE(\cdot)$ is an edge extracting function.In our approach, we use Canny Edge Detector (Canny 1986) as our edge extractor to extract edge maps, then compare the difference between two contours in L_2 distance.

Perceptual Loss. Similar to (Johnson, Alahi, and Fei-Fei 2016), we also penalize the problem of mismatching high-level details between the generated image and the real one. As we have also observed on the generated images of vanilla Stable Diffusion model, the high level details (such as colors and patterns) are often misaligned with the original ones. Therefore, we incorporate perceptual loss into our training process to tackle such issue, as perceptual loss works by mapping the images into a semantic space using a pretrained network, and then minimizing the difference between the high-level features of the generated image and the original image. The formula for perceptual loss is defined as

$$L_{\text{perceptual}} := ||\phi(I_i) - \phi(I_i')||^2 \tag{4}$$

where ϕ denotes a pretrained image encoder to extract the high level features of an image. This is applied to impose a stricter supervision on color, texture and other high level features. In our method, we use a *CLIP-ViT-L/14* (Radford et al. 2021) image encoder to act as our pretrained image encoder for perceptual loss.

4.4 Objective Functions

Combining all the extra multi-source multi-modal supervisions above, the overall training objective of our system is:

$$\begin{aligned} L :=& L_{SD} + \lambda_1 \, L_{\text{text}}(x_i, X) \\ &+ \lambda_2 \, L_{\text{edge}}(I_i, I_i') + \lambda_3 L_{\text{perceptual}}(I_i, I_i') \end{aligned} \tag{5}$$

where λ_1, λ_2 and λ_3 are hyperparameters controlling the weight of each supervision loss; x_i is the similarity between a positive sample pair yielded from T_i' and X is a set of similarities of sampled negative pairs; I_i and I_i' are the ground-truth and the restored image from our model's prediction.

Models	Prompt	λ_1	λ_2	λ_3	CLIP-VS ↑	SSIM ↑	LPIPS ↓
Taiyi-SD-finetuned-description	raw description	-	-	-	0.772	0.536	0.608
Taiyi-SD-finetuned-attributes	LLM enhanced attribute	-	-	-	0.792	0.554	0.598
OURS-attributes +text	LLM enhanced attributes	0.5	-	-	0.801	0.580	0.552
OURS-attributes +edge+perceptual	LLM enhanced attributes	-	0.3	0.1	0.815	**0.636**	**0.497**
OURS-attributes +text+edge+perceptual	LLM enhanced attributes	0.3	0.3	0.1	**0.831**	0.594	0.536

Table 2: Quantitative comparison of our models against the finetuned *Taiyi-SD* baselines over CLIP Visual Similarity (CLIP-VS), SSIM and LPIPS.

Prompt	CLIP-VS ↑	SSIM ↑	LPIPS ↓
Raw description	0.748	0.383	0.748
LLM enhanced attributes-sequence	**0.765**	**0.413**	**0.730**

Table 3: Quantitative comparison between zero-shot *Taiyi-SD* models using different textual prompts over CLIP Visual Similarity (CLIP-VS), SSIM and LPIPS.

5 Experiment

5.1 Experimental Setup

Dataset. In face of the sparsity of paired text-image data in the ancient artifact domain, we build our own text-to-image dataset by collecting artifact information from National Palace Museum open data platform (Taipei National Palace Museum 2019). After careful cleansing of available entries, we are left with 16,092 unique artifact samples with their descriptions and ground-truth images. We split the data by 80%/10%/10% for training, validation and testing.

Implementation Details. For our backbone model, we use a pretrained Chinese Stable Diffusion *Taiyi-Stable-Diffusion-1B-Chinese-v0.1* (Zhang et al. 2022a) (dubbed *Taiyi-SD*) which was trained on 20M filtered Chinese image-text pairs. *Taiyi-SD* inherits the same VAE and U-Net from *stable-diffusion-v1-4* (Rombach et al. 2022) and trains a Chinese text encoder from *Taiyi-CLIP-RoBERTa-102M-ViT-L-Chinese* (Zhang et al. 2022b) to align Chinese prompts with the images. Further training details are left in Appendix C.

Evaluation Metrics. To comprehensively evaluate our method for text-to-image synthesis quantitatively, we employ three commonly used metrics that measure image generation quality: **CLIP Visual Similarity**, **Structural Similarity Index (SSIM)** (Wang et al. 2004) and **Learned Perceptual Image Patch Similarity (LPIPS)** (Zhang et al. 2018). Each of them highlights different aspects of the generated image. Together, they provide a thorough judgement of a synthesized artifact image in terms of its overall resemblance to the ground-truth, accuracy of its shape and pattern, and its perceptual affinity to the target image. We leave a an extensive explanation of these metrics in Appendix D.

5.2 Main Results and Discussion

In Table 2, we compare the quantitative results of our approach with the baselines on our test set. The first column denotes the models we experimented with.

For the baselines, we use *Taiyi-SD* via two versions: ***Taiyi-SD*-finetuned-description:** the finetuned *Taiyi-SD* with the raw description (directly available from museum archives) as input prompt; ***Taiyi-SD*-finetuned-attributes:** the finetuned *Taiyi-SD* using LLM-enhanced prompt (a sequence of artifact attributes) as designed in Section 4.1.

For our approach, we apply the LLM-enhanced prompt by default and also explore three different versions of extra supervisions in addition to training the *Taiyi-SD* backbone: ***OURS*-attributes +text:** finetuning with additional text contrastive loss (see Section 4.2) to align the text representation of our model with domain expertise; ***OURS*-attributes +edge+perceptual:** finetuning with edge loss and perceptual loss (see Section 4.3) as additional supervision to enforce more visual-semantic constraints on the image generation process; ***OURS*-attributes +text+edge+perceptual:** finetuning with both text contrastive loss and the edge and perceptual loss as multi-source supervisions.

Overall, our proposed artifact image synthesis approach significantly outperforms the finetuned *Taiyi-SD*-baselines across all metrics. The improvement on SSIM indicates that images generated by our model better preserve the shapes and boundaries of the original artifacts. An increase in CLIP Visual Similarity also indicates that our approach produces images that are more closely aligned to the ground-truths.

Additional visual-semantic constraints in the form of edge loss and perceptual loss contribute greatly to boosting the SSIM and LPIPS score. This can be attributed to the fact that edge loss and perceptual loss put a stricter condition on both structural details like edge and contour (captured by SSIM) and the perceptual-level image features like color and texture (captured by LPIPS). These visual details are exactly much desired in our case of artifact image synthesis, as the shape, pattern and texture of artifacts are of vital importance for determining their historical position and status.

By further incorporating the text contrastive loss into the overall training objective, we observe a slight increase in CLIP Visual Similarity, yet a decrease in SSIM and LPIPS scores. We believe there are two reasons behind this phenomenon. For one, by aligning the text knowledge (descriptions with names) (see Section 4.2), the textual guidance for

Models	Material ↑	Shape ↑	Pattern/Color ↑	Size/Ratio ↑	Dynasty ↑	total avg. ↑
Taiyi-SD-finetuned	2.66	1.50	1.44	1.79	2.12	1.90
OURS	**3.94**	**3.38**	**3.25**	**3.30**	**3.20**	**3.41**

Table 4: Human evaluation of the quality of artifact images generated by the finetuned baseline and our model. The images are rated from 5 different aspects on a scale of 0 to 5 by 20 archaeology experts from top institutions.

generating the image is better represented and closer to the general visual content of the artifact, thus leading to a higher CLIP Visual Similarity. For another, the relative weight of edge and perceptual loss is reduced with the additional text contrastive loss, which might compromise the strict supervision on structural coherence and perceptual similarity and limit the model's performance on SSIM and LPIPS.

As is also evidently shown by just comparing the baselines finetuned with different prompt formats in Table 2, using LLM to enhance the prompt construction as a sequence of important artifact attributes effectively improves performance of the finetuned baseline model across all three metrics. More about the effects of our LLM-enhanced prompting method will be discussed in the following subsection 5.3.

5.3 Ablation Studies

To investigate the contribution of components proposed in our approach and for further studies, we conduct extensive ablation studies on two key designs of our model.

Effectiveness of LLM enhanced prompt. As is shown in Table 2, the finetuned model benefits from LLM-enhanced prompting, achieving better scores on all three quantitative metrics. To further illustrate the effectiveness of our proposed prompting method, we explore the zero-shot setting, where the baseline *Taiyi-SD* is directly prompted to generate artifact images without any training on our artifact dataset. We use either the raw description from the museum archives or the sequence of artifact attributes enhanced by LLM as prompt. The results, as shown in Table 3, again demonstrate the superiority of the LLM-enhanced prompting, which excels across all metrics. This can be credited to the organized information format in the attribute sequence and the additional knowledge provided by LLM (see Section 4.1).

Effectiveness of Edge Loss and Perceptual Loss. In Figure 4, we compare the artifact images generated from our model that uses edge loss and perceptual loss against the finetuned *Taiyi-SD* baseline that does not involve these visual semantic constraints. Evidently, the shape and patterns of the artifacts are more accurate and close to the ground-truth if the model is additionally supervised by edge and perceptual loss. On the other hand, the vanilla finetuning paradigm may easily lead to output objects that are out of shape. For example, in the second column of Figure 4, the "tall-footed" aspect of the target bowl is clearly neglected without edge and perceptual constraints.

5.4 User Study

In addition to quantitative evaluation, we conducted a user study involving archaeology experts to evaluate the generated images. This study is designed to assess various aspects of the generated artifacts, as outlined in our prompt design:

- *Material*: resemblance to the manufacturing material
- *Shape:* match with the described shape
- *Pattern/Color*: faithfulness to actual patterns and colors
- *Size/Ratio*: maintenance of height-width ratio
- *Dynasty*: reflection of era-characteristics

Each aspect is rated on a scale of **0** to **5**, with higher ratings indicating better quality. We randomly select 30 samples from the test set and provide the model-generated images to 20 graduate students of archaeology major from top institutions for assessment. The average ratings of images generated by our proposed method (*OURS*) are compared with those generated by the baseline Chinese SD model also finetuned on our data. The results are presented in Table 4.

Clearly, according to human experts, the artifact images generated by our method are much better in quality across all five important rating aspects, especially in terms of shape, pattern and color. These results of user study resonate with the findings from the automatic evaluation metrics and further highlight the superior performance of our model in generating artifact images that accurately align with history.

To offer a richer qualitative demonstration of our model's capabilities, we present a diverse collection of artifact images generated by our model, showcasing its remarkable fidelity across a broad spectrum of historical artifacts. Refer to Figure 6 in Appendix E for a comprehensive visual display.

6 Conclusion

In this paper, we present a novel approach to tackle the challenge of artifact image synthesis. Our method features three key techniques: 1). Leveraging a LLM to infuse textual prompts with archaeological knowledge, 2). Aligning textual representations with domain expertise via contrasting learning, and 3). Employing stricter visual-semantic constraints (edge and perceptual) to generate images with higher fidelity to visual details of historical artifacts. Quantitative experiments and user study confirm the superior performance of our approach compared to existing models, significantly advancing the quality of generated artifact images. Beyond technological contributions, our work introduces a profound societal impact. As the first attempt to restore lost artifacts from the remaining descriptions, our work empowers archaeologists and historians with a tool to resurrect lost artifacts visually, offering new perspectives on cultural heritage and enriching our understanding of history. We also hope that this work will open new avenues for further exploration, fostering deeper insights into our past and cultural legacy with the help of technical advances.

Expert Attribute	Example
Name	Yuhuchun vase in cobalt blue glaze
Material	Porcelain
Time Period	Qing Dynasty, Yongzheng reign, 1723-1735 AD
Type	Yuhuchun vase
Type Definition	Also known as "narrow-necked vase," yuhuchun vase is a practical commemorative ceramic widely popular in the northern regions. The vase consists of five parts: neck, shoulders, body, foot, and mouth. The neck is long and slender, the body is plump, and the foot can be a short circular footring or a horseshoe-shaped foot. Yuhuchun vases are created using various clay recipes and glaze techniques, resulting in distinct colors and surface effects for each piece
Shape	Flared mouth, slender neck, sloping shoulders, pear-shaped ample body, and a circular footring
Pattern	The body of the vase is adorned with a cobalt blue glaze, which shines with a bright indigo color. The interior and the base of the vessel are covered in white glaze. The footring reveals the white body of the vase
Size	Height of 30.3 cm, mouth diameter of 8.5 cm, base diameter of 12.0 cm
Enhanced Prompt Example	Yuhuchun vase in cobalt blue glaze *[SEP]* Porcelain *[SEP]* Qing Dynasty, Yongzheng reign, 1723-1735 AD *[SEP]* Yuhuchun vase *[SEP]* Also known as "narrow-necked vase," yuhuchun vase is a practical commemorative ceramic widely popular in the northern regions. The vase consists of five parts: neck, shoulders, body, foot, and mouth. The neck is long and slender, the body is plump, and the foot can be a short circular footring or a horseshoe-shaped foot. Yuhuchun vases are created using various clay recipes and glaze techniques, resulting in distinct colors and surface effects for each piece *[SEP]* Flared mouth, slender neck, sloping shoulders, pear-shaped ample body, and a circular footring *[SEP]* The body of the vase is adorned with a cobalt blue glaze, which shines with a bright indigo color. The interior and the base of the vessel are covered in white glaze. The footring reveals the white body of the vase *[SEP]* Height of 30.3 cm, mouth diameter of 8.5 cm, base diameter of 12.0 cm

Table 5: An example of the proposed expert attributes of artifacts. The last row forms the LLM-enhanced prompt input to our text-to-image model. The special delimiter *[SEP]* connecting attributes is implemented as a Chinese comma.

A Example of the Enhanced Artifact Attributes

As a supplement to Section 4.1 and Table 1 in the main text, we provide a concrete example of the proposed expert attributes of the historical artifact "*yuhuchun vase in cobalt blue glaze*" in Table 5 of this appendix[3]. The last row is the arranged sequence of the artifact attributes separated by *[SEP]* (implemented as a Chinese comma), which forms the LLM-enhanced prompt input to our text-to-image models.

B Example of the Prompt Template for Querying GPT-3.5

As specified in Section 4.1 in the main text, we use *GPT-3.5-TURBO* as our knowledge-base LLM. The prompt template for querying GPT-3.5 is designed with a similar format following self-instruct (Wang et al. 2022). It consists of three parts: 1). A task statement that describes to GPT-3.5 the task to be done; 2). Two in-context examples sampled from our labeled pool of 54 artifacts written by archaeology experts; 3). The target artifact whose "material", "shape", "pattern", "type" and "type defintion" are left as blank and need to be answered by GPT-3.5. In between the parts and the different artifact samples, we use "###" as a separator. Figure 5 shows an example of our prompt for querying information about the artifact "*snuff bottle with intertwined floral decoration in fencai polychrome enamels on a yellow ground*".

[3]The numbering of tables, figures and equations in this appendix follows that of the main text. So the number does not start fresh from 1 in this appendix.

C More on Implementation Details

Here we provide additional details with regard to the implementation of our method as a supplement to the model details specified in Section 5.1 of the main text: To get the canny edge map of images for the edge loss computation, we implement a canny filter (Canny 1986) with a Gaussian kernel of size 3, Sobel filter kernel size of 3, a low threshold on pixel intensity of 0.15. We compute perceptual loss (Johnson, Alahi, and Fei-Fei 2016) using visual features extracted by the image encoder of *CLIP-ViT-L/14* (Radford et al. 2021). The weights of additional losses used in the our model training objective defined in (5) of the main text are $\lambda_1 = 0.3$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.1$ (as can been seen in Table 2 in the main text). In addition, we employed the Min-SNR Weighting Strategy (Hang et al. 2023) to facilitate the train-

ing process, resulting in faster convergence using $\gamma = 5.0$. During training, we use an Adam (Kingma and Ba 2014) optimizer and set the batch size to 24 and learning rate to $1.e^{-6}$ for all our experiments, which run on dual NVIDIA A100 GPUs.

D More Details on Evaluation Metrics

In this section, we offer a more extensive explanation of the automatic metrics (**CLIP Visual Similarity**, **Structural Similarity Index (SSIM)** (Wang et al. 2004) and **Learned Perceptual Image Patch Similarity (LPIPS)**) employed to quantitatively evaluate the performance of artifact image generation models. This serves to provide additional technical details to the evaluation metrics stated in Section 5.1 in the main text.

CLIP Visual Similarity. Inspired by the CLIP Score (Hessel et al. 2021), which is widely employed to assess the similarity between a text-image pair, we compute the visual similarity of two images (*i.e.*, the ground-truth image and the generated image) with the visual module of a pretrained CLIP model (specifically, CLIP-ViT-L/14) (Radford et al. 2021). We refer to this metric as Clip Visual Similarity, which is also utilized in (Gal et al. 2022; Kumari et al. 2022; Wei et al. 2023). Since the CLIP model has shown a strong ability in capturing the overall image contents and mapping those into a feature space, a higher similarity of the encoded visual features suggests a generally closer resemblance of the generated image to the ground-truth.

Structural Similarity Index (SSIM). SSIM (Wang et al. 2004) is designed to primarily measure the similarity in terms of structural components between two images. It evaluates how well the synthesized output preserves the structural details present in the ground truth images. This includes important edges, boundaries and overall structural coherence. By quantifying the degree of structural resemblance, SSIM provides valuable insights into the accuracy of artifact image synthesis in terms of preserving crucial details related to the formative appearance of artifacts, *e.g.*, their shape and patterns.

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS (Zhang et al. 2018) judges the perceptual similarities between two images. In artifact image synthesis tasks, it is indispensable to assess not only the structural similarity but also the perceptual quality of the synthesized images. LPIPS is designed to align with human perception of image quality which also corresponds to the goal of artifact image synthesis tasks of generating historical objects that are visually convincing to human experts.

E More Examples of Artifact Images Generated by Our Model

Our model is capable of synthesizing images of a wide range of artifacts with high historical accuracy based on simple textual descriptions, as shown in Figure 6 in this appendix.

F Dataset and Code

In our commitment to reproducibility and to facilitating further research, we will readily make all our **dataset** and **codebase** publicly available upon publication of this paper. This will also include the **data preprocessing scripts** and a **license** that allows free usage for research purposes.

G Diffusion Models and Stable Diffusion

In this appendix, we present a mathematical introduction of diffusion models and the Stable Diffusion architecture which serves as the backbone of our method (see Section 3 in the main text).

Diffusion Models (Rombach et al. 2021; Sohl-Dickstein et al. 2015; Song et al. 2020; Ho, Jain, and Abbeel 2020; Song and Ermon 2020) are a family of probabilistic models which involves two processes: forward process and reverse process. Let $p(x_0)$ be the image distribution and $x_0 \sim p(x_0)$ be a real image in the distribution. The forward process q (also known as diffusion process) is a Markov chain of a length of a fixed timestep T that applies random noises from Gaussian distribution on to the previous state according to a variance schedule $\beta_1, \beta_2, ..., \beta_T$, where

$$q(x_t|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{6}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1}; \beta_t \boldsymbol{I}) \tag{7}$$

The reverse process p_θ is a Markov chain that aims to reverse the diffusion process by denoising the random noises at time step t where $1 < t \leq T$ to eventually restore the real image x_0. The goal is to learn the parameter θ for Gaussian transitions starting from $p(x_T) = \mathcal{N}(x_T|\mathbf{0}; \boldsymbol{I})$, where

$$p_\theta(x_0) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{8}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\boldsymbol{\mu_\theta}(x_t, t), \boldsymbol{\Sigma_\theta}(x_t, t)) \tag{9}$$

The original diffusion model (Ho, Jain, and Abbeel 2020) set $\boldsymbol{\Sigma_\theta}(x_t, t) = \sigma_t^2 \boldsymbol{I}$ and $\boldsymbol{\mu_\theta}(x_t, t)$ to predict noise ϵ_t at time step t with a noise predictor ϵ_θ parameterized by θ, leading to the loss function

$$L(\theta) := \mathbb{E}_{t,x_0,\epsilon}||\epsilon - \epsilon_\theta(\alpha_t x_0 + \sigma_t \epsilon, t)||^2 \tag{10}$$

where $\alpha_t = \sqrt{1-\sigma_t^2}$ and $\sigma_t^2 = \beta_t$ and ϵ_θ is a neural model using U-Net (Ronneberger, Fischer, and Brox 2015) as the backbone.

Stable Diffusion (SD) (Rombach et al. 2021) introduces perceptual image compression which first maps the image to a latent space, then applies diffusion process and reverse process on the latent space rather than the pixel space which was used in earlier diffusion models. Overall, a Stable Diffusion model can break down into three parts: a VAE (Kingma and Welling 2013), a text encoder and a U-Net (Ronneberger, Fischer, and Brox 2015).

现有3件文物，并有名称和描述对各件文物做出介绍。请根据文物的名称和描述，回答各件文物的材质、形态、纹饰、器型，并给出该种器型的器型基本外形定义。前2件是例子，第3件请按同样的格式输出。

###

名称：青花云龙纹双耳扁壶

描述：扁壶，小唇口，直颈，器腹呈扁圆形，浅圈足呈椭圆形，颈肩处饰以螭形双耳。口沿及圈足壁饰以青地白纹的海波纹样；颈饰变形蕉叶纹，肩部饰如意云头纹一周；瓶身腹部，正背两面皆饰圆形团龙云纹，外绕一圈留白。龙龙纹作正面龙形，四之四展，五爪犄张；龙身鳞片勾勒细腻，云纹穿绕隙地。满布缠枝花卉纹，近足处为一圈变形莲瓣纹。底有「大清乾隆年制」六字款。

材质：青花瓷

形态：扁壶，小唇口，直颈，器腹呈扁圆形，浅圈足呈椭圆形，颈肩处饰以螭形双耳。

纹饰：口沿及圈足壁饰以青地白纹的海波纹样；颈饰变形蕉叶纹，肩部饰如意云头纹一周；瓶身腹部，正背两面皆饰圆形团龙云纹，外绕一圈留白。龙龙纹作正面龙形，四之四展，五爪犄张；龙身鳞片勾勒细腻，云纹穿绕隙地。满布缠枝花卉纹，近足处为一圈变形莲瓣纹。

器型：双耳扁壶

器型基本外形定义：以扁壶而有双耳得名。小口，细颈，扁圆腹，颈肩处置双耳，椭圆形圈足或长方倭角圈足。

###

名称：茶叶末釉螭耳花浇

描述：花浇圆口，周缘切平，一侧向外突出形成短尖流。短颈，圆硕腹，口腹接一道边棱，侧置一螭形把，口下划一道，定出螭首衔贴位置。底内挖成卧足，底心浅印「雍正年制」两行四字篆款。全器罩施茶叶末釉，口部釉层下流，形成褐色边。釉面清楚可见橘皮棕眼。此器祖型为西亚玉器或金属器，十五世纪明朝永乐、宣德官窑临仿之烧制成青花瓷器。由于永乐作品较常出现双首螭纹柄的造型，故以为此品当是雍正仿永乐之作。回溯明朝，因有与西亚交流的背景，因此瓷花浇的烧制，可视为是反映史实的具体例证。相对于此，雍干两朝档案倒是出现有永宣花浇的纪录，遂让人从中推想今日将此类器皿称作花浇，当是沿袭清宫旧称。据此也能明白花浇加设出水流口的设计，或是为了进一步落实使用功能，而进行的改良。那么始见于康熙朝的茶叶末釉加施于外，也让此品兼具仿古器形和新釉彩的特色。

材质：茶叶末釉瓷

形态：花浇圆口，周缘切平，一侧向外突出形成短尖流。短颈，圆硕腹，口腹接一道边棱，侧置一螭形把，口下划一道，定出螭首衔贴位置。底内挖成卧足。

纹饰：全器罩施茶叶末釉，口部釉层下流，形成褐色边。釉面清楚可见橘皮棕眼。

器型：花浇

器型基本外形定义：室内浇花用的瓷壶、瓷杯。罐形腹，口沿处有流，相对处有把柄。

###

名称：粉彩黄地缠枝花卉纹鼻烟壶

描述：鼻烟壶，口外撇，短颈，扁圆腹下敛，上厚下薄，前后微鼓，平底，附红色盖，盖上突起小钮，下接木塞、牙匙，用来舀壶里的鼻烟，或盛在小碟上，或直接放在拇指背上，入鼻嗅用。全器满布黄地彩绘卷枝花纹，花心伸出桃实；盖沿、口沿及器缘均描金边。底外有一行四字篆款「嘉庆年制」。乾隆、嘉庆年间，常成套烧制釉上彩鼻烟壶，此器为一套十件的其中一件。

材质：

形态：

纹饰：

器型：

器型基本外形定义：

There are three cultural relics, each with a name and a description. Please answer the material, shape, pattern, and type of each relic based on its name and description, and give the basic external shape definition of the type of vessel. The first two are examples, and the third one should be in the same format.

###

Name: Flask with handles and decoration of cloud and dragon in underglaze blue

Description: Flat pot, small lip mouth, straight neck, flat oval belly, shallow round foot in an elliptical shape, and adorned with dragon-shaped double ears at the neck and shoulder. The mouth and foot are decorated with white wave patterns on a blue background; the neck is adorned with deformed banana leaf patterns, and the shoulder is adorned with ruyi cloud patterns all around; both the front and back of the belly are adorned with round dragon cloud patterns, surrounded by a circle of white space. The dragon pattern is in the form of a front-facing dragon, with four legs spread out and five claws open; the dragon's scales are delicately outlined, and the cloud pattern weaves through the gaps. The entire body is covered with entwined floral patterns, and near the foot is a circle of deformed lotus petal patterns. The bottom has a six-character inscription "Da Qing Qianlong Year Made".

Material: Blue and white porcelain

Shape: Flat pot, small lip mouth, straight neck, flat oval belly, shallow round foot in an elliptical shape, and adorned with dragon-shaped double ears at the neck and shoulder.

Pattern: Mouth and foot are decorated with white wave patterns on a blue background; the neck is adorned with deformed banana leaf patterns, and the shoulder is adorned with ruyi cloud patterns all around; both the front and back of the belly are adorned with round dragon cloud patterns, surrounded by a circle of white space.

Type: Double-eared flat pot

Basic external shape definition: Named for its flat pot with double ears. Small mouth, thin neck, flat round belly, double ears at the neck and shoulder, and elliptical or rectangular foot.

###

Name: Pitcher with a chi-dragon handles in tea-dust glaze

Description: The flower sprinkler has a round mouth, flat edge, and one side protrudes outward to form a short, pointed spout. Short neck, round and full belly, a rib connects the mouth and belly, and a dragon-shaped handle is placed on the side. The bottom is carved into a lying foot, and the shallow center is marked with “Yongzheng Year Made” in two rows of four-character seal script. The entire piece is covered with a tea-dust glaze, and the glaze layer flows down from the mouth, forming a brown edge. The glaze surface clearly shows orange-peel brown eyes. The prototype of this piece is a Western Asian jade or metal piece, and the Ming Dynasty official kiln in the 15th century made it into blue and white porcelain. As Yongle's works often feature double dragon pattern handles, it is believed that this piece is an imitation of Yongle during the Yongzheng period. Tracing back to the Ming Dynasty, due to the background of exchanges with Western Asia, the firing of porcelain flower sprinklers can be seen as a concrete example reflecting historical facts. In contrast, the Yong and Xuan archives have records of Yongxuan flower sprinklers, which led people to speculate that the name “flower sprinkler” today is inherited from the old name in the Qing Palace. It can also be understood that the design of adding a water outlet to the flower sprinkler is an improvement for further implementing the use function. Then, the tea-dust glaze applied from the Kangxi period onward gives this piece a combination of antique shape and new glaze color features.

Material: Tea-dust glazed porcelain

Shape: Round mouth flower sprinkler, flat edge, one side protrudes outward to form a short, pointed spout. Short neck, round and full belly, a rib connects the mouth and belly, and a dragon-shaped handle is placed on the side. The bottom is carved into a lying foot.

Pattern: The entire piece is covered with a tea-dust glaze, and the glaze layer flows down from the mouth, forming a brown edge. The glaze surface clearly shows orange-peel brown eyes.

Type: Flower sprinkler

Basic external shape definition: A porcelain pot or cup used for indoor flower watering. Jar-shaped belly, with a spout at the mouth and a handle on the opposite side.

###

Name: Snuff bottle with intertwined floral decoration in fencai polychrome enamels on a yellow ground

Description: Snuff bottle, outward-flaring mouth, short neck, flat round belly that tapers down, thick at the top and thin at the bottom, slightly bulging front and back, flat bottom, with a red lid, the lid has a small raised knob, connected to a wooden plug and ivory scoop, used to scoop snuff from the bottle, or put it on a small plate, or directly on the back of the thumb for sniffing. The entire piece is covered with yellow ground famille rose entwined floral patterns, with peach fruit extending from the flower center; the lid edge, mouth edge, and vessel edge are all outlined in gold. The bottom has a row of four-character seal script "Jiaqing Year Made". During the Qianlong and Jiaqing periods, sets of overglaze painted snuff bottles were often made, and this piece is one of a set of ten.

Material:

Shape:

Pattern:

Type:

Basic external shape definition:

Figure 5: An example of our GPT-3.5 querying prompt. We use Chinese by default because of the origin language of our data. The right-hand side is an English translation also done by the same *GPT-3.5-TURBO* engine.

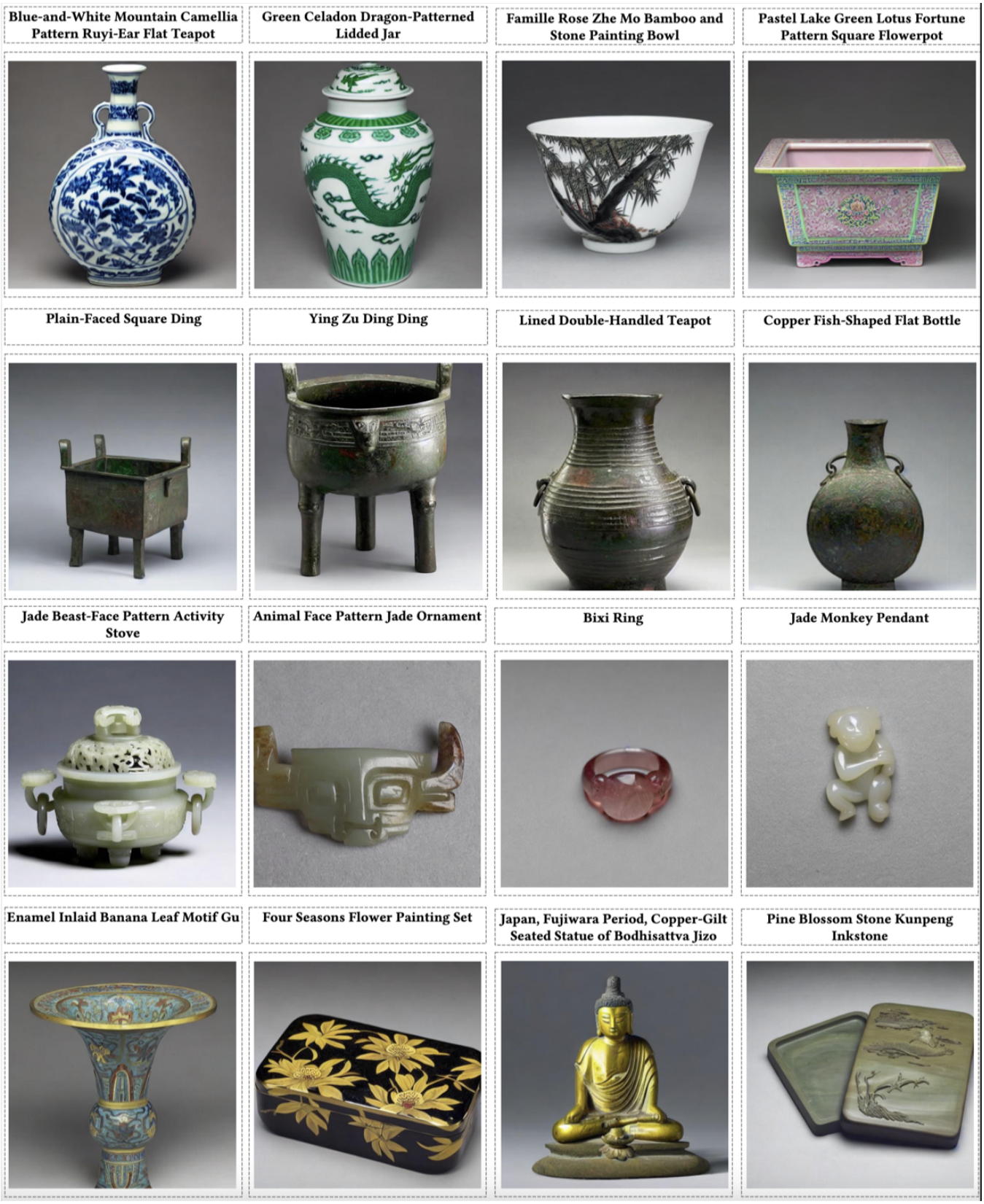

Figure 6: High-fidelity images of a wide range of artifacts generated by our model.

VAE (Kingma and Welling 2013) contains two parts: an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. In Stable Diffusion (Rombach et al. 2021), the encoder part of VAE (Kingma and Welling 2013) is used to encode the image x into latent space $\mathcal{Z}$ in the forward process during training. And the decoder part of VAE (Kingma and Welling 2013) is used to decode the the denoised latent representation into image at inference time.

Text Encoder $\mathcal{E}_{text}$ is responsible for mapping raw text into a embedding space that can be used to condition the backward denoising process. *i.e.*, For a raw text input S, the text encoder maps it to w such that $w = \mathcal{E}_{text}(S)$. Stable diffusion (Rombach et al. 2021) uses a pre-trained CLIP (Radford et al. 2021) as the text encoder.

U-Net (Ronneberger, Fischer, and Brox 2015) contains two parts: an encoder and a decoder, with ResNet (He et al. 2015) as the block structure. The encoder part projects the image into a low resolution image presentation and the decoder part aims to restore the original image. Similar to the original diffusion model, in Stable Diffusion model, the U-Net structure serves as ϵ_θ and aims to denoise the latent space at time step t where $1 < t \leq T$ during reverse process, conditioned on the text embedding using cross-attention (Vaswani et al. 2017).

During fine-tuning, the train objective for Stable Diffusion can be written as

$$L_{SD}(\theta) := \mathbb{E}_{t,\mathcal{E}(x_0),\epsilon}||\epsilon - \epsilon_\theta(z_t, t, w)||^2 \qquad (11)$$

where $z_t \in \mathcal{Z}$ is the representation of image in the latent space at time step t and $\mathcal{E}$ is the latent space encoder. w is the text representation encoded by $\mathcal{E}_{text}$.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Canny, J. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6): 679–698.

Chang, X.; Chao, F.; Shang, C.; and Shen, Q. 2022. SundialGAN: A Cascade Generative Adversarial Networks Framework for Deciphering Oracle Bone Inscriptions. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 8780–8794. Curran Associates, Inc.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. arXiv:2111.14822.

Hang, T.; Gu, S.; Li, C.; Bao, J.; Chen, D.; Hu, H.; Geng, X.; and Guo, B. 2023. Efficient Diffusion Training via Min-SNR Weighting Strategy.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up GANs for text-to-image synthesis.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. .

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2022. Multi-Concept Customization of Text-to-Image Diffusion. *arXiv preprint arXiv:2212.04488*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2022. Holistic Evaluation of Language Models.

Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Le Bras, R.; Choi, Y.; and Hajishirzi, H. 2022. Generated Knowledge Prompting for Commonsense Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3154–3169. Dublin, Ireland: Association for Computational Linguistics.

Nichol, A.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. arXiv:2102.09672.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. Hong Kong, China: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-resolution image synthesis with latent diffusion models.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and 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)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

Seif, G.; and Androutsos, D. 2018. Edge-Based Loss Function for Single Image Super-Resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1468–1472.

Sheng, S.; and Moens, M.-F. 2019. Generating Captions for Images of Ancient Artworks. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM.

Shi, D.; Diao, X.; Shi, L.; Tang, H.; Chi, Y.; Li, C.; and Xu, H. 2022. CharFormer: A Glyph Fusion based Attentive Framework for High-precision Character Image Denoising.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. .

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Song, Y.; and Ermon, S. 2020. Generative Modeling by Estimating Gradients of the Data Distribution. arXiv:1907.05600.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. .

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. arXiv:2011.13456.

Taipei National Palace Museum. 2019. National Palace Museum Open Data. Data retrieved from National Palace Museum Open Data, https://theme.npm.edu.tw/opendata/index.aspx?lang=2.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with Contrastive Predictive Coding.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need.

Wang, C.; Liu, P.; and Zhang, Y. 2021. Can Generative Pre-trained Language Models Serve As Knowledge Bases for Closed-book QA? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3241–3251. Online: Association for Computational Linguistics.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. arXiv:2212.10560.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2017. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2022. Diffusion models: A comprehensive survey of methods and applications.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. 2016. StackGAN: Text to photo-realistic image synthesis with Stacked Generative Adversarial Networks.

Zhang, J.; Gan, R.; Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; Huang, Y.; Li, X.; Wu, Y.; Lu, J.; Zhu, X.; Chen, W.; Han, T.; Pan, K.; Wang, R.; Wang, H.; Wu, X.; Zeng, Z.; and Chen, C. 2022a. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. .

Zhang, J.; Gan, R.; Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; Huang, Y.; Li, X.; Wu, Y.; Lu, J.; Zhu, X.; Chen, W.; Han, T.; Pan, K.; Wang, R.; Wang, H.; Wu, X.; Zeng, Z.; and Chen, C. 2022b. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. .