Master's Thesis

Diffusion-based Image Generation of Oracle Bone Inscription Style Characters

Xiaoxuan Xie

Supervisor Haoran Xie

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2025

Abstract

Oracle bone inscriptions, a corpus of ancient Chinese script carved onto animal bones and turtle shells, constitute one of the most valuable cultural assets in understanding the early formation of Chinese civilization. Dating back more than three millennia (circa the Shang Dynasty, c. 1600–1046 Before the Common Era), these inscriptions embody proto-forms of Chinese characters and reflect the beliefs, rituals, historical records, and sociopolitical structures of the period.

The distinct pictographic nature of oracle bone inscriptions, where glyphs represent objects or concepts, makes them an invaluable resource for archaeologists, epigraphers, historians, and artists. However, the comprehensive digital analysis, stylistic rendering, and generation of oracle bone inscription-style images present profound challenges. These challenges stem from limited datasets, intricate visual features, and the need to translate modern concepts into the archaic and stylistically rich oracle bone inscription's visual language.

However, applying state-of-the-art generative artificial intelligence (AI) to oracle bone inscriptions is nontrivial. Transformation of a modern object image into an oracle bone inscription-inspired glyph requires modeling a unique aesthetic that is neither purely symbolic nor entirely representational. oracle bone inscription glyphs often combine ideographic and pictographic elements, implying that their stylistic formation depends on both the object's semantic meaning and visual representation. Unlike conventional style transfer tasks that apply superficial filters or patterns, oracle bone inscription style transformation demands fidelity to ancient carving techniques, line thickness variations, spatial composition rules, and subtle textural cues that reflect inscription on hard surfaces rather than ink on paper. Moreover, the historical erosion of numerous oracle bone inscription samples introduces additional noise and uncertainty to the visual features.

To solve these issues, this thesis introduces a generation pipeline based on a diffusion model specifically tailored to generate images in the style of oracle bone inscriptions. The proposed approach builds upon three key components: (1) constructing a domain-specific dataset aligning ancient oracle bone inscription references, textual descriptions, and contemporary object images, it contains 44 categories of oracle bone inscription and 180 sets of data pairs; (2) fine-tuning a diffusion model enhanced by ControlNet to achieve controllable oracle bone inscription-style image generation aligned with both shape and semantic intent; and (3) refining generative outputs to better adhere to the structural norms, carving patterns, and stylistic conventions of authentic

oracle bone inscriptions. Evaluations using IP-Adapter, pix2pix, and CycleGAN demonstrate that the proposed method achieves superior results in generating semantically consistent oracle bone inscription-style images.

Moreover, the proposed method is evaluated with the IP-Adapter, pix2pix, and CycleGAN. In qualitative evaluation, this work shows excellent performance in reconstructing existing oracle bone inscriptions, generating new characters, and generating diverse stylistic variants. In quantitative evaluation this work achieves optimal scores in Fréchet Inception Distance, CLIP Image-Image Similarity, and Neural Image Assessment. In the user preference study, 44% of the users preferred the results generated by this work and also obtained the highest scores for original image similarity. All the evaluations show that this method generates semantically consistent oracle bone inscription-style images.

Contents

List of Figures

List of Tables

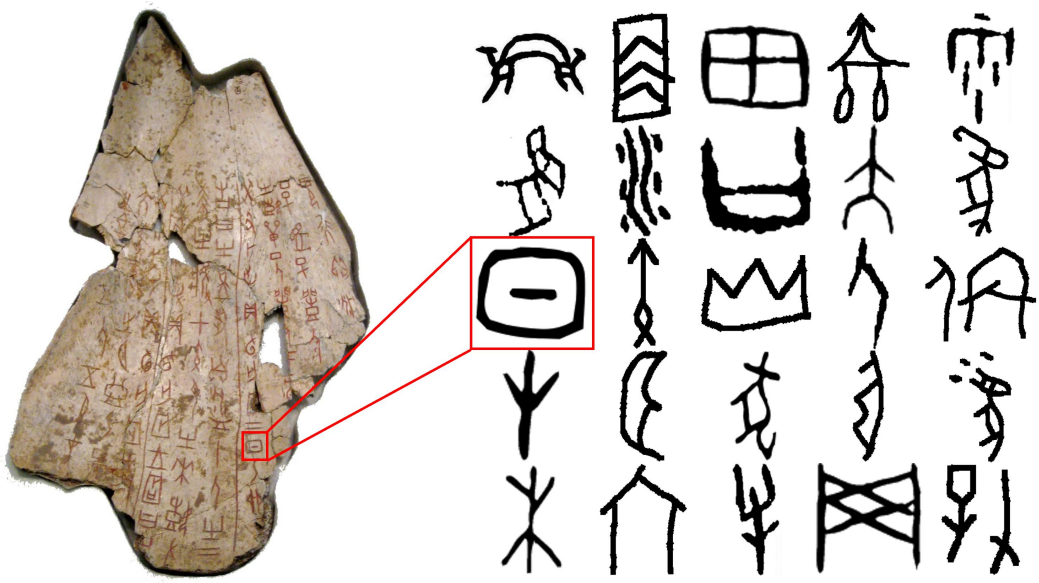
Chapter 1

Introduction

The study of oracle bone inscriptions (OBIs) represents a crucial intersection between historical linguistics, archaeology, and computational analysis. As the earliest known form of Chinese writing, OBIs not only provide insights into the evolution of the Chinese script but also serve as invaluable records of early human civilization. Their unique pictographic and ideographic nature differentiates them from modern writing systems, making their study both historically significant and technically challenging. However, despite their importance, OBIs remain difficult to analyze and reproduce due to their fragmented state, stylistic variations, and the lack of comprehensive datasets.

With advancements in artificial intelligence (AI) and generative models, there is growing interest in applying computational techniques to aid in the digital reconstruction and generation of OBI-style images. While traditional methods rely heavily on manual efforts by experts, machine learning approaches—particularly those leveraging generative adversarial networks (GANs) and diffusion models—offer promising new directions for automated OBIs synthesis. These methods have the potential to facilitate historical preservation, stylistic exploration, and artistic reinterpretation of OBIs. However, existing generative approaches often struggle with maintaining the stylistic authenticity and semantic integrity required for OBIs representation.

This study addresses these challenges by introducing a novel diffusion model-based framework tailored specifically for OBI-style image generation. By leveraging a domain-specific dataset and incorporating structured constraints, this research seeks to balance generative flexibility with stylistic fidelity, ensuring that generated OBI-style images remain both visually compelling and historically accurate. The following sections provide a detailed examination of the historical background of OBIs, the motivations behind this research, and the technical contributions of this study.



(a) (b)

Figure 1.1: An oracle bone fragment (a) and oracle bone inscriptions (b). The inscription in the red box is "sun". Images license under creative commons of Wikipedia.

1.1 Background

Oracle bone inscriptions (OBIs), dating back over 3,500 years, represent one of the earliest known forms of written communication in human history. These inscriptions, etched onto turtle shells or animal bones, were primarily used for divination purposes during the Shang Dynasty (c.1250 –c.1046 BC) in ancient China. The historical significance of OBIs lies not only in their status as an early form of script but also in their role as a record of political, social, and religious practices of the time. They offer unique insights into early Chinese civilization, contributing to research fields such as archaeology, linguistics, and cultural studies.

Unlike linear and abstract modern writing systems, such as the Roman alphabet or Arabic script, OBIs exhibit unique pictographic and ideographic features. These inscriptions often visually resemble the objects or concepts they represent, blending artistic abstraction with communicative intent. For instance, as shown in the red box in Figure 1.1 the character for "sun" takes the form of a circular representation with a dot at the center, closely mirroring the natural form of the sun.

In the realm of art and design, OBIs inspire modern reinterpretations,

influencing calligraphy, graphic design, and cultural exhibitions [9] [10]. Their distinctive aesthetic qualities, combining simplicity and abstraction, make them a cornerstone of visual culture in East Asia. However, the lack of standardized forms and the significant variations in style present challenges for consistent reproduction and analysis.

Despite their significance, OBIs present substantial challenges for both archaeological research and computational analysis. More than 160,000 fragments have been uncovered, as shown in Figure 1.1(a), yet only 4,500 unique characters have been cataloged, of which two-thirds remain undeciphered [11]. The inscriptions lack a standardized writing system, and their stylistic variations reflect differences in authorship, periods, and intended usage. These complexities make OBIs difficult to interpret, analyze, and reproduce, particularly when adapting them for modern applications such as digital art or cultural preservation. Moreover, the irregular brushstrokes and degraded nature of numerous inscriptions add layers of difficulty to computational processing.

Nowadays, artificial intelligence technology has been explored extensively to understand and explore this ancient script. Researchers have utilized generative adversarial networks (GANs) to address the challenges of data scarcity and stylistic variability. For example, GAN models were employed to directly train existing OBIs for recognition and style adaptation [12]. Additionally, GAN-based approaches extended OBIs writing styles through handwritten adaptations, enabling the generation of synthetic OBIs to augment training datasets and improve recognition accuracy [13]. Despite their success in generating a sufficient number of synthetic OBIs, GANs often face issues related to controllability and unstable training, limiting their practical applications for tasks requiring precise stylistic and semantic alignment. In contrast, diffusion models (DMs) have emerged as a more robust and controllable alternative for image synthesis. Models such as Stable Diffusion [14] and ControlNet [8] leverage stable learning objectives and text-prompt-based controllability to produce high-quality, semantically consistent images. By iteratively refining noisy latent variables, diffusion models achieve superior stability and flexibility compared to GANs, making them particularly well-suited for tasks involving complex visual styles and semantic precision [15]. However, existing research in diffusion models for OBIs primarily focuses on recognition and detection tasks and often has difficulty to decipher inscriptions without integrating text prompts for generating content or exploring interpretations of OBI-style images.

Despite these technological advancements, applying state-of-the-art generative models to generate OBI-style images remains a non-trivial task. Challenges include the scarcity of annotated data, the unique pictographic struc-

ture of OBIs, and the need to preserve their stylistic integrity while adapting them for modern contexts. These factors underscore the importance of developing specialized approaches that can balance fidelity to the original style with creative flexibility.

1.2 Research Motivation

Oracle bone inscriptions (OBIs) hold immense historical, cultural, and linguistic value, yet their study and preservation present significant challenges. One of the key issues lies in the limited availability of high-quality OBIs datasets, as only a small subset of the total discovered inscriptions has been digitized, annotated, or analyzed[16, 1, 1, 4]. Furthermore, the incomplete and eroded state of many inscriptions adds noise and uncertainty to their visual features, complicating efforts to standardize and interpret these ancient scripts. The irregular shapes and diverse stylistic variations of OBIs—resulting from differences in authorship, carving techniques, and historical periods—further exacerbate the difficulty of analysis. As shown in Figure 1.2, this is a set of the same oracle bone inscription "Jian", meaning two people are holding a dagger-axe together.

Figure 1.2: Different styles of the same oracle bone inscription. The thickness, position, and even direction of the strokes differ from each other but have the same meaning.

Generating OBI-style images introduces a unique set of technical challenges. Unlike traditional style transfer tasks [17], OBI-style generation may go beyond applying superficial visual effects or filters. Instead, it demands fidelity to the intricate visual characteristics of OBIs, including their irregular line thickness, spatial composition, and the distinct pictographic features that connect abstract glyphs to the forms or ideas they represent. These inscriptions blend semantic meaning with visual representation, requiring a nuanced approach that preserves their ideographic essence while accurately reflecting their stylistic identity. Capturing these elements ensures that generated images are both visually compelling and contextually faithful to their origins.

The conventional generative models, such as GANs [18], have been employed to address some of these challenges. However, GAN-based approaches often suffer from unstable training and limited controllability, which hinder their ability to generate stylistically precise and semantically consistent OBI-style images. Meanwhile, diffusion models offer promising improvements in stability and controllability, their application to OBIs has largely been limited to recognition and detection tasks [19, 20, 21, 22], without addressing the generation of OBI-style images from contemporary inputs, such as textual descriptions or object images.

To summarize the issues and their technical challenges, Figure 1.3 provides an outline of the inputs and outputs involved in this work. The inputs consist of modern object images and textual descriptions, while the outputs are OBI-style images that adhere to the semantic and stylistic characteristics of the originals.

1.3 Technical Contributions

This work addresses the significant challenges in generating high-quality oracle bone inscription (OBI)-style images by introducing a novel diffusion model-based framework, DiffOBI [23]. This work makes the following key contributions:

Domain-Specific Dataset Construction: This work constructed a unique dataset tailored to the requirements of OBI-style image generation. This dataset aligns ancient OBIs references with contemporary object images and textual descriptions, serving as a foundation for robust model training and evaluation.

Diffusion model-Based Generative Framework: DiffOBI is proposed as a two-stage diffusion model-based pipeline enhanced with ControlNet for style control over generated outputs. The framework enables the transformation of modern object images into semantically meaningful and stylistically accurate OBI-style images, going beyond conventional style transfer methods by incorporating both semantic and stylistic fidelity.

Refinement Module for Enhanced Quality: To address the stylistic and structural complexities of OBIs, a refinement module was developed to iteratively improve the raw outputs of the diffusion model. This module ensures that the final results adhere closely to the structural norms, line thickness, and spatial compositions characteristic of OBIs.

Comprehensive Evaluation Framework: Extensive evaluations were conducted to compare DiffOBI with state-of-the-art generative models, including GAN-based approaches (e.g., pix2pix [24], CycleGAN [25]) and re-

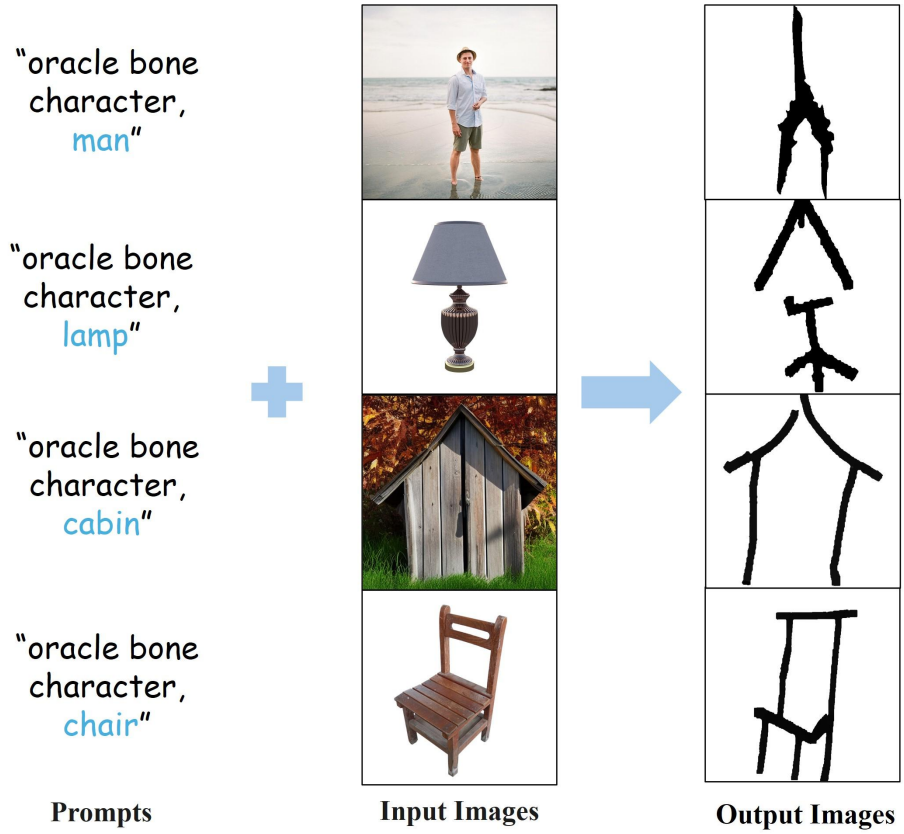


Figure 1.3: Overview of the input and output in the proposed framework. Modern object images and textual prompts serve as inputs, while the generated outputs are oracle bone inscription (OBI)-style images that retain semantic and stylistic fidelity.

cent diffusion models (e.g., IP-Adapter [26]). Both qualitative results and user preference studies demonstrate the superiority of the proposed method in producing semantically consistent and visually compelling OBI-style images.

Advancing Cultural Preservation through AI: This work contributes to the preservation and reinterpretation of ancient Chinese cultural heritage by leveraging modern artificial intelligence techniques. By bridging the gap between historical artifacts and contemporary applications, the proposed approach provides a novel tool for cultural preservation, artistic expression, and education.

1.4 Outline of Thesis

The remainder of this paper is organized as follows:

In Chapter 2, this thesis explores the historical and cultural significance of oracle bone inscriptions (OBIs) and examines their current relevance in research. It provides an overview of advancements in generative modeling techniques, including GANs, VAEs, and diffusion models, and introduces conditional image generation methods. Additionally, this chapter highlights the limitations of existing datasets and methods in OBIs studies, emphasizing the need for semantic alignment and stylistic fidelity in generated OBIs representations.

In Chapter 3, this thesis provides foundational insights into the three core domains relevant to this research: OBIs, diffusion models, and ControlNet. It details the origins, characteristics, and cultural significance of OBIs, followed by an in-depth explanation of Denoising Diffusion Probabilistic Models (DDPMs) and Latent Diffusion Models (LDMs). Lastly, it introduces ControlNet as an innovative framework for controlled generation, providing the theoretical basis for the proposed method.

In Chapter 4, this thesis introduces DiffOBI, the proposed framework for OBI-style generation. The framework comprises a two-stage architecture: the generation stage employs a ControlNet-enhanced diffusion model to produce semantically and stylistically consistent OBI-style images, while the refinement stage applies advanced optimization techniques for binary filtering, impurity removal, edge enhancement, and resolution improvement.

In Chapter 5, this thesis outlines the construction of a novel OBI-specific dataset tailored to support the proposed method. It details the selection of data sources, the preprocessing pipeline for aligning textual descriptions with images, and the use of ControlNet for generating aligned object images.

In Chapter 6, this thesis evaluates the effectiveness of DiffOBI through qualitative and quantitative analyses. It presents the model's performance in reconstructing existing OBIs, generating novel glyphs, and producing stylistically diverse outputs. Comparative studies with baseline models such as pix2pix, CycleGAN, and IP-Adapter are included, alongside user preference surveys and metric-based performance evaluations.

In Chapter 7, this thesis discusses the broader implications of the findings, evaluating the strengths and limitations of DiffOBI in the context of OBIs research and generative modeling. Future research directions, such as expanding dataset coverage, enhancing computational efficiency, and exploring additional applications, are proposed. The chapter concludes by summarizing the study's contributions to the field of OBIs preservation and generative modeling.

Chapter 2

Related Works

In this chapter, an overview of existing works that are related to the proposed framework DiffOBI is provided. In Section 2.1 first discusses the study of oracle bone inscriptions (OBIs), reviewing key advances in character detection, recognition, and evolutionary analysis. In Section 2.2, Few-Shot Font Generation (FFG) is examined, highlighting its shared challenges with OBIs and how these two domains may inform each other. Section 2.3 explores related datasets for OBIs, focusing on their different properties and roles in driving the oracle decipherment and oracle image generation fields. Finally, Section 2.4 explores recent developments in generative models, emphasizing their potential to support complex, historical script tasks.

2.1 Oracle Bone Inscriptions

In the domain of oracle bone inscriptions (OBIs) research, various methods have been proposed to facilitate character detection, recognition, and interpretative analysis. OBIs were used approximately 3000 years ago, provide critical insights into ancient Chinese writing and early societal structures [27, 22]. Despite their value, many inscriptions remain challenging to interpret because of degraded bones, background noise, and incomplete glyph shapes, prompting researchers to pursue automated techniques for accurate and efficient OBIs analysis [28, 21]. Early work often relied on handcrafted graph-based or morphological features, but the rise of deep learning has led to significant improvements in robustness and processing efficiency when dealing with noisy, fragmentary data[22, 29].

Two-stage paradigms, in which a detection model isolates inscription regions before a classifier recognizes individual glyphs, have become common for dealing with cluttered or incomplete bone surfaces [27, 29]. Furthermore,

generative adversarial networks (GANs) were introduced to aid interpretation by producing morphological shape cues or even simulating character evolution. For example, Gao et al. [30] leverage GAN-based translation models to generate shape hints for partially annotated OBIs, enabling their system to handle incomplete or noisy strokes; in doing so, they provide clearer glyphs that experts can reference for further decoding. Meanwhile, Sundial-GAN proposed by Chang et al. [19] adopts a cascaded structure of multiple GANs, each targeting a different historical phase so that an oracle bone inscription image can be gradually transformed toward a recognizable modern form, capturing both coarse and finer evolutionary traits in the process. In tasks with limited samples or under-interpreted characters, specialized frameworks based on few-shot learning or pseudo-category labeling can mitigate data imbalance, leveraging relationships between known and unknown categories to improve generalization [20, 21]. While certain networks rely on direct convolutional neural network (CNN) backbones or Inception-style models, others incorporate attention modules or metric learning, helping address the unbalanced distribution of OBIs classes and the diversity of stroke patterns [22].

Gao et al. also benefited from newly established datasets and annotation software, accelerating the creation of large-scale labeled corpora [30]. These datasets often include real rubbing images with severe cracks and occlusions, thereby testing the resilience of deep neural models [28, 22]. To address the challenge of fragment rejoining, which is essential for reconstructing ancient bones, Zhang et al. [29] introduced a set of strategies designed to align broken edges and extract local shape features. These strategies are complemented by an interface that ranks potential fragment matches, facilitating the reassembly process. Their pipeline can process boundary curves of cracks, match fragment contours, and produce a short list of possible rejoin pairs for expert validation.

2.2 Few-Shot Font Generation (FFG)

Font generation aims to transfer a set of references in a particular style into novel character shapes with the same style while preserving the original content. Recently, few-shot font generation (FFG) has emerged as a vital direction in this domain, aiming to synthesize new fonts using as few exemplars as possible (e.g., below 10 reference glyphs). As shown in figure 2.1, a small number of reference fonts (a) are input as a style cue while a standard font (b), and other characters (c) with the style of font (a) are output. Prior research can be broadly categorized into global style representation and

localized style representation.

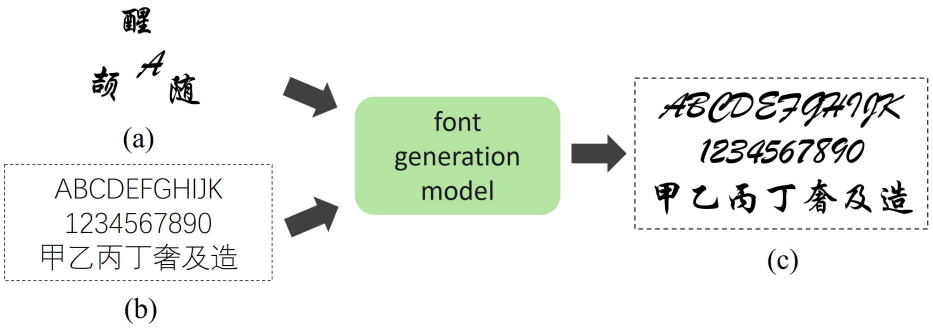


Figure 2.1: The process of few-shot font generation. (a) denotes a small number of reference fonts, (b) is a standardized font, and (c) is generated new fonts.

A significant number of existing few-shot font generation approaches adopt a global style representation for each font. For instance, AGIS-Net encodes the shape and texture of glyph images to address style transfer under a very limited set of references [31]. Similarly, Zhang et al. [32] proposed the EMD framework, which learns a universal style embedding along with content features to achieve better style-content disentanglement. These methods are effective for transferring relatively simple font styles but often struggle with capturing the complex local details of characters, particularly for non-linear scripts like Chinese, which include sophisticated radicals and strokes. This limitation becomes even more pronounced in OBIs, where glyphs are not simply textual characters but pictographic sketches closely tied to objects or concepts.

To capture the internal structure of glyphs, several works propose learning localized style features. LF-Font factorizes the style embeddings into component and style factors so that the model can reconstruct the entire vocabulary even with incomplete references [33]. DG-Font integrates deformable convolutions in the generator and uses a multi-task discriminator to handle large stroke-level variations without direct style label supervision [34]. Similarly, CF-Font and XMP-Font further refine content-style fusion through basis font features [35] and cross-modality pre-training [36], respectively. These approaches excel in scenarios where precise alignment between radical-level structure and style is required. However, applying such techniques directly to OBIs faces unique challenges, as OBIs are not merely structured characters but highly irregular glyphs influenced by artistic abstraction and ancient carving tools. Unlike font glyphs that follow consistent compositional rules,

OBIs often resemble freehand sketches with varying spatial arrangements and degraded edges[37].

To address the limitations of global and localized style representations, recent studies focus on fine-grained local-style mining. For example, Tang et al. [38] propose an attention-based style aggregation mechanism to precisely extract subtle patterns from references, enabling their system to generate glyphs that faithfully capture both style and content. While this approach shows promise for detailed scripts, OBIs require even higher levels of adaptability due to their hieroglyphic nature, where glyphs are conceptual representations of objects (e.g., "sun," "rain") rather than abstract characters. Fine-grained methods for OBIs may account for this pictographic complexity.

While few-shot font generation shares certain characteristics with OBIs image generation such as the need for style transfer with limited references—key differences exist. Generate OBIs are fundamentally different from generate fonts. Fonts typically emphasize uniformity and regularity, whereas OBIs embody a pictographic and artistic nature, more akin to ancient sketches or conceptual drawings. The irregular stroke patterns, spatial variability, and semantic alignment of OBIs introduce challenges that go beyond those encountered in traditional font generation. The proposed methods developed for FFG, such as deformable convolutions and attention-based style aggregation [34, 38], may inspire, but adaptations are required to accommodate the unique visual and semantic characteristics of OBIs.

2.3 Datasets for OBIs

The study of oracle bone inscriptions (OBIs) has been significantly advanced by various datasets that differ in scale, data types, and annotation granularity. These datasets primarily focus on deciphered characters, rubbings, handwriting, and even multi-modal representations, catering to diverse research tasks such as recognition, denoising, and decipherment. Early datasets, such as the one presented by Guo et al. [16], introduced approximately 20,000 character images collected from non-public resources, focusing exclusively on deciphered OBCs and proposing a hierarchical framework to bridge shape-based recognition and sketch analysis. Similarly, datasets like OBC306 [4], comprising over 300,000 images from scanned rubbings and authoritative works, emphasize large-scale raw data but cover only 306 deciphered categories, reflecting a long-tail distribution where frequent characters dominate. These efforts highlight the foundational role of deciphered datasets, despite their limited coverage of the many undeciphered OBCs. To address the challenges posed by real rubbings, recent datasets prioritize multi-modal and real-

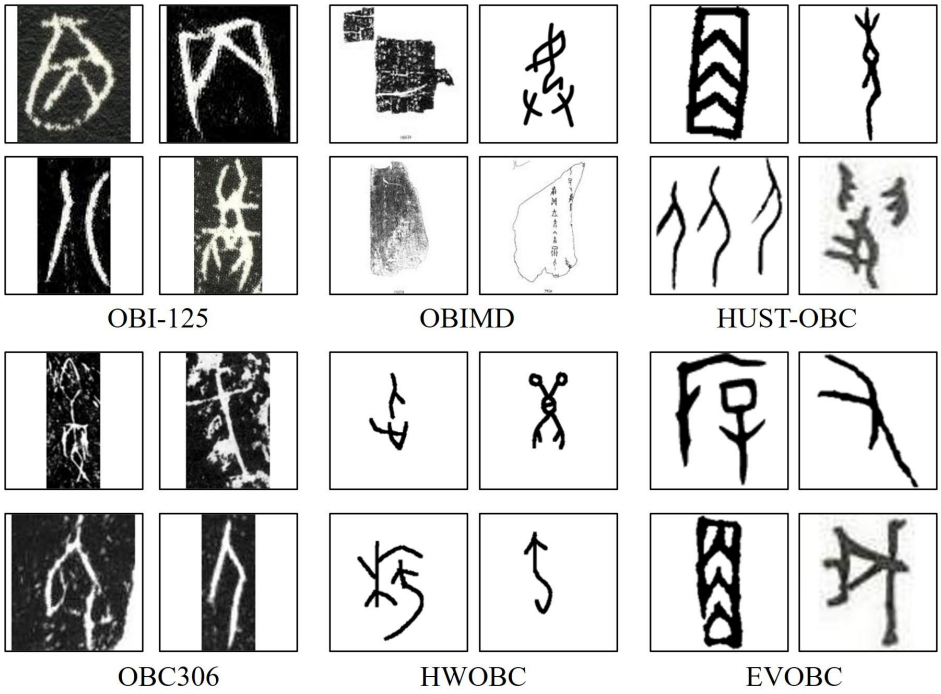


Figure 2.2: Sample images from publicly available Oracle datasets, listed in order, are OBI-125 [1], OBIMD [2], HUST-OBC [3], OBC306 [4], HWOBC [5], EVOBC [6].

world data characteristics. OBIMD [2] combines over 10,000 pieces of oracle bones with pixel-aligned facsimiles, rubbings, bounding boxes, and reading sequences, facilitating tasks such as enhancing rubbing clarity and predicting character reading order. In contrast, OBI-125 [1] focuses specifically on rubbing-type data, offering 125 categories with dynamic data augmentation to handle imbalance and overfitting issues. These datasets capture the noise, cracks, and occlusions inherent in rubbings, thereby enhancing the resilience of models trained for real-world conditions.

Other datasets expand the scope to handwriting-based or evolutionary studies. HWOBC [5], containing 83,245 samples across 3,881 categories, emphasizes calligraphy and handwriting recognition. On the other hand, EVOBC [6]focuses on the historical evolution of Chinese characters, linking oracle bone inscriptions to later scripts such as bronze inscriptions and seal scripts. With 229,170 images spanning 13,714 categories, EVOBC provides a comprehensive resource for studying morphological lineages and facilitates zero-shot analyses of glyph evolution. Moreover, HUST-OBC [3] combines deciphered and undeciphered OBCs, offering 77,064 deciphered and 62,989

undeciphered images across multiple modalities, aiming to bridge the gap between known and unknown glyphs for decipherment tasks.

To illustrate the differences in visual characteristics across datasets, Figure 2.2 presents a curated compilation of sample images from all datasets except Oracle-20K, which is excluded due to its non-open-source status. This figure highlights variations in noise, preprocessing, and handwriting inclusion, emphasizing the strengths and challenges posed by each dataset.

In summary, while existing datasets have made significant contributions to OBIs research, they primarily focus on textual, structural, or evolutionary characteristics. Few datasets explicitly explore the pictographic nature of OBIs or their relationships to real-world imagery. This work addresses this gap by emphasizing the visual and semantic alignment between OBI glyphs and corresponding real-world objects. By integrating annotated examples that highlight these connections, the proposed approach facilitates the study of OBIs as both ancient sketches and meaningful linguistic artifacts, pushing forward AI-based oracle bone decipherment and visual interpretation.

2.4 Generative Models

Generative models aim to create new data samples resembling a given dataset, rather than simply classifying or predicting. They learn the distribution of training data and then generate new samples from it. Early generative methods relied on simple probabilistic assumptions, limiting their performance in complex domains. With deep neural networks, generative models can now produce realistic images, fluent text, and coherent audio. However, challenges like mode collapse [39] and capturing data variability persist. Despite these issues, deep generative models are widely applied in creative content generation, simulation, and data recovery. Representative generative models include normalized flows (NFs) [40], energy-based models (EBMs) [41], generative adversarial networks (GANs) [42], variational autoencoders (VAEs) [43], and diffusion models [15], with GANs, VAEs, and diffusion models especially influential in Conditional Image Generation.

Generative Adversarial Networks (GANs) [42] employ an adversarial setup of a generator and a discriminator to synthesize high-quality data. Figure 2.3 illustrates this framework, where the generator attempts to fool the discriminator by producing realistic samples. Following their introduction by Goodfellow et al. [42], GANs rapidly advanced, with DCGAN [44] and ProGAN [45] improving training stability and image resolution. Beyond static images, VGAN [46] targeted video synthesis, while SeqGAN [47] explored text generation via reinforcement learning. In medical imaging, adversarial

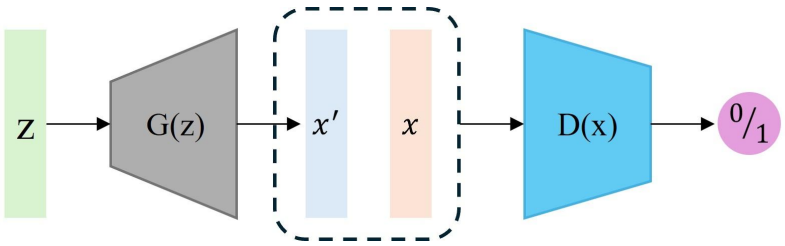


Figure 2.3: Overview of the Generative Adversarial Network (GAN) framework. The generator $G(z)$ transforms random noise z into realistic samples x', while the discriminator $D(x)$ evaluates the realism of generated data.

strategies aided lung segmentation [48] and brain tumor segmentation [49]. More recently, Li et al. [12] addressed long-tailed oracle character recognition through tailored adversarial techniques, boosting tail-class accuracy. Despite their versatility, GANs still face issues like training instability, mode collapse, and limited evaluation metrics.

Variational Autoencoders (VAEs) [43] are probabilistic models that map inputs to a latent Gaussian space for both reconstruction and new sample generation. As shown in Figure 2.4, the encoder produces parameters of a latent distribution, while the decoder attempts to reconstruct the input. A

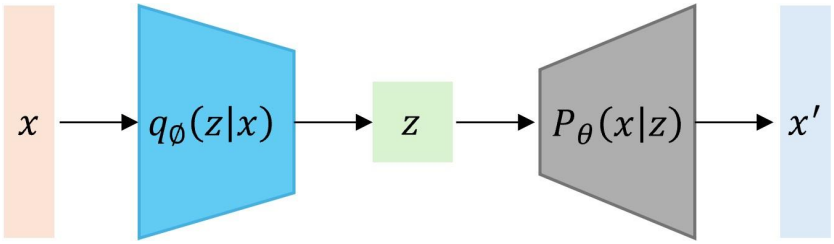


Figure 2.4: Overview of the Variational Autoencoder (VAE) architecture. The encoder $q_\phi(z|x)$ maps input x into a latent space z, while the decoder $p_\theta(x|z)$ reconstructs x' from z.

VAE's loss function combines a reconstruction term and a Kullback-Leibler (KL) divergence term, which regularizes the latent space toward a Gaussian prior. Conditional VAEs (CVAEs) [50] incorporate labels to direct generation, while VFAE [51] disentangles noise from meaningful features. Hybrid approaches, such as CVAE-GAN [52] and PixelVAE [53], combat blurriness by introducing adversarial or autoregressive components. Wasserstein Autoencoders (WAE) [54] adopt the Wasserstein distance for smoother optimization, and NVAE [55] leverages hierarchical structures for high-resolution tasks. VAEs have broad applications in image synthesis, NLP, and anomaly

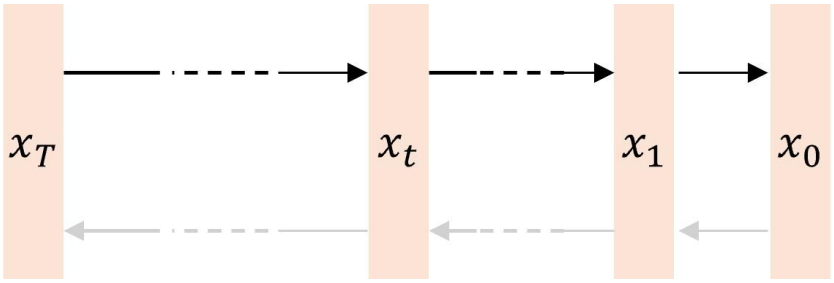


Figure 2.5: Overview of the diffusion model process. Starting from a noisy input x_T, the model iteratively denoises to generate a high-quality output x_0.

detection, though they can produce blurry outputs under pixel independence assumptions and remain computationally intensive.

Diffusion models iteratively refine noisy data into high-fidelity samples, adding noise in a forward process and learning to reverse it. Figure 2.5 shows the progression from noisy input to a clean output. Denoising Diffusion Probabilistic Models (DDPM) [56] and continuous-time SDE approaches [57] laid the groundwork for diffusion-based generative modeling. Text-to-image systems like DALL-E 2 [58], Imagen [59], and Stable Diffusion [14] generate photorealistic images from text. Refinements include InstructPix2Pix [24], Palette [60], and DiffEdit [61] for precise editing, as well as eDiff-I [62] and DreamBooth [63] for higher-resolution and personalized outputs. Beyond images, text-to-3D methods such as Point-E [64] and DreamFusion [65] generate point clouds and neural radiance fields, with Magic3D [66] improving fidelity. Diffusion has also shown promise in text generation [67, 68], medical imaging [69, 70], and molecular design [71, 72, 73]. By iterative denoising, diffusion models mitigate mode collapse and often yield high-quality outputs, though sampling can be slow [74]. Accelerations like DPM-Solver [75] offer faster inference but real-time deployment remains challenging. Overall, diffusion models stand out for their stability, control, and impressive performance across diverse tasks.

2.5 Conditional Image Generation

Conditional image generation has proven to be a transformative tool across diverse applications, ranging from image editing and restoration to more specialized tasks such as personalization, composition, and layout control. These applications leverage the ability of generative models to integrate specific conditional inputs, enabling precise control over the synthesized outputs.

Beyond the commonly used text prompts, modern conditional generation methods introduce additional controls to guide the output more precisely. In sketch-guided flow field generation, Chang et al.[76] utilized a latent diffusion model (LDM) to generate 2D velocity fields constrained by sketches, improving robustness compared to cGAN-based methods. Similarly, Zhang et al.[77] proposed a sketch-guided spatial control framework for text-to-image diffusion models, where segmented sketches provide precise spatial guidance for generating complex multi-object images. In the realm of fashion design, TexControl by Zhang et al. [78] employs a two-stage pipeline integrating sketch-based ControlNet and texture optimization to produce high-quality clothing images with fine-grained details.

A foundational contribution to conditional image generation is the introduction of conditional adversarial networks (cGANs) by Isola et al. [79]. This framework uses a generator-discriminator setup where the generator synthesizes outputs and the discriminator evaluates their realism. The integration of a U-Net-based generator and a PatchGAN discriminator allows cGANs to handle tasks such as converting edge maps to photos, colorizing grayscale images, and translating semantic labels into photorealistic scenes. In recent years, diffusion models have largely replaced GANs as the backbone of conditional image generation tasks due to their superior stability and sample quality. Unlike GANs, diffusion models leverage an iterative denoising process, allowing for more controlled and precise image generation.

The T2I-Adapter improves text-to-image generation by using lightweight convolutional adapters to encode visual inputs, like sketches and depth maps, into multi-scale features. These features are then injected into the U-Net backbone of the diffusion model, allowing it to handle additional conditional inputs beyond text descriptions. This flexibility makes T2I-Adapter particularly useful for multi-modal creative design tasks [80]. Similarly, ControlNet improves conditional control by cloning the deep encoding layers of the diffusion model's U-Net architecture, enabling it to process complex visual inputs like pose and segmentation maps. This design enhances alignment between input conditions and generated images, making ControlNet effective in high-precision applications such as virtual try-on and lighting control [8].

In scenarios requiring fine-grained control over image-based inputs, the IP-adapter introduces cross-attention layers to inject image embeddings directly into the T2I backbone. This method is especially effective for customization and advanced editing tasks, where detailed control is crucial [26]. Meanwhile, GLIGEN focuses on layout control, incorporating gated self-attention mechanisms to process spatial layout information like bounding boxes. By ensuring precise spatial organization of objects, GLIGEN proves invaluable in tasks such as scene generation and compositional editing

[7].Figure 2.6 illustrates two examples of such methods: GLIGEN and ControlNet. GLIGEN incorporates bounding box annotations to define spatial layouts alongside textual prompts, enabling structured generation based on the explicit positional information. Similarly, ControlNet integrates sketch-based conditions, using hand-drawn outlines to ensure the generated images adhere to the desired shapes and structures.

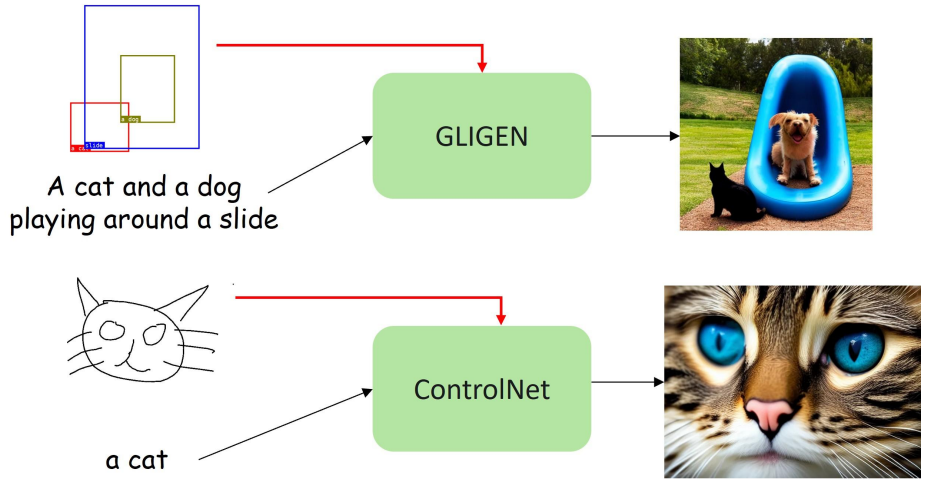


Figure 2.6: Examples of conditional image generation. GLIGEN [7] (top) introduces bounding box control in addition to text prompts to manage spatial layouts, while ControlNet [8] (bottom) uses sketch inputs to provide structural guidance. Both demonstrate how additional conditions enhance control over generated outputs.

For semantic image editing, Imagic utilizes inversion and interpolation techniques to achieve intuitive edits while preserving the original image's core structure and details [81]. It optimizes text embeddings for the source image and interpolates between source and target text descriptions, allowing for high-fidelity edits aligned with user intentions. In the realm of personalization, DreamBooth fine-tunes text-to-image models using a few reference images, embedding unique user-specific objects or subjects into generated images. This approach ensures that personalized elements retain their distinct characteristics, enabling tailored content creation for artistic and professional applications [63].

These models highlight the versatility and power of conditional diffusion frameworks in addressing diverse and complex image-generation tasks. Despite their successes, challenges remain, such as the computational intensity of iterative sampling and the difficulty of aligning new conditional inputs with

pre-trained models. However, ongoing advancements continue to expand the possibilities for diffusion models, paving the way for further innovation in personalized, efficient, and precise image synthesis.

Chapter 3

Prior Knowledge

This chapter presents the essential background knowledge that underpins this research. It begin by examining oracle bone inscriptions (OBIs) in Section 3.1, highlighting their historical importance and unique linguistic characteristics. Next, in Section 3.2 explores existing OBIs databases, discussing their respective scopes, data preprocessing methods, and annotation strategies. Sections 3.3 and 3.4 introduce denoising diffusion probabilistic models (DDPM) and latent diffusion models (LDM), explaining their foundational principles and their role in facilitating efficient, high-fidelity generative tasks. Finally, Section 3.5 presents ControlNet [8] as an extension of stable diffusion models, demonstrating how external guidance signals can steer the generative process toward specific, task-oriented outputs.

3.1 Oracle Bone Inscriptions (OBIs)

Oracle Bone Inscriptions (OBIs), also referred to as "Jiaguwen", "oracle script", or "inscriptions on tortoise shells and animal bones", represent the earliest known system of mature Chinese writing. These inscriptions, primarily found on turtle plastrons and ox scapulae, date back to the Shang and early Zhou Dynasties (16th to 11th centuries BCE), marking a significant milestone in the evolution of Chinese characters and serving as a vital cultural legacy. The OBIs were first discovered in 1899, with major findings concentrated in the ruins of Yin at Anyang for the later Shang period and in Zhengzhou for earlier Shang remnants [82].

OBIs are deeply rooted in the religious and sociopolitical fabric of Shang Dynasty society, where divination played a central role in decision-making. Shang kings and their officials inscribed questions concerning state affairs, agriculture, warfare, weather, and personal matters onto these bones and

shells. The process involved heating pre-drilled pits on the bones to produce cracks, known as "omen cracks", which were interpreted to derive divine guidance. The inscriptions recorded the question, the resulting cracks, and occasionally the outcome. These inscriptions, therefore, form a detailed historical record, reflecting social, political, economic, and spiritual dimensions of life during the Shang Dynasty.

The OBIs, totaling over 160,000 fragments discovered by 2019 [83], contain more than 4,500 unique characters. Among these, approximately 1,500 have been deciphered, revealing a linguistic system that blends pictographic and ideographic elements. These inscriptions employ various structural mechanisms aligned with the "Six Principles of Chinese Character Formation" (Pictographs, Compound ideographs, Phono-semantic compounds, Indicatives, Derivative cognates, and Loangraphs). Compared to contemporary writing systems, OBIs show significant variability in form and function, yet already demonstrate a mature writing system. In 2017, OBIs were inscribed into the UNESCO Memory of the World Register, recognizing their global cultural significance [84].

OBIs mark the inception of a continuous tradition of Chinese writing, which evolved through subsequent stages such as Bronze Inscriptions (BI), Spring and Autumn Characters (SAC), Warring States Characters (WSC), Seal Script (SS), and Clerical Script (CS), eventually giving rise to modern Chinese characters. While OBIs are known for their flexibility and pictorial nature, they also demonstrate early examples of structural conventions that later solidified into more formalized scripts. These inscriptions, unlike the rigid and highly stylized Seal Script, retain a fluidity akin to sketches, emphasizing their dual roles as practical communication tools and artistic expressions. The evolution from OBIs to modern characters illustrates a gradual shift from symbolic depictions to phonetic and structural regularity, bridging the gap between abstract representation and linguistic precision. This transformative journey is visually captured in Figure 3.1, which outlines the sequential progression from OBIs to contemporary Chinese script.

OBIs continue to inspire interdisciplinary research, bridging fields such as archaeology, linguistics, artificial intelligence, and digital humanities. Their irregularities, while challenging for modern analysis, provide a unique opportunity to study the linguistic creativity and adaptability of early Chinese civilization.

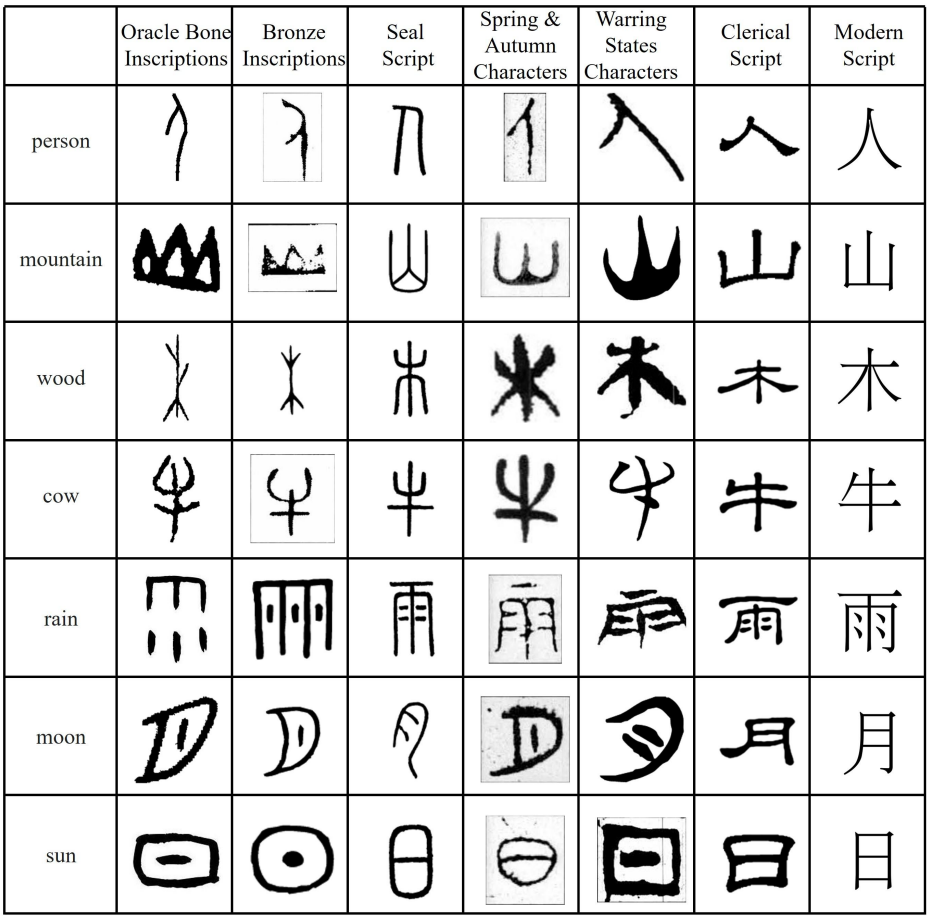


Figure 3.1: Evolution of Chinese Characters from oracle bone inscriptions (OBIs) to Modern Script. Data except modern scripts are from the EVOBC dataset[6]. This diagram illustrates the sequential development of Chinese writing, starting from OBIs and tracing through key historical script forms, including Bronze Inscriptions, Spring and Autumn Characters, Warring States Characters, Seal Script, and Clerical Script, to the contemporary system, with the first column showing the interpretations of their expressions.

3.2 OBIs Datasets

To provide a comprehensive overview of the major publicly discussed datasets for oracle bone characters, this section analyzes the key properties of each resource. The comparison involves aspects such as the total number of images and character classes, whether those character classes have been deciphered, the presence or absence of handwriting or real rubbings, the degree of pre-

processing (e.g., denoising, binarization, image cropping), and the level of annotation (e.g., image-level labels, detection bounding boxes, multi-modal alignment). The following will analyze and summarize each dataset.

The dataset referred to as Oracle-20K [16] consists of 20,039 oracle bone character images spanning 261 deciphered categories. It was compiled by extracting oracle bone characters and their corresponding labels from an online repository that derived its data from an authoritative oracle character lexicon. Initially, a total of 31,876 samples covering 952 unique characters were gathered. However, to maintain consistency and reliability in experimentation, characters with fewer than 25 occurrences were excluded. As a result, the dataset contains characters distributed across deciphered categories, with the most frequently occurring category comprising 291 samples, while the least populated categories contain 25 samples each.

Another dataset is OBI-125 [1], containing 4,257 images of oracle bone characters categorized into 125 deciphered classes. Its images come primarily from scanned books of rubbings, with each image retaining the textures and artifacts of real fragments. The dataset performs only basic image cropping without strong binarization or enhancement, so various forms of noise or incomplete strokes remain. Furthermore, it includes no handwritten samples and only offers image-level class annotations. Although the dataset's scale is smaller compared to some others, it captures genuine rubbing noise and is commonly used for studying oracle bone character recognition under real-world quality constraints. This dataset has a limited category count, making it less suitable for very large-scale tasks.

The OBI-125 dataset [1] comprises 4,275 oracle bone character images categorized into 125 deciphered classes. These images were derived from rubbings scanned from the "Oracle Bone Inscriptions Collection" held by the Shanghai Museum. All characters were manually segmented and classified according to their annotations. The dataset retains the textures and artifacts of real fragments, with minimal preprocessing limited to basic cropping. Noise and incomplete strokes are preserved, and the dataset does not include handwritten samples. It provides only image-level class annotations, focusing on real-world quality constraints for studying oracle bone character recognition. Although the size of this dataset is small compared to other datasets, it emphasizes real friction noise.

OBIMD (Oracle Bone Inscriptions Multi-modal Dataset) [2] focuses on entire bone fragments. Specifically, This dataset includes 10,077 rubbings, each paired with a pixel-aligned facsimile image. These facsimiles were manually created by scholars based on rubbings and historical sources. OBIMD provides comprehensive annotations, including bounding boxes for all visible characters, inscription group information, and reading sequences. Ad-

ditionally, it incorporates a two-level character category structure from the "Oracular Digital Platform" and includes attributes marking contentious or missing parts, and a large fraction of these characters remain undeciphered. Because OBIMD provides pairs of rubbings and facsimiles, it is considered a multi-modal resource, facilitating tasks such as the detection of inscriptions on a full bone piece, multi-modal alignment, and group reading order analysis. Because the data is not uniform single-character images, it is well-suited for layout analysis, line/sequence recognition, or domain adaptation studies.

The OBC306 dataset [4] offers 309,551 single-character images spanning 306 deciphered categories. These images were cropped from scanned or photographed rubbings, covering diverse noise artifacts. The dataset demonstrates a severe long-tail distribution: a few categories account for the majority of samples, while many have fewer than 100 instances. The most frequent category contains 25,898 samples, while 29 categories have only one sample each. It does not include handwritten data or any border annotations but rather provides image-level class labels referring to deciphered modern Chinese counterparts. This dataset has been used primarily for building and evaluating classification models for deciphered oracle bone characters.

Distinct from the above datasets, HWOBC [5] is a handwriting-oriented dataset, featuring 83,245 images produced by twenty-two participants who rendered each oracle bone character on a plain 400×400 white canvas, resulting in uniform, noise-free images. It includes 3,881 classes of deciphered characters, with each class on average having over twenty samples. Since HWOBC is fully handwritten, it bypasses the complexities of rubbings or fragment noise yet loses direct contact with the real surfaces of ancient bones. With only image-level labeling, it offers no bounding boxes or multi-modal references. The dataset contains only image-level labels, making it suitable for use in accelerating the digitization of oracle characters, with text-level parsing, and for future oracle decipherment research.

Another recent dataset is EVOBC [6], which covers six historical script phases, from oracle bone script to clerical script. It has 229,170 images grouped into 13,714 modern Chinese categories, meaning each modern Chinese character is linked to multiple historical shapes. Its data come from a mix of scanned texts and partial manual reproductions across various dynasties, reflecting morphological evolutions. The dataset attempts some standardized contextualization, merges simplified/traditional forms, and maps correspondences between deciphered texts and images, but does not strongly denoise or unify the topographies for each era. It is partially relevant to oracle bone research but focuses on tracing cross-time evolution, so the oracle bone portion is only one subset. EVOBC can thus facilitate studies on script lineage or zero-shot recognition across eras, rather than emphasizing a single

era's noise or data distribution.

Lastly, HUST-OBC [3] consolidates diverse images from scanned books, multiple websites, and existing databases (including HWOBC). This leads to 10,999 classes, of which 1,588 are deciphered and 9,411 remain undeciphered. The total image count is 140,053, with some images being genuine rubbings, others being digital facsimiles, and some being handwriting. The pipeline used for HUST-OBC attempts to standardize backgrounds and remove duplicates across sources, while also separating deciphered from undeciphered classes. Its manual review by experts ensures relatively high annotation quality. HUST-OBC, therefore, is notable for including a large number of undeciphered categories, which can be valuable for zero-shot or decipherment-focused studies.

To better understand the datasets available for oracle bone character (OBC) research, Table 3.1 provides a comprehensive comparison of the key publicly discussed datasets. These datasets vary widely in scale, standardization practices (e.g., denoising, cropping, or background unification), the inclusion of handwriting samples, and the availability of undeciphered characters. For instance, while Oracle-20K [16] provides a foundational dataset of 20,000 images, its closed-source nature limits its accessibility for broader research. In contrast, the remaining datasets, such as OBI-125, OBIMD, OBC306, HWOBC, EVOBC, and HUST-OBC, offer open or partially open resources with diverse attributes and use cases.

Among the datasets compared, EVOBC provides a comprehensive resource with 13,714 categories and a total of 229,170 images. This dataset undergoes a rigorous standardization process, ensuring uniformity in background and preprocessing, making it well-suited for tasks requiring clean and consistent data. In contrast, OBC306, despite offering 309k images, includes only 306 categories and lacks any substantial data preprocessing, which can pose challenges for downstream tasks involving noisy or fragmented samples. On the other hand, HWOBC, a handwriting-based dataset, features entirely redrawn oracle bone characters. While this contributes to the standardization and digitization of OBCs, it omits the critical depth and thickness variations inherent to the original carvings. Considering these factors, the EVOBC dataset was selected as the foundational dataset for this study due to its balance of scale, standardization, and relevance to the objectives of this thesis.

Table 3.1: Comparison of Oracle Bone Inscription Datasets

Name	Scale	Std.	HW	Open
Oracle-20K [16]	20k / 261	No	No	No
OBI-125 [1]	4.2k / 125	No	No	Yes
OBIMD [2]	10.1k fragments	Yes	No	Partially
OBC306 [4]	309.6k / 306	No	No	Yes
HWOBC [6]	83.2k / 3,881	Yes	Yes	Yes
EVOBC [6]	229.1k / 13.7k	Yes	No	Yes
HUST-OBC [3]	140.1k / 11k	Yes	Partial	Yes

Notes:

- **Scale:** Total number of images and categories in the dataset (e.g., "20k / 261" means 20,000 images across 261 classes).
- **Std. (Standardization):** Indicates whether the dataset underwent denoising, cropping, or background unification.
- **HW (Handwriting):** Specifies whether the dataset includes handwritten characters ("Partial" means some handwritten samples are included).
- **Open:** Denotes the dataset's accessibility—whether it is fully open, partially available, or closed.
- For **OBIMD**, "fragments" represent oracle bone fragments rather than single-character images.

3.3 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPM) are landmark generative models that utilize a probabilistic framework to generate high-quality data. By employing a two-phase process inspired by nonequilibrium thermodynamics, DDPM involves a forward diffusion process that incrementally corrupts data with noise and a reverse process that learns to denoise and reconstruct the data, as shown in Figure 3.2. This iterative framework ensures stable training and facilitates the synthesis of realistic samples, outperforming traditional generative models such as GANs and VAEs in various tasks [56].

In the forward diffusion process, the data is progressively perturbed by

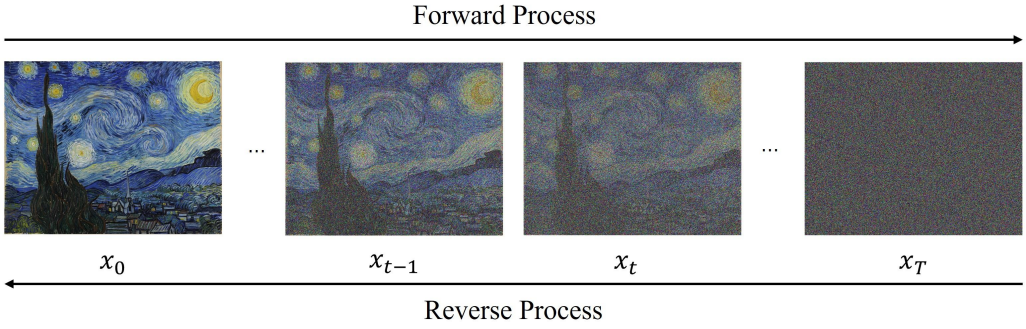


Figure 3.2: An illustration of the diffusion model. In the forward process, noise is progressively added to the data over T steps. In the reverse process, noise is gradually removed until a clean image is recovered.

adding Gaussian noise at each timestep t, transforming the original structured data into nearly pure noise by the final timestep T. Mathematically, this process is represented as:

$$r(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_{t-1}, \beta_t I). \tag{3.1}$$

Here, $r(z_t|z_{t-1})$ represents the transition probability of the data from timestep $t-1$ to t during the forward diffusion process. The term z_t denotes the corrupted data at timestep t, which depends on z_{t-1}, the data from the previous timestep. The term $\alpha_t = 1 - \beta_t$ represents the proportion of the previous timestep's contribution that remains. The term β_t represents the variance of the added Gaussian noise. The symbol $\mathcal{N}$ represents a Gaussian distribution, with the mean $\sqrt{\alpha_t} z_{t-1}$ scaling the contribution of the prior timestep's data and the variance $\beta_t I$ introducing isotropic Gaussian noise. The identity matrix I ensures that the noise affects all dimensions equally.

The reverse denoising process aims to reconstruct the original data by progressively removing the noise added during the forward process. This process is parameterized as:

$$s_\phi(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\phi(z_t, t), \Sigma_\phi(z_t, t)). \tag{3.2}$$

Here, $s_\phi(z_{t-1}|z_t)$ denotes the probability of reconstructing z_{t-1} from the noisy data z_t. The mean $\mu_\phi(z_t, t)$, predicted by a neural network parameterized by ϕ, determines the most likely denoised value. The variance $\Sigma_\phi(z_t, t)$ quantifies the uncertainty in this prediction and is often simplified during training to reduce computational complexity. This iterative denoising transforms the noisy data back into a coherent and realistic sample.

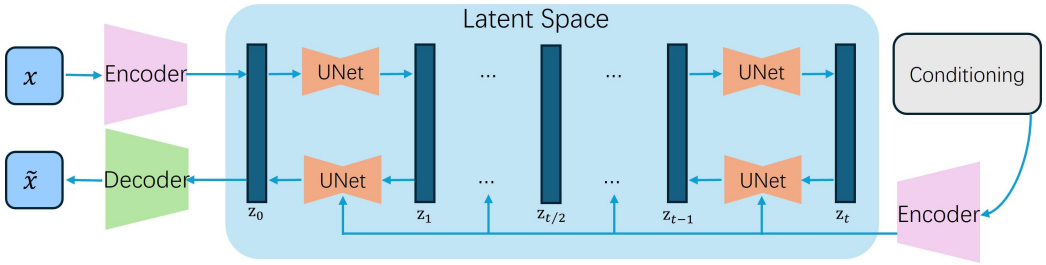


Figure 3.3: The framework of the Latent Diffusion Model (LDM), highlighting its forward and reverse processes across latent and pixel spaces.

DDPM offers several advantages, including robust training dynamics due to its non-adversarial framework and the ability to model complex data distributions with exceptional fidelity. The progressive refinement of samples ensures detailed and realistic outputs. However, the computational demands of DDPM are significant, with slow sampling speeds requiring hundreds of timesteps for generating a single data point.

3.4 Latent Diffusion Models

The Latent Diffusion Model (LDM) [14] introduces a groundbreaking approach to generative modeling by performing the diffusion process in a perceptually compressed latent space. This method significantly reduces computational costs while preserving the high fidelity of generated images. Figure 3.3 illustrates the LDM framework, which integrates a pre-trained autoencoder for compression and a U-Net-based diffusion model for generation. This structure enables LDM to efficiently handle complex tasks such as high-resolution image synthesis, super-resolution, inpainting, and text-to-image generation.

During the forward process, an input image $x \in \mathbb{R}^{H\times W\times 3}$ in pixel space, where H and W represent the height and width of the image, respectively, is first compressed into a latent representation $z \in \mathbb{R}^{h\times w\times c}$ using the encoder $\mathcal{E}$ in the following form:

$$z = \mathcal{E}(x), \tag{3.3}$$

where $h = H/f$ and $w = W/f$, with f as the downsampling factor, c denotes the number of feature channels in the latent space, which captures essential semantic information while reducing spatial dimensions. This compression abstracts high-frequency details while retaining semantic content. The diffusion process then operates in this latent space, where Gaussian noise is iteratively added to z, creating a sequence of noisy representations $\{z_0, z_1, \ldots, z_t\}$.

A time-conditional U-Net architecture is employed to predict and refine the noise at each timestep, leveraging time embeddings to maintain temporal coherence.

A key innovation of LDM is the introduction of conditioning mechanisms through cross-attention. These mechanisms enable multi-modal training and support a variety of conditional image generation tasks. For example, text prompts can be encoded and injected into the U-Net using cross-attention, facilitating text-to-image synthesis. Similarly, class labels or layout information can condition the generation process, supporting tasks like class-conditional image generation and layout-to-image synthesis. The cross-attention mechanism ensures that conditioning information is effectively integrated at every stage of the reverse diffusion process.

In the reverse process, the model begins with a noisy latent variable z_t and iteratively denoises it to recover the clean latent representation z_0. This process is expressed as:

$$z_{t-1} = \text{Denoise}(z_t, t, \text{Conditioning}). \tag{3.4}$$

Finally, the decoder $\mathcal{D}$ reconstructs the image $\tilde{x}$ from the denoised latent representation:

$$\tilde{x} = \mathcal{D}(z_0). \tag{3.5}$$

The separation of perceptual compression and generative modeling offers several key benefits. By performing the diffusion process in a compressed latent space, LDM reduces the computational overhead associated with high-dimensional pixel space. This allows the model to leverage the inductive biases of convolutional U-Net architectures effectively. Furthermore, the use of cross-attention for conditioning inputs enables flexible and precise control over the generated outputs, making LDM highly versatile across a range of applications.

While LDMs demonstrate exceptional capabilities in high-fidelity image generation, generating oracle bone inscriptions (OBIs) remains a significant challenge for existing commercial diffusion models. Figure 3.4 showcases several examples generated using Stable Diffusion 1.5, Stable Diffusion 3.0, Stable Diffusion XL, Flux, and DALL·E 3 models with the text prompt "Oracle Bone Characters, Tree.". These outputs consistently fail to capture the intricate pictographic and stylistic characteristics of OBIs. The generated images lack the semantic and structural precision required for accurate OBIs representation, producing results that neither reflect the intended glyph shapes nor exhibit the aesthetic qualities of oracle bone inscriptions.

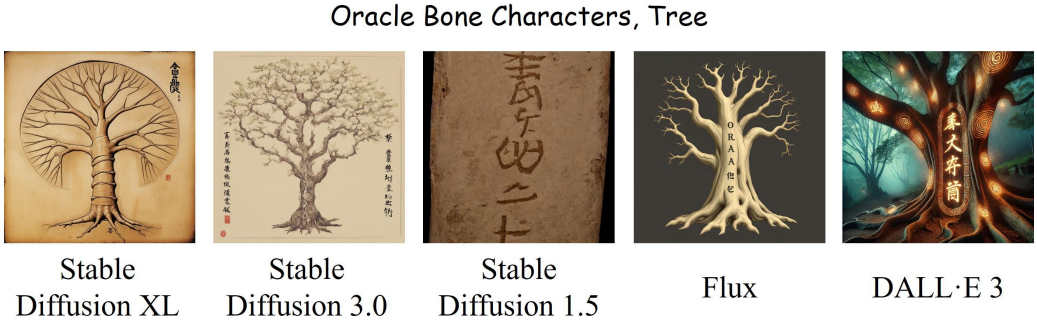


Figure 3.4: Bad examples of OBI-style glyphs generated using Stable Diffusion 1.5, Stable Diffusion 3.0, Stable Diffusion XL, Flux, and DALL·E 3 with the text prompt "Oracle Bone Characters, Tree." These examples highlight the models' inability to capture the stylistic and semantic characteristics of OBIs.

3.5 ControlNet

ControlNet [8] builds upon the foundational work of Latent Diffusion Models (LDMs) and its practical implementation, Stable Diffusion. While LDMs introduced the idea of performing diffusion in a perceptually compressed latent space, Stable Diffusion enhanced this concept by leveraging significantly larger datasets, refined text encoders, and multi-resolution training strategies. Specifically, Latent Diffusion was trained on the LAION-400M dataset, whereas Stable Diffusion utilized the much larger LAION-2B-en dataset. The latter also incorporated data curation techniques, such as filtering out watermarked images and prioritizing images with high aesthetic scores, to improve data quality. In addition, Stable Diffusion adopted a pretrained CLIP text encoder for text conditioning, which proved superior to the randomly initialized transformer used in LDMs. Another key improvement was the training strategy: while Latent Diffusion was trained only at 256×256 resolution, Stable Diffusion was first pretrained at 256×256 and then fine-tuned at 512×512, resulting in significantly better high-resolution outputs.

ControlNet extends these advancements by enabling task-specific conditions to be incorporated into pretrained diffusion models, allowing for more precise control over image generation. Figure 3.5 illustrates the ControlNet framework, which introduces and integrates task-specific conditional inputs, such as edge maps, depth maps, or keypoint annotations, into the intermediate layers of the U-Net architecture. This integration is carefully designed to preserve the pretrained capabilities of Stable Diffusion while allowing the model to adapt flexibly to diverse tasks. The additional conditions are pro-

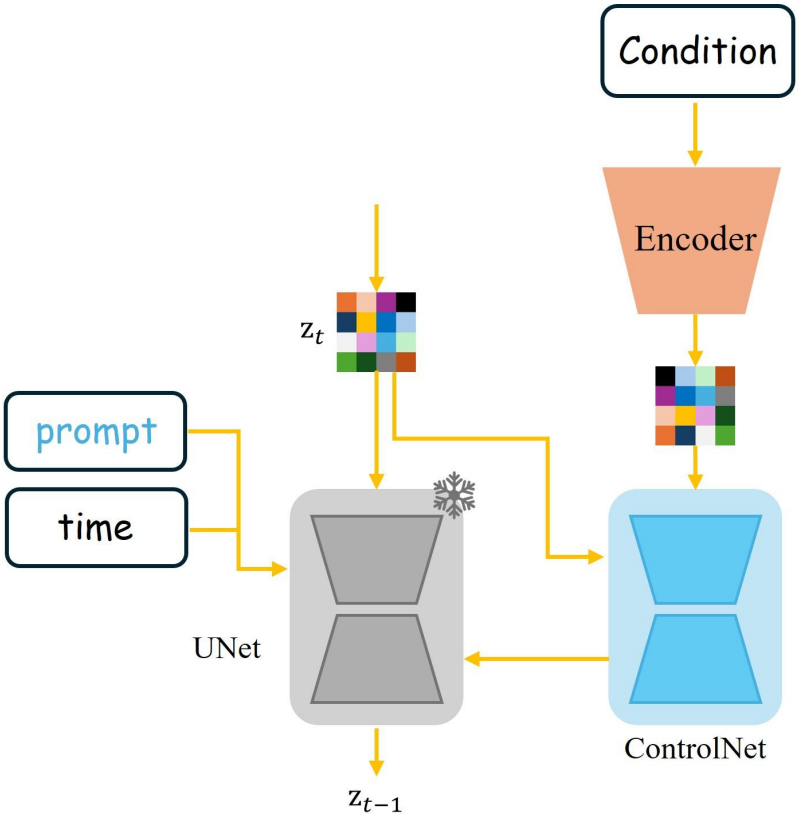


Figure 3.5: The framework of ControlNet, illustrates how task-specific conditions are injected into the network via the training copy to control the generative process. This mechanism preserves the pretrained model's capabilities while adapting to new tasks.

cessed through dedicated layers that inject external guidance signals directly into the generative process, ensuring seamless adaptation to new inputs without disrupting the original model's behavior.

The results of incorporating various conditional inputs into ControlNet are illustrated in Figure 3.6. For example, when edge maps are provided as conditions, ControlNet effectively reconstructs structural details while generating realistic textures and colors that adhere to the given edges. Similarly, depth maps allow the model to synthesize images with coherent depth and perspective, enhancing scene realism. Keypoint annotations, such as those produced by OpenPose, guide the generation of human figures, ensuring accurate poses and limb placements. Segmentation maps enable layout-specific synthesis, ensuring that different regions in the image correspond to their intended semantic labels.

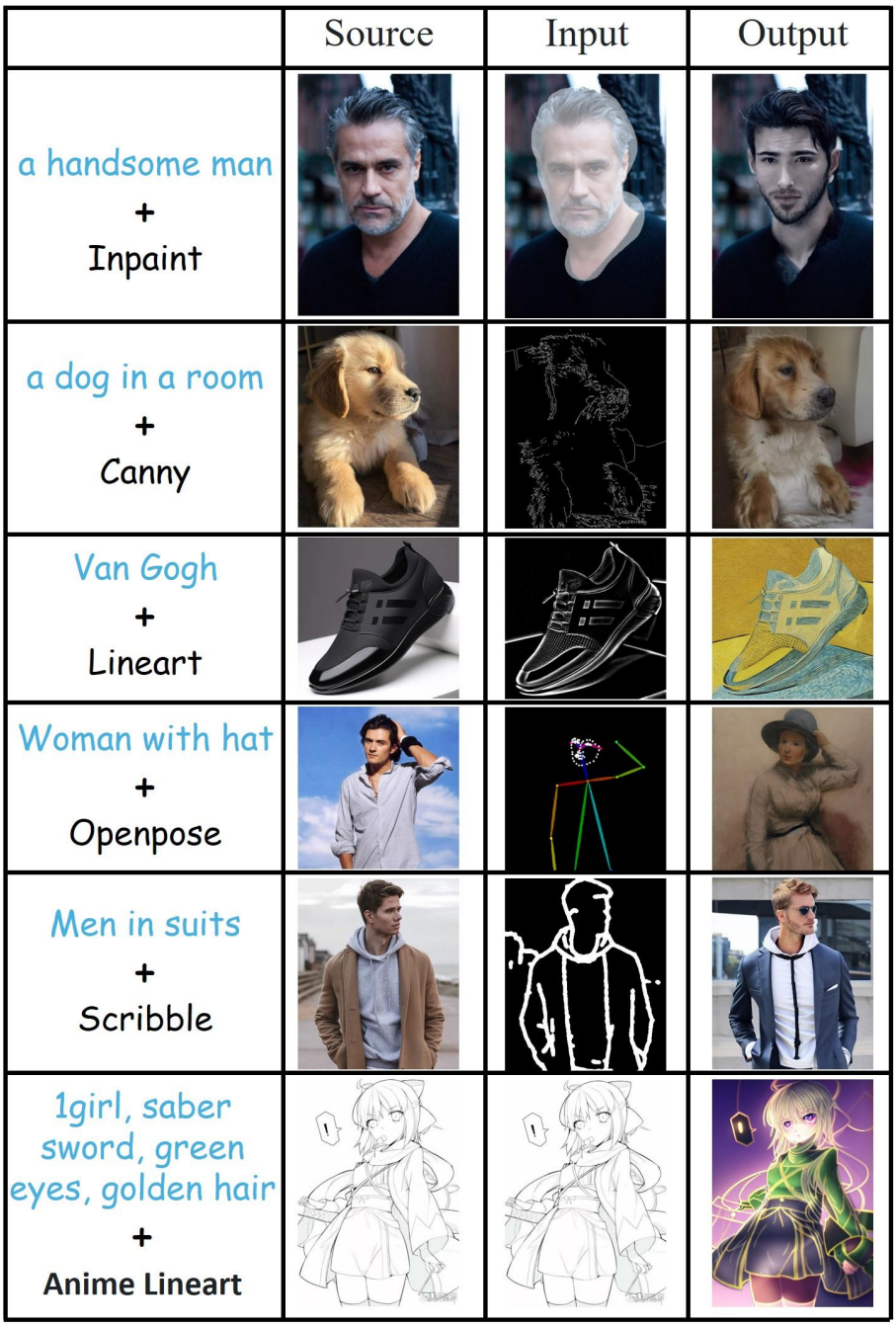


Figure 3.6: Examples of ControlNet results with various conditional inputs, including Inpaint, Canny, Lineart, OpenPose, Scribble, and Anime Line-art. These conditions enable diverse and precise control over the image generation process.

These examples demonstrate ControlNet's remarkable versatility and adaptability. The ability to condition the diffusion process with multimodal inputs allows for a wide range of creative and practical applications. For instance, edge-to-image generation can assist in artistic design, depth-to-image synthesis can aid in realistic scene rendering, and pose-to-image generation is particularly valuable for character design and animation. The seamless integration of these diverse modalities highlights the robustness and flexibility of ControlNet as an extension of Stable Diffusion, making it a powerful tool for controlled and specialized image synthesis.

3.5.1 ControlNet Scribble

ControlNet Scribble is a powerful feature of the ControlNet framework that uses freehand sketches as guiding inputs for image generation. Unlike other conditional inputs, such as edge maps or depth maps, scribbles provide a highly intuitive and flexible way for users to define the structure and layout of the desired image. This functionality proves especially useful in contexts requiring rapid conceptualization and prototyping of visual ideas, establishing it as an essential tool in both artistic and scientific workflows.

The core functionality of ControlNet Scribble lies in its ability to interpret rough, hand-drawn sketches as structural constraints during the diffusion process. By injecting these sketches into the intermediate layers of the U-Net architecture, ControlNet ensures that the generated image adheres closely to the provided outline while simultaneously producing realistic textures, colors, and details. This approach bridges the gap between rough drafts and polished outputs, enabling precise and controlled image synthesis from highly abstract inputs.

Chapter 4

Proposed Model

To solve the challenge of maintaining both stylistic fidelity and semantic accuracy in OBI-style image generation, this thesis proposes DiffOBI, a two-stage framework that integrates modern generative diffusion models with refinement techniques, which represents a significant advancement in generating OBI-style images. The first stage of the framework, the generation module, is built upon ControlNet, which incorporates conditional inputs such as text prompts and object images into the diffusion process. This ensures that the generated images align with both the stylistic characteristics of OBIs and the contextual requirements specified by the user. The second stage, the refinement module, applies a suite of post-processing techniques, including binary optimization, impurity filtration, edge refinement, and resolution enhancement. These processes collaboratively enhance both the visual quality and structural precision of the generated images, refining initial outputs into well-defined results. As illustrated in Figure 4.1, this iterative refinement ensures that the final images adhere closely to the stylistic and structural characteristics of OBIs.

4.1 Generation Module

The generation module is the key contribution of DiffOBI, leveraging the capabilities of ControlNet [8] to integrate conditional controls into the image generation process. ControlNet can enhance the functionality of Stable Diffusion by introducing additional inputs, such as text descriptions and reference images, which guide the diffusion process and ensure that the generated outputs meet specific stylistic and contextual requirements.

ControlNet operates by freezing the weights of the pretrained model $N(x;\Theta)$, where x represents the input features and Θ are the model pa-

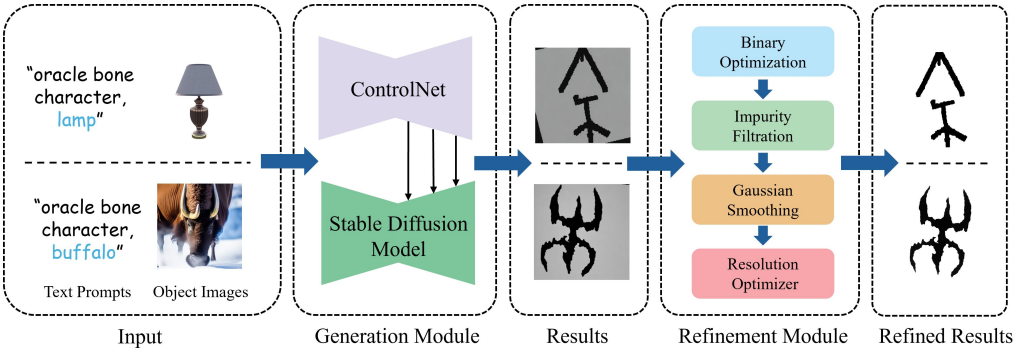


Figure 4.1: Overview of the DiffOBI framework. The two-stage process includes a generation stage using ControlNet for conditional OBIs image synthesis and a refinement stage to enhance visual quality and structural accuracy.

rameters. A trainable copy $N(x;\Theta_c)$ is introduced, with Θ_c representing the trainable parameters adapted for task-specific conditions. Zero convolution layers, represented as $Z(c;\Theta_z)$, are used to incorporate control conditions c into the model. The final output y_c is calculated as follows:

$$y_c = N(x;\Theta) + Z\left(N\left(x + Z(c;\Theta_{z1});\Theta_c\right);\Theta_{z2}\right), \tag{4.1}$$

where Θ_{z1} and Θ_{z2} are the parameters of the zero convolution layers responsible for injecting and processing the control conditions.

During the generation process, text prompts are encoded into latent representations τ_θ, which capture the semantic context of the desired output. Simultaneously, reference object images are encoded into latent features c_f, providing structural guidance. These inputs are combined with random latent noise z_t, which is iteratively refined by the diffusion model:

$$z_{t-1} = \phi_\theta(z_t, t, c_f, \tau_\theta), \tag{4.2}$$

where z_t represents the noisy latent state at time step t, ϕ_θ is the denoising function parameterized by θ, and t denotes the timestep. This iterative process ensures that the generated image transitions from noise to a coherent and detailed OBI-style representation. By the end of the diffusion process, the latent representation z_0 is decoded into an initial OBI-style image.

The generation module is designed to capture the intricate patterns and textures characteristic of OBIs. By incorporating additional controls, such as reference images and text prompts, DiffOBI provides the flexibility to tailor the outputs to specific artistic or contextual requirements. This module forms the foundation for producing high-quality initial results, which are further refined in the subsequent module.

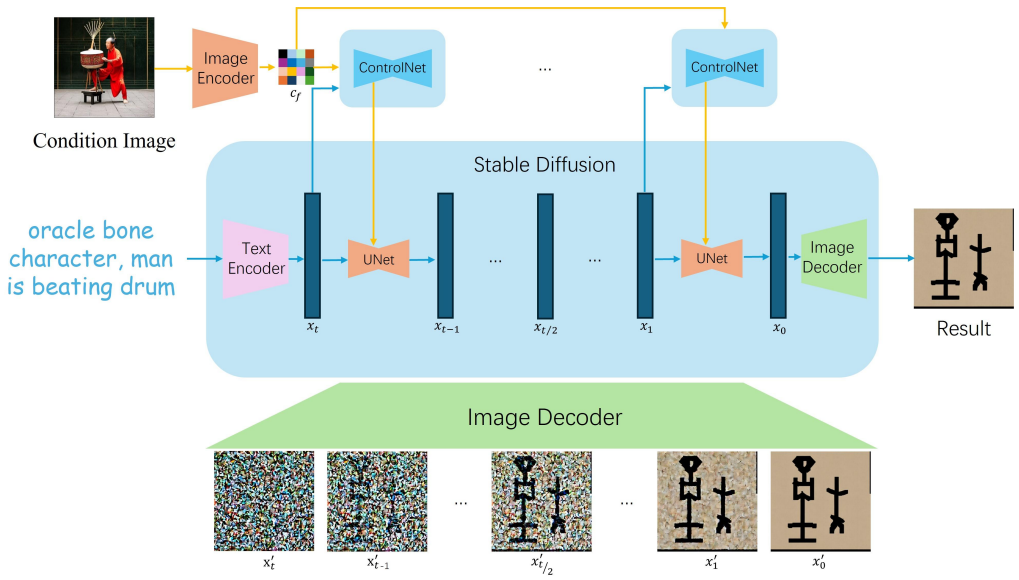


Figure 4.2: The generation module of DiffOBI utilizes ControlNet to integrate textual prompts and conditional image inputs into the stable diffusion. This process ensures the production of preliminary OBI-style images with structural and semantic guidance.

4.2 Refinement Module

As shown in Figure 4.3, the refinement module applies sequential processing steps, including binary optimization, impurity filtration, Gaussian smoothing, and resolution optimization. Each step improves the quality and clarity of the OBI-style images, with intermediate results visualized for better understanding. The refinement module addresses the limitations and imperfections of the initial results produced by the generation module. This module employs a series of post-processing techniques to enhance the visual quality, structural integrity, and overall appeal of the generated images.

The first step in the refinement process is binary optimization. This step separates the OBIs patterns from the background using a pixel-based binarization technique. Each pixel value $P(x, y)$ in the image is compared to a predefined threshold T and transformed as follows:

$$P'(x,y) = \begin{cases} 0, & \text{if } P(x,y) < T, \\ 255, & \text{if } P(x,y) \geq T. \end{cases} \tag{4.3}$$

Here, $P(x, y)$ is the original pixel intensity, $P'(x, y)$ is the binarized intensity, and T is the threshold value. After testing multiple thresholds, $T = 50$ was

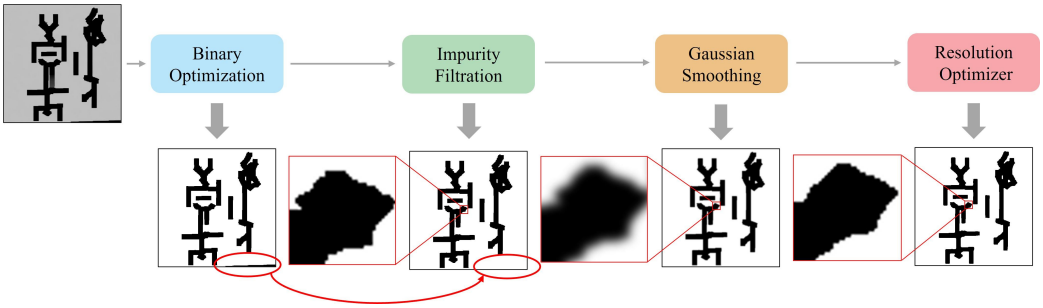


Figure 4.3: The refinement module applies a sequence of post-processing steps: binary optimization separates patterns from the background, impurity filtration removes artifacts, Gaussian smoothing enhances edges, and resolution optimization improves clarity. Each step's output is shown alongside a zoomed-in view of a specific region.

found to provide the optimal balance between detail preservation and noise reduction in this work.

Impurity filtration removes artifacts such as dark corners introduced during training. This process is based on a flood-filling algorithm implemented as:

$$\text{if } I(y,x) < 50, \text{ then floodFill}(I, M, (x,y), 255, \text{loDiff} = 0, \text{upDiff} = 100), \tag{4.4}$$

where $I(y,x)$ is the pixel intensity at coordinates (x,y), M is the mask array, and (x,y) is the seed pixel. The flood fill algorithm begins at a specified seed point and replaces connected pixels that satisfy the intensity difference constraints defined by `loDiff` and `upDiff`. The seed point is set to four vertices, and the replacement value is set to 255 (white), ensuring that unwanted dark regions are effectively removed, resulting in a cleaner image.

Edge refinement is implemented using Gaussian smoothing to enhance the continuity and coherence of the OBIs patterns. The Gaussian function is defined as:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{4.5}$$

where $G(x,y)$ represents the Gaussian kernel value at position (x,y), and σ controls the degree of smoothing. This operation reduces pixel-level irregularities while preserving the overall structure of the inscriptions, ensuring smooth and visually appealing edges.

Resolution enhancement is achieved through bilateral filtering, which combines spatial and range kernels to smooth intensity differences while pre-

serving edges. The filtering process is expressed as:

$$I^*(p) = \frac{1}{W_p} \sum_{p_i \in S} I(p_i) \cdot f_r(|I(p_i) - I(p)|) \cdot f_s(|p_i - p|), \tag{4.6}$$

where $I(p)$ and $I^*(p)$ represent the original and filtered intensities, p_i denotes neighboring pixels, S is the spatial domain, f_r is the range kernel controlling intensity similarity, f_s is the spatial kernel controlling proximity, and W_p is a normalization factor ensuring the sum of the weights is equal to 1.

Combining all the above functions together, these refinement techniques transform the preliminary results into polished and high-quality OBI-style images. By addressing common artifacts and imperfections, the refinement module ensures that the outputs are visually consistent, structurally accurate, and contextually appropriate. This comprehensive approach makes DiffOBI a reliable and versatile tool for generating OBI-style images for applications in cultural heritage preservation, artistic reinterpretation, and historical research.

Chapter 5

Dataset Construction

The construction of a novel dataset forms the cornerstone of this research, addressing the limitations of existing datasets in the domain of OBIs. Previous datasets, while instrumental in advancing the study of OBIs, have notable shortcomings. They often lack meaningful contextual alignment between OBIs and real-world representations, making it difficult to leverage modern generative models for stylistically faithful and semantically coherent image generation. Another significant limitation is the absence of detailed textual explanations for deciphered OBIs. This motivated the creation of a dataset that bridges these gaps, enabling conditional generation by aligning OBIs with textual prompts and real-world representations.

To construct the dataset, the EVOBC dataset [6] was utilized, a rich repository containing diverse OBIs and their associated meanings. A total of 44 different groups of OBIs were selected from this resource, each consisting of 3-5 images, resulting in 180 unique OBIs. These groups were carefully chosen to represent a wide range of glyph styles and semantic contexts, ensuring comprehensive coverage of both pictorial and textual diversity.

For each oracle bone inscription, textual meanings were researched by consulting authoritative books and verified online resources. These meanings were then paired with their corresponding OBIs images to create semantic-text prompts. Using these text prompts and the OBIs images as inputs, a pre-trained ControlNet Scribble model was employed to generate aligned images. This approach enabled the synthesis of object images that remain faithful to the semantic essence and stylistic characteristics of the original OBIs. Figure 5.1 illustrates this process, showcasing the workflow for generating aligned images from OBIs data and their associated meanings.

The generated dataset comprises 180 pairs, each consisting of the original OBIs image, its textual meaning as a prompt, and the corresponding aligned image generated by ControlNet. To facilitate training and evaluation, the

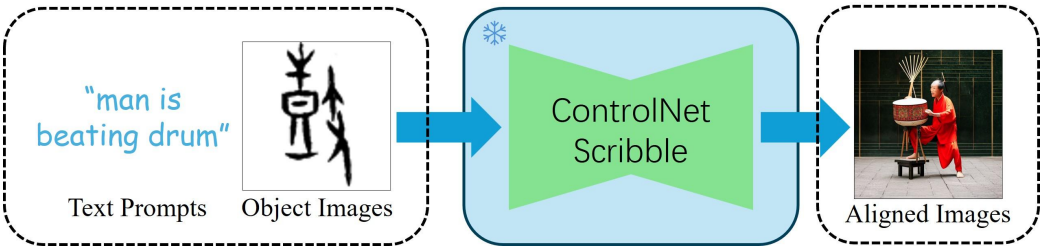


Figure 5.1: Framework for generating aligned images from original OBIs and their textual meanings using ControlNet. The pipeline ensures semantic and stylistic alignment between OBIs and their generated representations.

dataset was divided into two subsets: 150 pairs were allocated to the training set, and 30 pairs were reserved for the test set. This stratified division ensures a robust experimental setup, allowing for both model training and the assessment of generalization performance.

To provide an overview of the dataset, a subset of the training and testing images is shown in Fig. 5.2. This visualization highlights the diversity of the dataset, showcasing the alignment between OBIs, their textual prompts, and the generated images.

By constructing this dataset, the critical need for aligning OBIs with real-world representations is addressed, paving the way for improved conditional generative modeling. The dataset's combination of original OBIs, textual prompts, and aligned images may play a significant advancement in facilitating research on OBIs and their stylistic and semantic transformations.

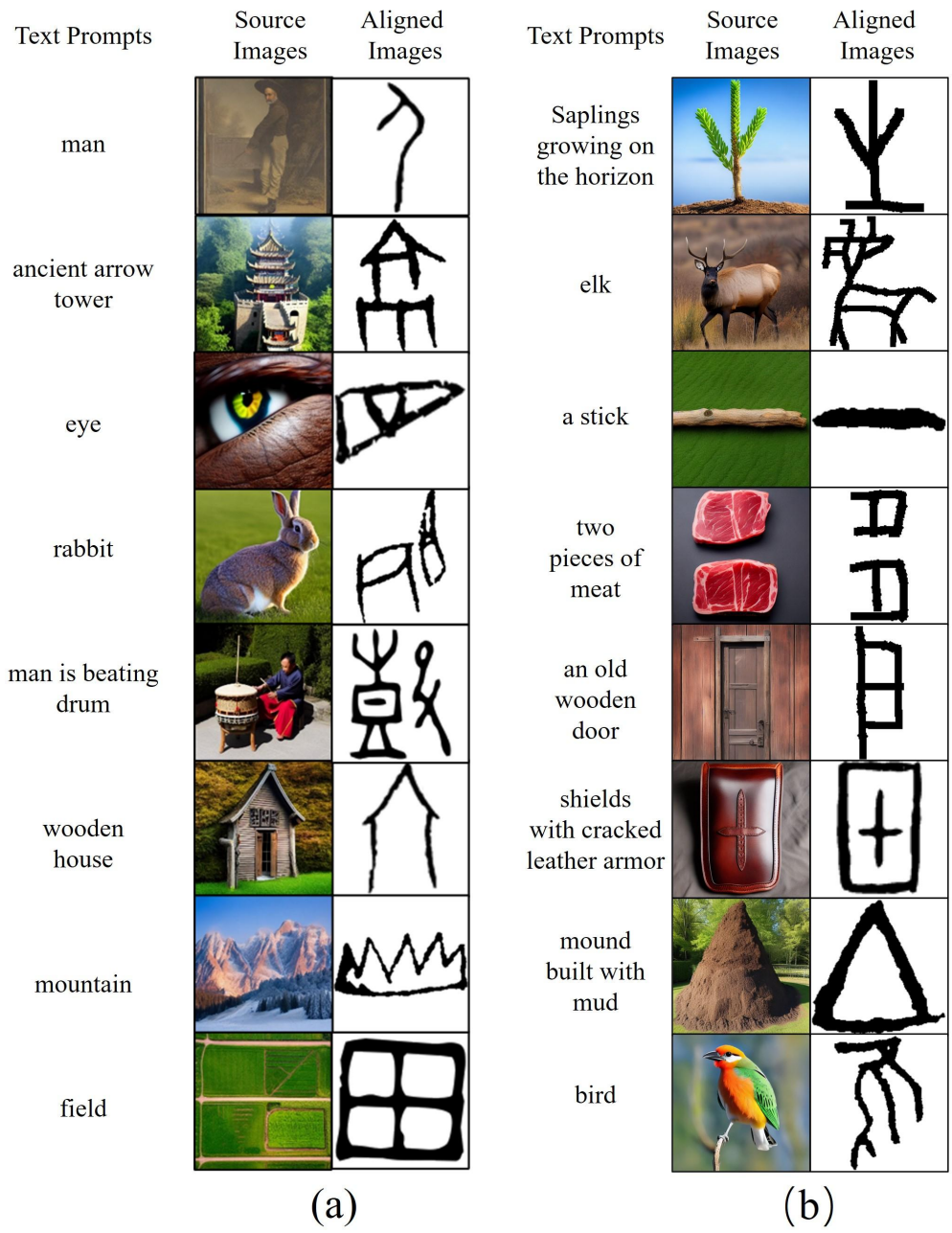


Figure 5.2: Sample images from the training set (a) and test set (b) of the constructed dataset, demonstrating the diversity of OBIs and its consistency with textual prompts and the aligned images generated from ControlNet.

Chapter 6

Results and Evaluation

In this chapter, the implementation of the generative model is detailed, focusing on training strategies, data augmentation, and the computational setup used to achieve high-quality results. The performance of the proposed DiffOBI framework is evaluated through both qualitative and quantitative analyses. The qualitative evaluation, presented in Section 6.2, examines the model's ability to reconstruct existing oracle bone characters, generate novel characters, and produce diverse stylistic variations, emphasizing its fidelity to the aesthetic and semantic features of OBIs. The quantitative evaluation, detailed in Section 6.3, compares the proposed model against baseline approaches, including pix2pix[17], CycleGAN[25], and IP-Adapter[26], using metrics such as Fréchet Inception Distance (FID), CLIP Image-Image Similarity (CLIP-I), and Neural Image Assessment (NIMA). Furthermore, Section 6.4 highlights the results of a user preference study, providing insights into the subjective evaluations of image quality, stylistic fidelity, and semantic consistency from participants. Finally, these evaluations demonstrate the robustness and versatility of DiffOBI in generating high-quality OBI-style images while preserving their cultural and artistic significance.

6.1 Implementation Details

The process of training ControlNet to generate OBI-style images involved multiple stages to ensure alignment between the textual and visual characteristics of the dataset. During the training phase, the previously generated aligned images were used as source images, while the corresponding OBIs served as target images. The textual meanings associated with the "Oracle Bone Character" were combined with the original text prompts to create enhanced prompts, improving the semantic relevance and diversity of the

training data.

The stable diffusion model (v2-1-512-ema-pruned) was employed as the foundational architecture for training the ControlNet. This pre-trained model provided a robust generative framework, enabling the integration of OBIs characteristics while retaining the semantic fidelity of the aligned images. To address the challenges posed by the limited dataset size, data augmentation techniques were applied, including random horizontal and vertical flips, rotations in multiples of 45°, random cropping, and resizing operations. Additionally, random adjustments to brightness, contrast, saturation, and tone were introduced to create variations in lighting and color distribution. These augmentations not only increased the effective size of the training dataset but also improved the model's generalization to diverse visual conditions. Figure 6.1 presents a sample of the augmented training dataset, where the top row displays real-world aligned images, and the bottom row shows the corresponding OBIs.

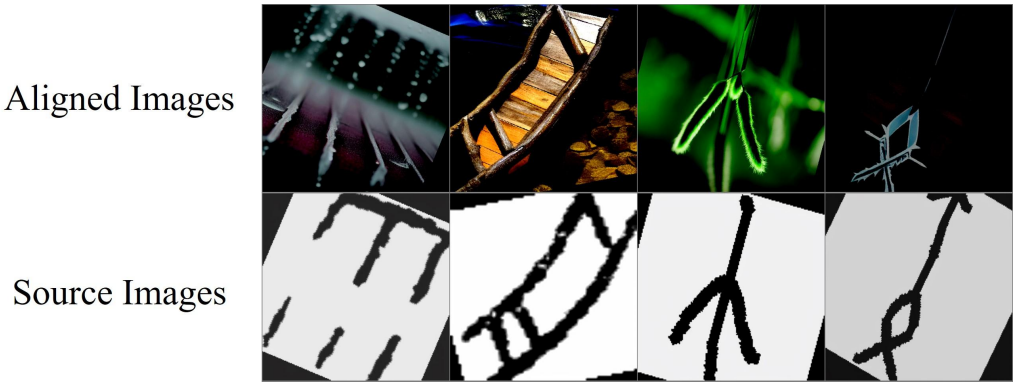


Figure 6.1: Sample images from the training dataset showcasing both the generated aligned real-world images and their corresponding oracle bone inscriptions (OBIs).

To ensure optimal learning, the model underwent 1,500 epochs of training, with a batch size of 4, leveraging the computational power of an NVIDIA A100 GPU.

6.2 Qualitative Evaluation

In this section, a qualitative evaluation of the proposed DiffOBI framework is presented, focusing on three critical aspects: the reconstruction of existing characters, the generation of novel characters, and the diversity of styles in the generated outputs. These evaluations demonstrate the versatility and

fidelity of the model in producing stylistically consistent and semantically accurate OBIs.

6.2.1 Reconstruction of Existing Characters

To evaluate the model's performance in reconstructing existing characters, its outputs were compared with those from IP-Adapter[26], pix2pix[17], and CycleGAN[25]. As shown in Fig. 6.2, the proposed model effectively preserves both the semantic integrity and stylistic characteristics of OBIs. This high fidelity in reconstruction highlights the model's ability to retain essential features of the original inscriptions.

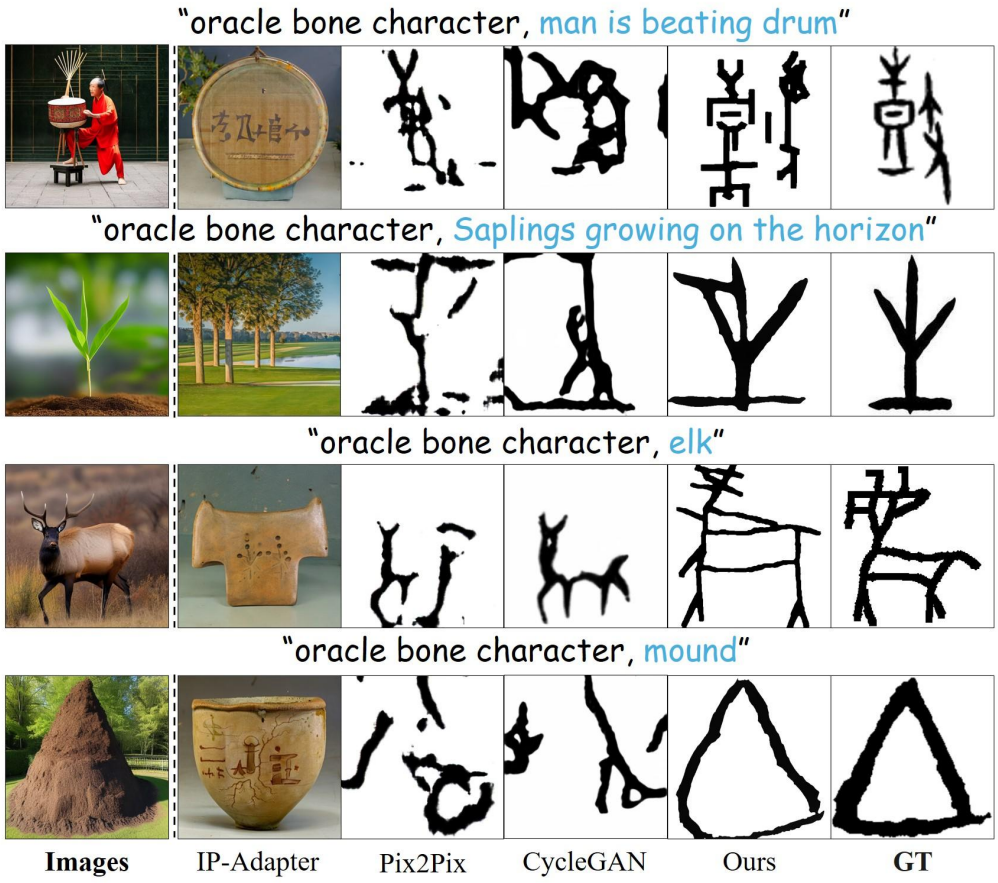


Figure 6.2: Results of generated images of objects (first row from training dataset, the others from the test dataset).

6.2.2 Generation of Novel Characters

The ability of the proposed model to generate novel characters not present in the original dataset was also assessed, such as "bike", "fuji", "ice cream", "Eiffel Tower", "ice coffee". The results, depicted in Fig. 6.3, showcase the model's capacity to produce outputs that adhere to the stylistic norms of OBIs while introducing new semantic content. This demonstrates the model's potential to expand the creative possibilities of OBIs generation by synthesizing unique objects that align with traditional aesthetics.

6.2.3 Diversity of Styles in Generated Outputs

The diversity of styles in the generated outputs was evaluated by examining the model's ability to produce stylistic variations for the same item using identical prompts. For example, variations were explored for items such as "oracle bone character, lamp." As shown in Fig. 6.4, the model successfully captured both the semantic essence and stylistic diversity of the inputs. These results highlight the model's capability to generate outputs that are not only diverse but also consistent with the traditional OBIs aesthetic.

6.3 Quantitative Evaluation

To assess the performance of the proposed model, a quantitative evaluation was conducted using three widely recognized metrics: Fréchet Inception Distance (FID), CLIP Image-Image similarity (CLIP-I), and Neural Image Assessment (NIMA). These metrics were used to compare the proposed model with IP-Adapter[26], pix2pix[17], and CycleGAN[25].

Table 6.1: Quantitative comparison results.

	pix2pix	CycleGAN	IP-Adapter	Ours
FID (↓)	367.39	316.90	404.08	**249.63**
CLIP-I (↑)	0.79	0.76	0.56	**0.81**
NIMA (↑)	4.34	4.75	4.48	**5.03**

The Fréchet Inception Distance (FID) metric quantifies the distance between the generated images and the original dataset in terms of feature distributions, with lower scores indicating higher visual fidelity. As shown in

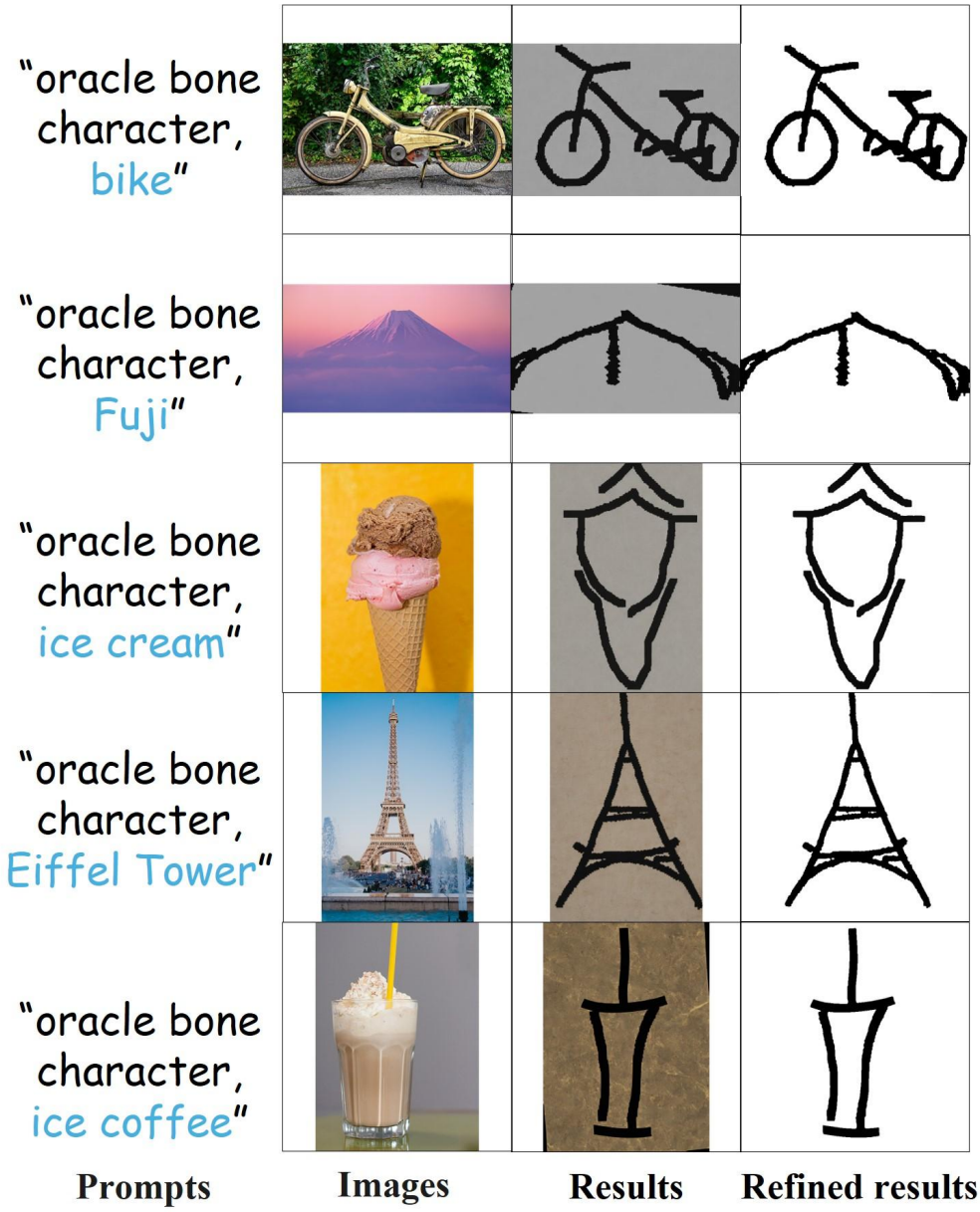


Figure 6.3: Results of generated images of objects not present in the original dataset.

Table 6.1, the proposed model achieved the lowest FID score of 249.63, significantly outperforming pix2pix (367.39), CycleGAN (316.90), and IP-Adapter (404.08), highlighting the superior visual quality of the generated images.

The CLIP Image-Image Similarity (CLIP-I) metric measures the semantic similarity between the generated and reference images, where higher scores

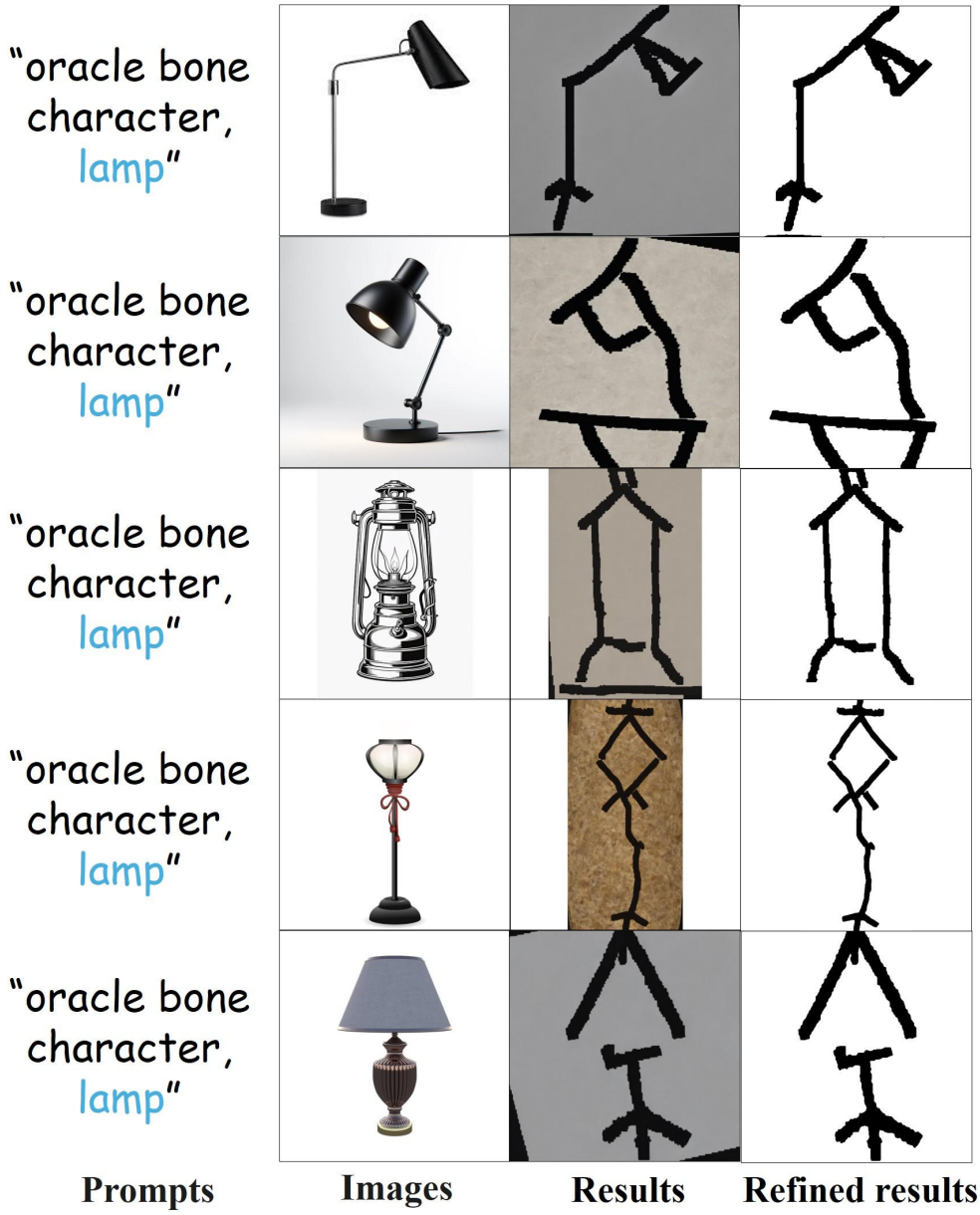


Figure 6.4: Examples of generated images showing stylistic variation within a single category.

indicate better alignment with the input prompts and reference images. The proposed model achieved the highest CLIP-I score of 0.81, surpassing pix2pix (0.79), CycleGAN (0.76), and IP-Adapter (0.56). This result demonstrates the model's capability to produce semantically coherent images that closely

align with the intended input descriptions.

The Neural Image Assessment (NIMA) metric evaluates the aesthetic quality of the generated images on a scale from 1 to 10, with higher scores reflecting better visual presentation. The proposed model achieved a NIMA score of 5.03, outperforming pix2pix (4.34), CycleGAN (4.75), and IP-Adapter (4.48), indicating that the generated images exhibit the highest aesthetic quality among the compared methods.

The quantitative evaluation demonstrates the superiority of the proposed model across all three metrics. By achieving the lowest FID score, highest CLIP-I score, and best NIMA score, this approach establishes a new benchmark for generating high-quality, semantically coherent, and aesthetically refined images in the domain of OBIs generation.

6.4 User Preference Study

To further assess the effectiveness of the proposed model, a user preference study was conducted, comparing it with three other generative models: IP-Adapter (based on diffusion models), pix2pix, and CycleGAN (both utilizing GAN structures). Each model was trained using the same dataset to ensure a fair evaluation. The study aimed to assess two key criteria: fidelity to the oracle bone inscription (OBI) style and the similarity between the original and generated OBI-style images. The questionnaire used for this study can be found in the Appendix.

Table 6.2: The comparison results from user preference study.

	pix2pix	CycleGAN	IP-Adapter	Ours
Selection rate/%	33.75	17.75	4.50	**44.00**
Average score	2.46	3.11	1.69	**3.54**

Initially, 100 participants were introduced to the distinctive morphology and characteristics of OBIs, as detailed in Appendix 7.2. This familiarization process ensured that they could effectively assess the generated images. Participants were then presented with a set of 16 images (four from each model) and asked to select the four images that best represented the OBI-style, as outlined in Appendix 7.2. The proposed model was selected in 44.00% of the choices, significantly outperforming pix2pix (33.75%), CycleGAN (17.75%), and IP-Adapter (4.50%). This strong preference underscores the model's superior ability to capture the essence of OBIs.

To further validate the results, participants rated the similarity between the generated OBI-style images and their corresponding originals on a scale of 1 (very poor) to 5 (very good), as described in Appendix 7.2. The proposed model achieved the highest average score of 3.54, surpassing pix2pix (2.46), CycleGAN (3.11), and IP-Adapter (1.69). These ratings, summarized in Table 6.2, confirm that this approach not only adheres to the stylistic norms of OBIs but also maintains a closer resemblance to the original images.

These results highlight the effectiveness of the proposed model in generating high-quality OBI-style images. By achieving the highest selection rate and average similarity score, this approach demonstrates its potential for practical applications in digital art and cultural preservation.

Chapter 7

Conclusion

In this thesis, DiffOBI, a novel two-stage framework for generating OBI-style images using diffusion models, was proposed. The approach ensures stylistic accuracy and semantic relevance by leveraging a dataset comprising original OBIs images, corresponding real-world objects, and descriptive textual prompts. The refinement module further enhances the initial outputs, aligning them more closely with the traditional structure and norms of OBIs. Experimental results demonstrate the high visual quality and strong user preference for the generated images, underscoring the effectiveness of the proposed method.

Despite these achievements, DiffOBI faces certain challenges, particularly in generating multi-object images and maintaining stylistic consistency. Addressing these limitations in future work will further enhance the capabilities of OBI-style image generation. Ultimately, DiffOBI represents a step forward at the intersection of cultural heritage preservation and generative AI, offering new possibilities for both artistic and academic exploration.

7.1 Limitations

While the DiffOBI framework demonstrated significant advancements in generating OBI-style images, certain limitations remain. One of the primary challenges lies in generating multi-object images as shown in Fig 7.1, where "man" is missing in the upper case and only "house" is generated in its corresponding position. In the lower case, "man" and "dog" are interpreted as one object resulting in only one OBI-style image being generated as well. This limitation arises from the single-object-focused nature of the training data, which constrains the model's ability to seamlessly synthesize complex compositions. Additionally, variations in brushstrokes and stylistic inconsistencies

can lead to discrepancies between the generated images and the ground truth. These issues highlight the need for further refinement in handling stylistic diversity and structural complexity inherent to OBIs.

Figure 7.1: Example of a failure case generating inconsistent results for the number of targets in the multi-target case.

7.2 Future Work

To address the limitations identified in this study, several avenues for future work can be considered. First, improving the stability of multi-object generation is a key objective. This involves leveraging the newly developed dataset, which integrates semantic and contour-based segmentation tools to facilitate the segmentation and annotation of multi-object compositions. As shown in Fig. 7.2, this dataset includes progress on multi-object segmentation, displaying original characters (first column), segmented regions (subsequent columns), and annotations for each segmented part. These advancements lay the groundwork for enhancing the model's ability to generate complex multi-object OBIs compositions.

Additionally, Additionally, exploring suitable evaluation metrics that better quantify the diversity and fidelity of generated OBI-style is a key objective. Metrics tailored to capture the subtleties of brushstroke variations and stylistic coherence will be essential for evaluating future iterations of DiffOBI.

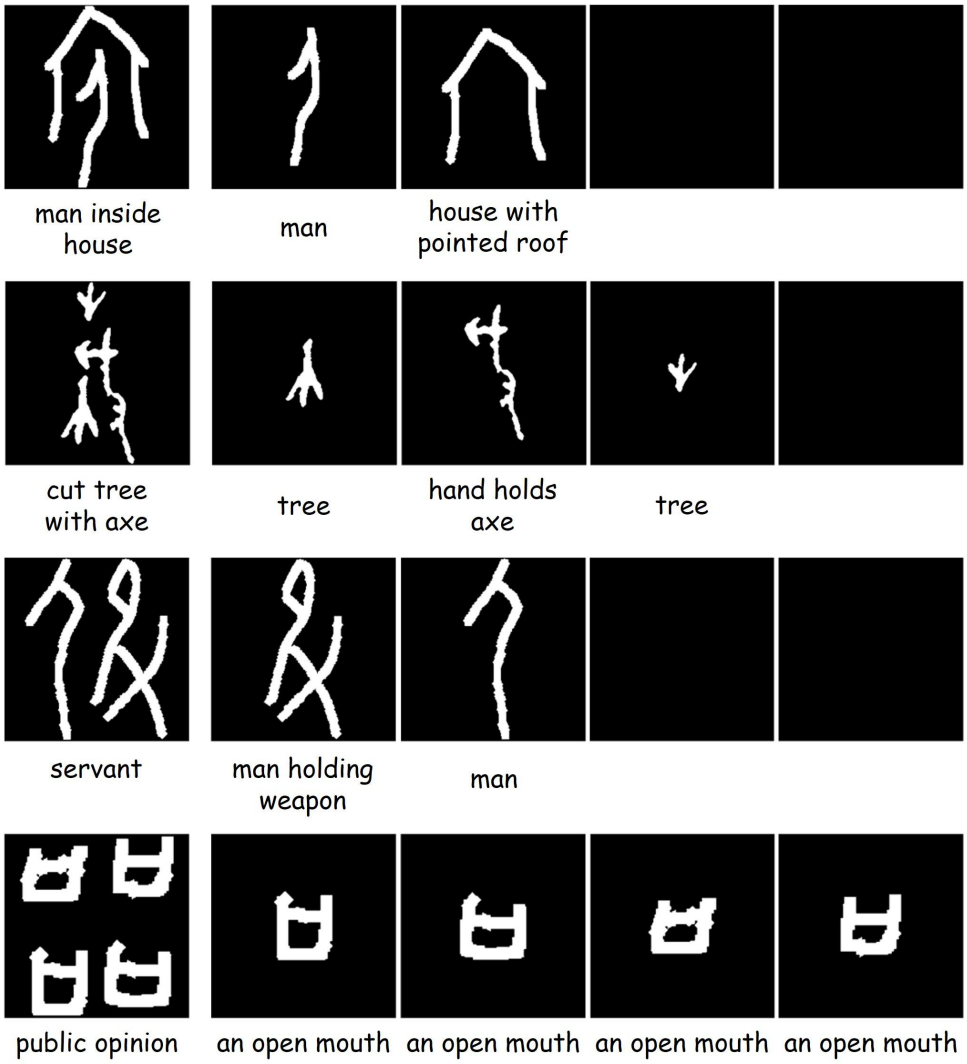


Figure 7.2: Visualization of the multi-object dataset progress, including original characters (first column), segmented regions (subsequent columns), and their respective annotations.

Furthermore, incorporating the historical evolution of OBIs characters into the generation process presents an intriguing research direction. By modeling the transformations of OBIs characters over time, the cultural and historical significance of the generated outputs can be further enriched.

Acknowledgement

I would like to begin by expressing my deepest gratitude to my supervisor, Prof. Haoran Xie, for his invaluable guidance and unwavering support throughout my research journey. Prof. Xie's profound knowledge and insightful feedback have significantly shaped the direction and quality of this work. His emphasis on rigor and clarity has instilled in me a deeper appreciation for academic excellence and meticulous research.

I am also sincerely grateful to all the members of our lab for their collaboration and support. Their constructive feedback during lab meetings and willingness to share ideas have greatly contributed to this research. Special thanks to my peers who provided assistance and encouragement, creating a collaborative and motivating environment.

Finally, I wish to extend my heartfelt thanks to my family, whose encouragement and unwavering support have been a constant source of strength and inspiration. Their belief in my abilities has been instrumental in helping me navigate the challenges of this journey.

References

[1] X. Yue, H. Li, Y. Fujikawa, and L. Meng, "Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition," *ACM Journal on Computing and Cultural Heritage*, vol. 15, no. 4, pp. 1–20, 2022.

[2] B. Li, D. Luo, Y. Liang, J. Yang, Z. Ding, X. Peng, B. Jiang, S. Han, D. Sui, P. Qin *et al.*, "Oracle bone inscriptions multi-modal dataset," *arXiv preprint arXiv:2407.03900*, 2024.

[3] P. Wang, K. Zhang, X. Wang, S. Han, Y. Liu, J. Wan, H. Guan, Z. Kuang, L. Jin, X. Bai *et al.*, "An open dataset for oracle bone character recognition and decipherment," *Scientific Data*, vol. 11, no. 1, p. 976, 2024.

[4] S. Huang, H. Wang, Y. Liu, X. Shi, and L. Jin, "Obc306: A large-scale oracle bone character recognition dataset," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 681–688.

[5] B. Li, Q. Dai, F. Gao, W. Zhu, Q. Li, and Y. Liu, "Hwobc-a handwriting oracle bone character recognition database," in *Journal of Physics: Conference Series*, vol. 1651, no. 1. IOP Publishing, 2020, p. 012050.

[6] H. Guan, J. Wan, Y. Liu, P. Wang, K. Zhang, Z. Kuang, X. Wang, X. Bai, and L. Jin, "An open dataset for the evolution of oracle bone characters: Evobc," *arXiv preprint arXiv:2401.12467*, 2024.

[7] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," *arXiv preprint arXiv:2301.07093*, 2023.

[8] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[9] M. L. Huang, R. Zhao, J. Hua, Q. V. Nguyen, W. Huang, and J. Wang, "Designing infographics/visual icons of social network by referencing to the design concept of ancient oracle bone characters," in *2020 24th International Conference Information Visualisation (IV)*. IEEE, 2020, pp. 694–699.

[10] Y. Chen and S. A. Sharudin, "Research on the application of chinese traditional carved symbols in cultural and creative product design," *International Journal of Education and Humanities*, vol. 13, no. 2, pp. 92–95, 2024.

[11] C. Bazerman, *Handbook of research on writing: History, society, school, individual, text*. Routledge, 2009.

[12] J. Li, Q.-F. Wang, K. Huang, X. Yang, R. Zhang, and J. Y. Goulermas, "Towards better long-tailed oracle character recognition with adversarial data augmentation," *Pattern Recognition*, vol. 140, p. 109534, 2023.

[13] J. Li, Q.-F. Wang, K. Huang, R. Zhang, and S. Wang, "Diff-oracle: Diffusion model for oracle character generation with controllable styles and contents," 2023.

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

[15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[16] J. Guo, C. Wang, E. Roman-Rangel, H. Chao, and Y. Rui, "Building hierarchical representations for oracle character and sketch recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 104–118, 2015.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[19] X. Chang, F. Chao, C. Shang, and Q. Shen, "Sundial-gan: A cascade generative adversarial networks framework for deciphering oracle bone inscriptions," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1195–1203.

[20] M. Wang, Y. Cai, L. Gao, R. Feng, Q. Jiao, X. Ma, and Y. Jia, "Study on the evolution of chinese characters based on few-shot learning: From oracle bone inscriptions to regular script," *Plos one*, vol. 17, no. 8, p. e0272974, 2022.

[21] X. Fu, R. Zhou, X. Yang, and C. Li, "Detecting oracle bone inscriptions via pseudo-category labels," *Heritage Science*, vol. 12, no. 1, p. 107, 2024.

[22] Z. Guo, Z. Zhou, B. Liu, L. Li, Q. Jiao, C. Huang, and J. Zhang, "An improved neural network model based on inception-v3 for oracle bone inscription character recognition," *Scientific Programming*, vol. 2022, no. 1, p. 7490363, 2022.

[23] X. Xie, X. Du, M. Li, X. Yang, and H. Xie, "Diffobi: Diffusion-based image generation of oracle bone inscription style characters," in *SIGGRAPH Asia 2024 Technical Communications*, 2024, pp. 1–4.

[24] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020.

[26] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023.

[27] Y. Fujikawa, H. Li, X. Yue, C. Aravinda, G. A. Prabhu, and L. Meng, "Recognition of oracle bone inscriptions by using two deep learning models," *International Journal of Digital Humanities*, vol. 5, no. 2, pp. 65–79, 2023.

[28] X. Fu, Z. Yang, Z. Zeng, Y. Zhang, and Q. Zhou, "Improvement of oracle bone inscription recognition accuracy: A deep learning perspective," *ISPRS International Journal of Geo-Information*, vol. 11, no. 1, p. 45, 2022.

[29] C. Zhang, R. Zong, S. Cao, Y. Men, and B. Mo, "Ai-powered oracle bone inscriptions recognition and fragments rejoining," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 5309–5311.

[30] F. Gao, J. Zhang, Y. Liu, and Y. Han, "Image translation for oracle bone character interpretation," *Symmetry*, vol. 14, no. 4, p. 743, 2022.

[31] Y. Gao, Y. Guo, Z. Lian, Y. Tang, and J. Xiao, "Artistic glyph image synthesis via one-stage few-shot learning," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 6, pp. 1–12, 2019.

[32] Y. Zhang, Y. Zhang, and W. Cai, "Separating style and content for generalized style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8447–8455.

[33] S. Park, S. Chun, J. Cha, B. Lee, and H. Shim, "Few-shot font generation with localized style representations and factorization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2393–2402.

[34] Y. Xie, X. Chen, L. Sun, and Y. Lu, "Dg-font: Deformable generative networks for unsupervised font generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5130–5140.

[35] C. Wang, M. Zhou, T. Ge, Y. Jiang, H. Bao, and W. Xu, "Cf-font: Content fusion for few-shot font generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1858–1867.

[36] W. Liu, F. Liu, F. Ding, Q. He, and Z. Yi, "Xmp-font: Self-supervised cross-modality pre-training for few-shot font generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7905–7914.

[37] A. Egorov, M. Egorova, and T. Orlova, "The use of a comparative analysis of the connection between ancient and modern chinese languages in the process of teaching students chinese characters," in *Proceedings of the 2nd International Conference on Education: Current Issues and Digital Technologies (ICECIDT 2022)*. Atlantis Press, 2022, pp. 10–19. [Online]. Available: https://doi.org/10.2991/978-2-494069-02-2_3

[38] L. Tang, Y. Cai, J. Liu, Z. Hong, M. Gong, M. Fan, J. Han, J. Liu, E. Ding, and J. Wang, "Few-shot font generation by learning fine-grained local styles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7895–7904.

[39] M. Jaweed and R. F. Shaikh, "Optimizing generative ai by overcoming stability mode collapse and quality challenges in gans and vaes," *MSW Management Journal*, vol. 34, no. 2, pp. 497–507, 2024.

[40] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[41] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang *et al.*, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[43] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[44] A. Radford, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[45] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 70–85.

[46] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 391–408.

[47] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[48] J. Tan, L. Jing, Y. Huo, L. Li, O. Akin, and Y. Tian, "Lgan: Lung segmentation in ct scans using generative adversarial network," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101817, 2021.

[49] S. Nema, A. Dudhane, S. Murala, and S. Naidu, "Rescuenet: An unpaired gan for brain tumor segmentation," *Biomedical Signal Processing and Control*, vol. 55, p. 101641, 2020.

[50] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014.

[51] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[52] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2745–2754.

[53] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville, "Pixelvae: A latent variable model for natural images," *arXiv preprint arXiv:1611.05013*, 2016.

[54] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," *arXiv preprint arXiv:1711.01558*, 2017.

[55] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in neural information processing systems*, vol. 33, pp. 19 667–19 679, 2020.

[56] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[57] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

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

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

[60] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[61] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," *arXiv preprint arXiv:2210.11427*, 2022.

[62] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine *et al.*, "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[63] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "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*, 2023, pp. 22 500–22 510.

[64] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022.

[65] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

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

[67] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4328–4343, 2022.

[68] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 981–17 993, 2021.

[69] H. Chung and J. C. Ye, "Score-based diffusion models for accelerated mri," *Medical image analysis*, vol. 80, p. 102479, 2022.

[70] Y. Yang, H. Fu, A. I. Aviles-Rivero, C.-B. Schönlieb, and L. Zhu, "Diffmic: Dual-guidance diffusion network for medical image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 95–105.

[71] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang, "Geodiff: a geometric diffusion model for molecular conformation generation," 2022. [Online]. Available: https://arxiv.org/abs/2203.02923

[72] B. Jing, G. Corso, J. Chang, R. Barzilay, and T. Jaakkola, "Torsional diffusion for molecular conformer generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 240–24 253, 2022.

[73] H. Huang, L. Sun, B. Du, Y. Fu, and W. Lv, "Graphgdp: Generative diffusion processes for permutation invariant graph generation," in *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022, pp. 201–210.

[74] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[75] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.

[76] H. Chang, Y. Peng, S. Sato, and H. Xie, "Sketch-guided flow field generation with diffusion model," in *International Workshop on Advanced Imaging Technology (IWAIT) 2024*, vol. 13164. SPIE, 2024, pp. 290–295.

[77] T. Zhang and H. Xie, "Sketch-guided text-to-image generation with spatial control," in *2024 2nd International Conference on Computer Graphics and Image Processing (CGIP)*, 2024, pp. 153–159.

[78] Y. Zhang, T. Zhang, and H. Xie, "Texcontrol: Sketch-based two-stage fashion image generation using diffusion model," 2024. [Online]. Available: https://arxiv.org/abs/2405.04675

[79] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[80] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4296–4304.

[81] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.

[82] G. Yuan, "Observing the traces and defining the book: The inheritance and development of xia and shang scripts in the archaeological perspective," 2021. [Online]. Available: https://epaper.gmw.cn/gmrb/html/2021-05/12/nw.D110000gmrb_20210512_1-11.htm

[83] C. C. C. I. Communication and Z. U. Education Research Center, *Spring and Autumn of Oracle Bones - Commemorating the 120th Anniversary of the Discovery of Oracle Bones*, 2019.

[84] Xinhua, "Chinese oracle bone inscriptions added to unesco world memory register," 2017. [Online]. Available: http://english.scio.gov.cn/chinavoices/2017-12/27/content_50168640.htm

Appendix

User Study Questionnaire

To evaluate the quality of OBI-style images generated by different models, participants completed a structured questionnaire divided into three parts: an introduction to OBI, an image selection task, and a similarity rating task. The detailed questionnaire is provided below:

Part 1: Introduction

Participants were introduced to the distinctive morphology and stylistic characteristics of oracle bone inscriptions (OBIs). The purpose of this introduction was to provide them with sufficient background to assess the generated images accurately. Figure A.1 illustrates typical examples of OBI glyphs.

Part 2: Image Selection Task

Participants were presented with multiple sets of generated images, with four images per set corresponding to different generative models. For each set, participants were asked to select the image that best represented the OBI-style.

Questionnaire 1: Select the image that best represents the OBI-style from the following options:

- (a) Image 1
- (b) Image 2
- (c) Image 3
- (d) Image 4

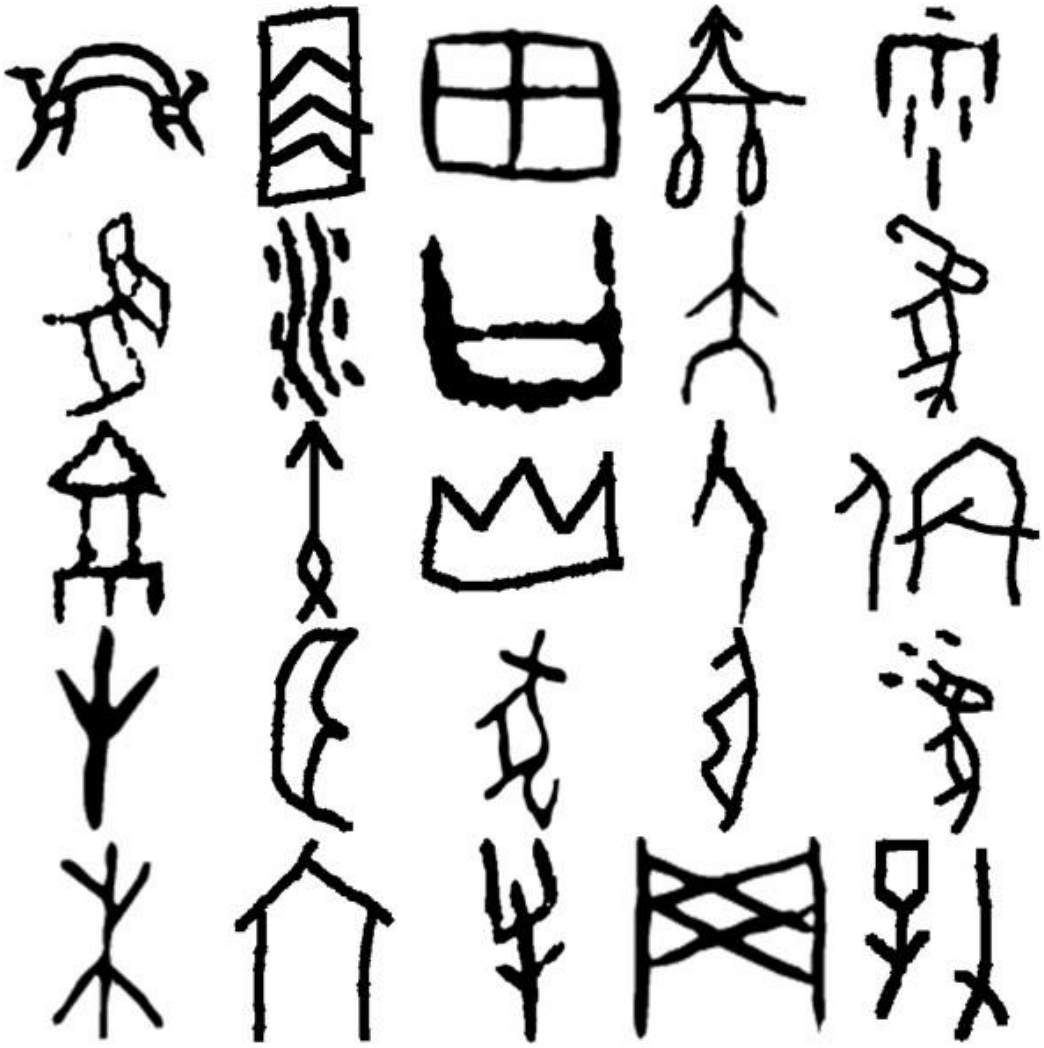


Figure A.1: Examples of oracle bone inscriptions (OBI) style. These images were shown to participants to highlight the structural and stylistic characteristics of OBIs.

To ensure unbiased evaluation, the order of the options (a, b, c, d) was randomized for each participant and each question. This randomization prevents participants from developing systematic biases toward specific positions.

The problem was repeated for 5 sets of images. Each set of images consisted of one image generated by the proposed model and three images generated by other models. Figure A.2 shows all the images in the problem.

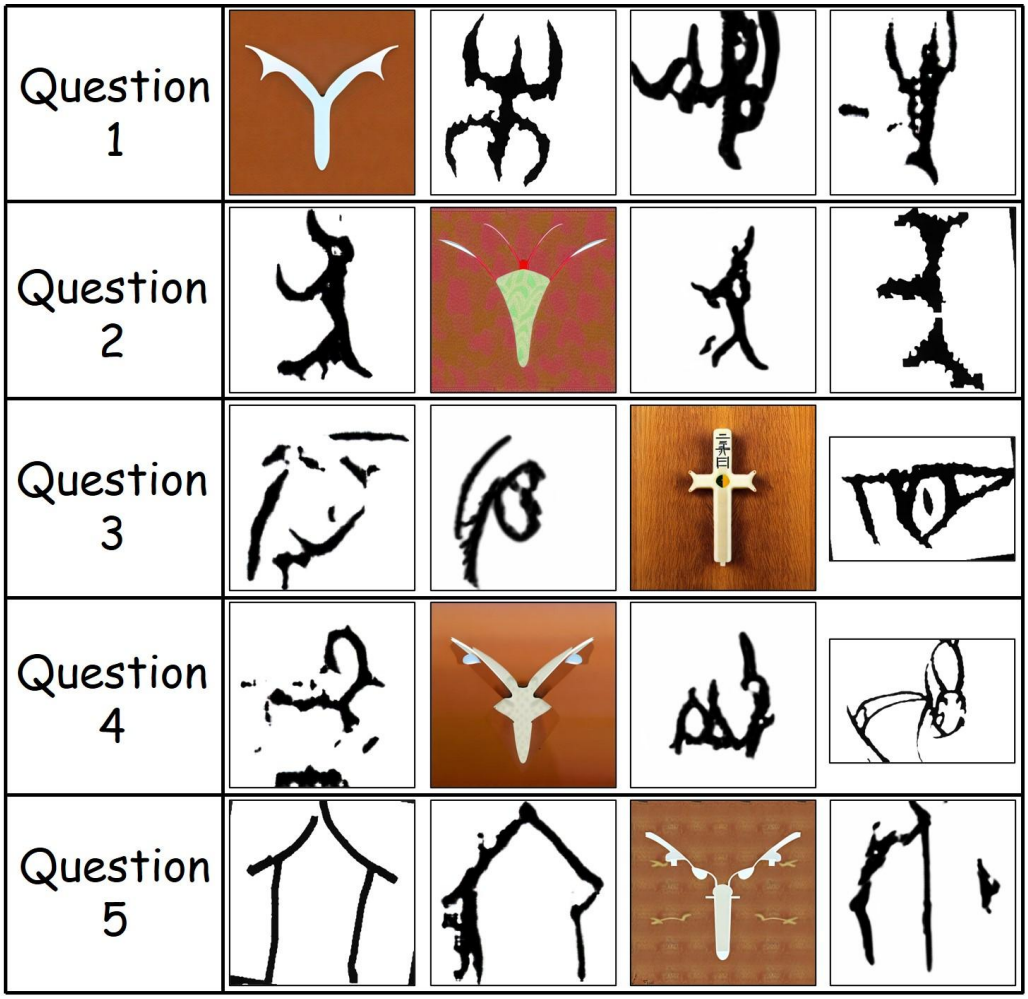


Figure A.2: Question for the Image Selection Task. Participants were asked to select the image that best represented the OBI-style.

Part 3: Similarity Rating Task

Participants were asked to evaluate the similarity between the generated OBI-style images and the original input images (e.g., object or textual inputs). The task involved rating the similarity on a scale of 1 to 5, where 1 represents "very poor" and 5 represents "excellent."

Questionnaire 2: Rate the similarity between the generated OBI-style image and the original input image based on the following criteria:

- 1: Very Poor
- 2: Poor

- 3: Fair
- 4: Good
- 5: Very Good

The problem was repeated for 5 sets of images. As shown in Figure A.3, "Refer" corresponds to "original input", "Options" corresponds to "generated OBI-style image".

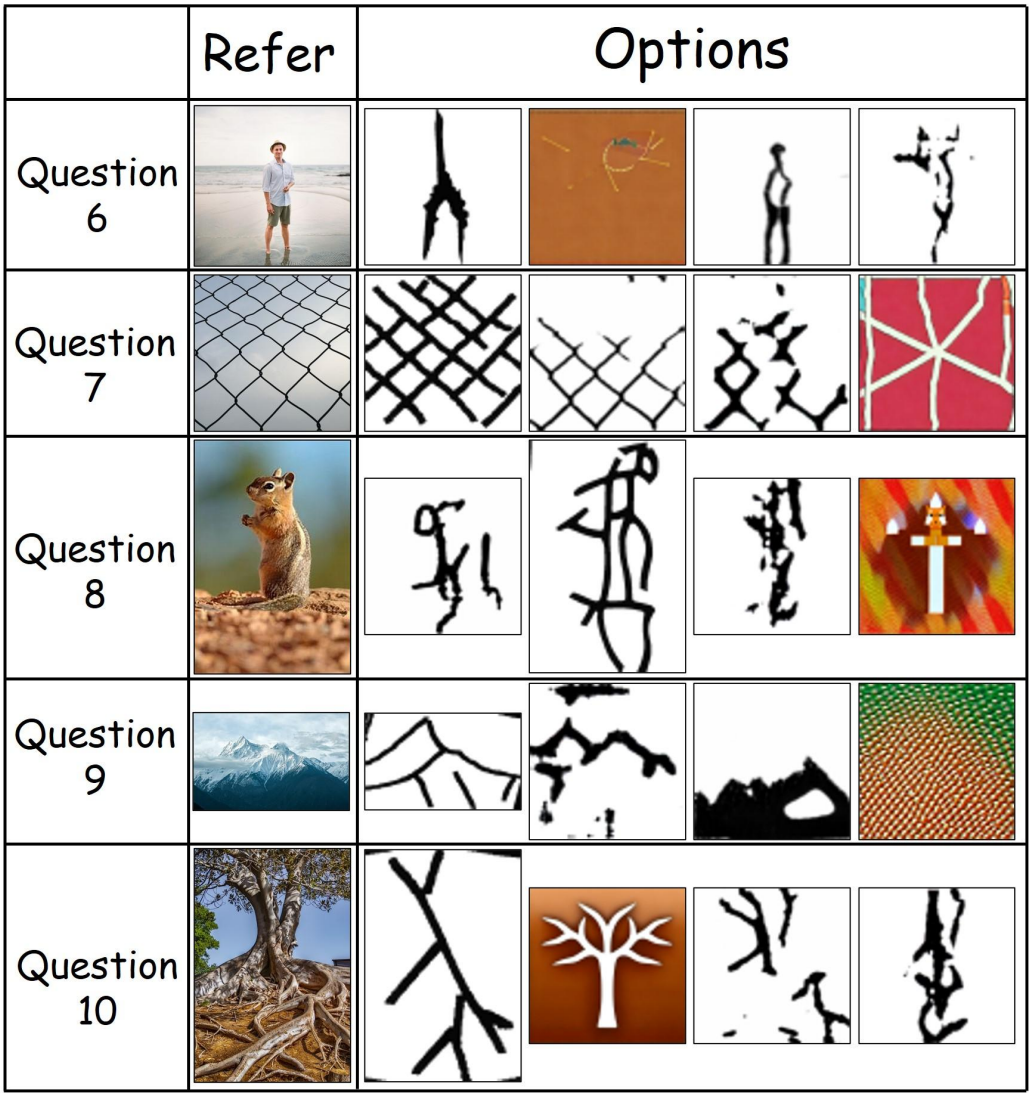


Figure A.3: Question of an Original and Generated Image Pair used for the Similarity Rating Task. Participants rated the similarity between the "Refer" and "Options" on a scale of 1 to 5.