Making Visual Sense of Oracle Bones for You and Me

Runqi Qiao[1] Lan Yang[1*] Kaiyue Pang[2] Honggang Zhang[1]
[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]SketchX, CVSSP, University of Surrey
{qrq, ylan, zhhg}@bupt.edu.cn k.pang@surrey.ac.uk

Abstract

Visual perception evolves over time. This is particularly the case of oracle bone scripts, where visual glyphs seem intuitive to people from distant past prove difficult to be understood in contemporary eyes. While semantic correspondence of an oracle can be found via a dictionary lookup, this proves to be not enough for public viewers to connect the dots, i.e., why does this oracle mean that? Common solution relies on a laborious curation process to collect visual guide for each oracle (Fig. 1), which hinges on the case-by-case effort and taste of curators. This paper delves into one natural follow-up question: can AI take over?

Begin with a comprehensive human study, we show participants could indeed make better sense of an oracle glyph subjected to a proper visual guide and its efficacy can be approximated via a novel metric termed TransOV (Transferable Oracle Visuals). We then define a new conditional visual generation task based on an oracle glyph and its semantic meaning and importantly approach it by circumventing any form of model training in the presence of fatal lack of oracle data. At its heart is to leverage foundation model like GPT-4V to reason about the visual cues hidden inside an oracle and take advantage of an existing text-to-image model for final visual guide generation. Extensive empirical evidence shows our AI-enabled visual guides achieve significantly comparable TransOV performance compared with those collected under manual efforts. Finally, we demonstrate the versatility of our system under a more complex setting, where it is required to work alongside with an AI image denoiser to cope with raw oracle scan image inputs (cf. processed clean oracle glyphs). Code is available at https://github.com/RQ-Lab/OBS-Visual.

1. Introduction

Not everyone can read art, but with proper guide, they might have a better chance. It is an experience that you and me, as museum and exhibition visitors, are so used to that generally takes place without notice - - to first form an initial impression of an art piece or installment, then habitually look for the information label nearby to seek for interpretation and finally with those information bearing in mind, to reappreciate the art with a deeper or different perspective. Lighthearted as this process might sound, these information labels, often varying across art forms and key to enabling public understanding and immersion of art, require joint hard efforts from both the curators and the artists themselves. As an AI practitioner, the immediate question becomes: can I program this? Not to completely drive humans of course, but to use AI to ease manual labour by generating guides from minimum cues provided by humans.

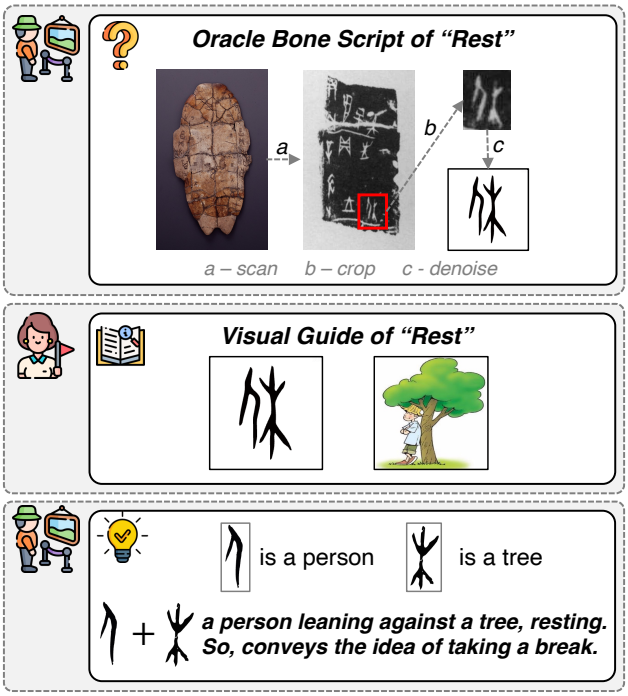

Figure 1. A well visual guide can effectively bridge the gap between modern viewers and the visual glyphs from 3,000 years ago.

* Corresponding author.

This paper does exactly that. In particular, we look at the visual guide problem in the domain of oracle bone scripts, a form of visual art that provides a unique peek into how people, before the advent of language, understand and record the visual world around them. Given an oracle glyph singled out and processed from the bone and their semantic meaning decoded by the experts, the guide is seen critical to enabling visual sense between the two and therefore getting to the end audience, *i.e.*, why this (oracle glyph) is that (semantic interpretation)? After all, with thousand years apart, vision changes and evolves. A few scribbles that seem rightfully meaningful to the prehistoric eyes might not be transferable to you and me (Fig. 2(c)).

We present a generative framework for acquiring oracle visual guides that provide essential cues to help connect abstract oracle glyphs with their modern semantics. Our ambition is simple: specify the oracle glyph and the semantics, and bingo, they are ready for display! AI takes over the hard example generation/curation part. Apart from being a practically useful case of AI for art, a successful framework of this kind has its scientific values on two fronts. First, it stepped onto the uncharted territory of a unique form of visual art, challenging the boundary of computer vision. Second, it contributes to the ongoing dialogue of expanding the benefits of generative foundation models (Stable Diffusion [26] or ControlNet [38]) for more downstream vertical tasks, especially the one with the dramatic level of visual abstraction present in oracles.

At the top of our to-do list is to first have a metric that can quantify the effect of a visual guide in enabling people towards a better understanding of an oracle - so that we are actually informed on the quality of any generation results designed later. To achieve this, we, in Sec. 3 conduct an extensive human study to evaluate the change of perspectives among 28 participants and 9434 trials when participants are offered a choice of visual guide. It follows that we design a metric, namely *TransOV*, that allows us to approximate these human results for a more scalable probe into this problem. *TransOV* is highly efficient, which in our 5-fold cross validations achieves 84.75% accuracy compared with those of human labels. We therefore adopt *TransOV* as our main evaluation tool throughout the empirical analysis.

On the outset, solving this new conditional visual generation task (oracle, text $\rightarrow$ image) would require a model no more complex than the ControlNet [38]. It just needs a text-to-image (T2I) agent with extra visual control. But in hindsight, the "text" and "visual" part along with an oracle is in no way comparable to the image caption and layout/edgemap visual that these existing T2Is adopt as inputs: the text is merely a character ("rest") with potentially multiple plausible visual scenes that explain it, and the visual resembles free-hand scribbles that are abstract and iconic. To address these challenges, we introduce *GenOV*, a comprehensive framework consisting of four distinct components tailored for oracle image generation. In "Textual Contextualisation", we enhance a single character into a detailed caption by leveraging the pre-trained language capabilities of a Vision-Language Model (VLM). Next, in "Visual Constraint", we address the challenge of the oracle glyph's abstraction by a coarse-to-fine layout generation. This is achieved through activating the VLM's spatial analysis ability. Following this, the "Candidates Generation" step involves utilising a foundational generation model, which, informed by the textual and visual conditions, produces extensive visual guide candidates. Lastly, in the "Guide Finalisation", we introduce an innovative method tailored to oracle. This method selects the most fitting visual guide by adeptly weighing and integrating both the semantic content and visual appearance of the oracle.

Our qualitative and quantitative experiments compare *GenOV's* outputs with those from generative models like ControlNet and Stable Diffusion, as well as expert-curated visual guides, evaluated using the *TransOV* metric. The results indicate that *GenOV* not only outperforms these foundational models but also achieves performance comparable to expert-curated guides. This is particularly promising for oracles lacking expert guides, as our method can rapidly generate effective visual guides. Additionally, *GenOV* includes a denoising capability, transforming raw, noisy scans of oracle bone scripts into clean glyphs, facilitating semantic lookup for previously unseen oracles.

In summary, our contributions are:

- Pioneering the use of AI generation models for oracle visual understanding and introducing *TransOV* (Transferable Oracle Visuals), a novel metric that quantitatively measures the impact of a visual guide on enhancing oracle comprehension.
- Proposing *GenOV*, a framework for interpreting highly abstract visual information and limited textual data from oracle bone scripts to generate appropriate visual guides.
- Demonstrating that our generated visual guides can match the quality of expert-curated images, showcasing *GenOV's* potential for oracles without existing expert guides. We also show that *GenOV* can handle more complex scenarios involving raw, noisy oracle glyphs.

2. Related Work

Oracle Character Processing. The task of processing oracle bones presents significant challenges within the computer vision community [6, 7, 9, 12, 13, 16, 19, 29, 30, 33, 37], due to the prevalence of noisy data, the scarcity of available datasets, and the labor-intensive nature of the annotation process. STSN [34] introduces a novel unsupervised domain adaptation network designed to facilitate knowledge transfer from hand-printed oracles to their scanned counterparts. Sundial-GAN [5] introduces a GAN-

based architecture that simulates the visual evolution of characters from ancient oracle bones to contemporary Chinese script, offering a novel framework for character transformation studies. To our knowledge, this work represents the first endeavor to create a visual guide for individual oracle bone scripts, aimed at aiding the general populace in understanding these ancient characters.

Conditional Image Generation. Generating high-quality images from different conditions has gained popularity in recent times [10, 11, 14, 15, 25–28, 31, 32, 35, 36, 38]. StableDiffusion [26] proposes a text-to-image generative model that operates by denoising in the latent space, followed by a decoding step that reconstructs the denoised latents into high-resolution images. LDM [18] proposes a two-stage approach for text-to-image generation: initially, a LLM produces a spatial layout from a user's prompt, which subsequently directs an off-the-shelf diffusion model to generate images anchored to the layout. Distinct from prior studies, the conditions of our method are a solitary character and an abstract glyph, rather than detailed image contexts. Our ambition is to go beyond the generation of high-fidelity image; we seek to create a visual guide that is not only of superior quality but also serves a pedagogical function, facilitating the understanding of ancient scripts for users.

3. Human Study

We conducted a rigorous human study, establishing that: (i) with proper guide, people have a better chance to read the oracle bone scripts; (ii) the effect of a visual guide in enabling people towards a better understanding of oracle can be quantified.

3.1. Data Preparation

Oracle Scripts. We crawled 364 oracle bone scripts from two reputable oracle websites [1, 2]. Each oracle $o_i|_{i=1,\cdots,X}$ is accompanied by a glyph image o_{gi} and its equivalent modern Chinese simplified character o_{si}. Moreover, from the website[2], we procured a textual description o_{ti} detailing the association between the oracle's glyph o_{gi} and its semantic o_{si}. Our oracle dataset is represented as $\mathcal{O} = \{o_i\}|_{i=1}^{364} = \{(o_{gi}, o_{si}, o_{ti})\}$.

Visual Guide. Drawing inspiration from the concept of "pictographic teaching", wherein children learn to recognise characters through associations with images, we aim to provide an apt visual guide to assist human comprehension. To secure a comprehensive and diverse visual guide for each oracle, we utilise two approaches to procure the visual guide: manual collection and AI generation. **Manual collection:** we extracted a singular reference image corresponding to each oracle script from the website [2]. **AI generation:** Conditional AI generation is now very powerful and capable of producing high-quality images, so we control them to output a wide variety of visual guides for each oracle by providing different text/visual conditions. Specifically, we utilise ControlNetv1.1 [38] as our conditional generative model $\mathcal{G}(\mathcal{T}^*, \mathcal{V}^*)$, concurrently accept both textual and visual conditions. The visual condition $\mathcal{V}^*$ is set to the glyph image o_{gi}, and to introduce variability, we implement a $condition\_scale$ with values of $\{1, 0.75, 0.5\}$. For textual stimuli $\mathcal{T}^*$, we examine three options: $\{o_{si}, o_{ti}, [o_{si}, o_{ti}]\}$. To further the heterogeneity of the generated images, GPT-4 [20] is harnessed to rewrite three semantically akin but distinct text prompts based on $\mathcal{T}^*$. Consequently, for each oracle, we yield a total of 28 visual guides[1] for subsequent human evaluation. The dataset employed in our subsequent human studies is denoted as $\mathcal{X} = \{x_i^k\}_{i=1,\cdots,X}^{k=1,\cdots,28} = \{(o_{gi}, o_{si}, o_{ti}, v_i^k)\}$.

3.2. Experiment Setup

Warm-up Study. Correctly associating oracle glyphs with their semantics is a significant challenge, especially for the uninitiated. In our pursuit of more accurate human-derived data, we commence with a foundational briefing for each participant. By showcasing a few representative o_is, as shown in Fig.2(a), our objectives are twofold:

(i) articulating the task's objectives and setting clear expectations regarding the desired output format,

(ii) leveraging these representative examples to illuminate typical thought processes and directions, thus enabling participants to adopt similar logical frameworks in their answers.

Two-stage Study. To better understand the effect of a visual guide, our study has been divided into two stages, as illustrated in Fig.2(b).

Stage 1, participants are shown an oracle glyph o_{gi} paired with its associated semantic character o_{si}. They are prompted to use their imagination to put themselves in the social context and daily routines of individuals living 3,000 years ago. They are then asked to analyse the possible reasoning behind using this oracle glyph to represent o_{si}. Participants are given the following response options:

- *Yes, I can*. [type your insight here]
- *Sorry, I cannot*. Remember, choosing this option will result in no payment.

If participants choose "*Sorry, I cannot*" they proceed to stage 2 of the study where they receive an additional visual guide v_i^k related to the oracle glyph.

Stage 2, participants will receive instructions similar to those in stage 1 with the added prompt: "Please refer to the additional visual guide we are providing to help envision the societal context and daily life from 3,000 years ago." Once again, participants will attempt to analyse why the oracle glyph represents o_{si}, and following options are provided:

[1]Here, X=1+3*3*3, accounting for 1 from manual collection, 3 from $\mathcal{V}^*$), 3 from $\mathcal{T}^*$, and 3 from GPT-4 rewrites.

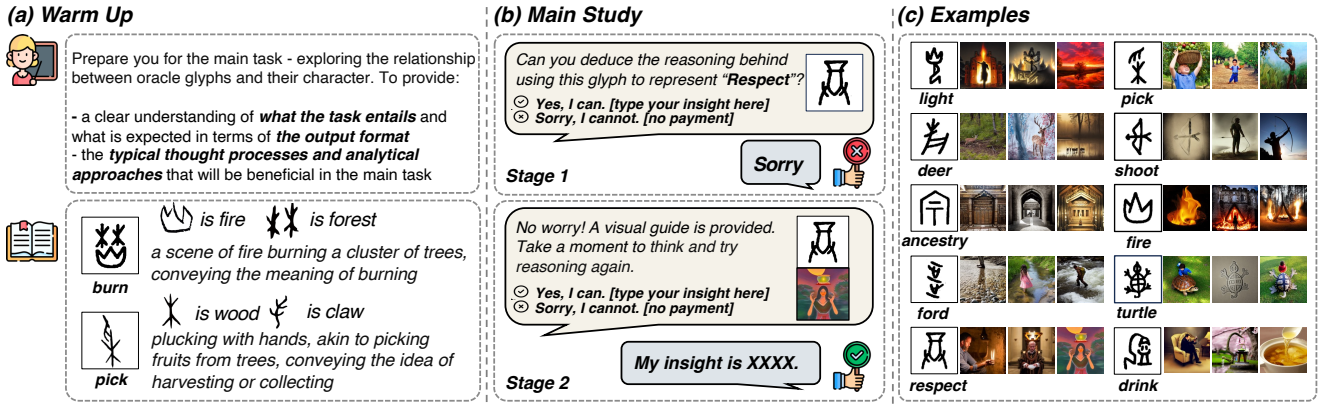

Figure 2. The overview of our human study: (a) Depicts the Warm-Up Stage, (b) Shows the pipeline of main human study, and (c) Presents randomly sampled examples of $\{o_{gi}, o_{si}, v_i^k\}$.

- *Yes, I can*. [type your insight here]
- *Sorry, I cannot*. Remember, choosing this option will result in no payment.

Participants. Each participant will iterate the entire $\mathcal{O}$. We ensured that at least 50 participants were assigned different visual guides in stage 2. To evaluate the consistency of participant responses, we included 10 sentinel trials where previously answered oracles were reintroduced. If a participant's answers differed from their original response for 8 or more sentinel trials, their entire response sheet was considered unreliable and discarded. Ultimately, 28 participants successfully passed the sentinel test.

3.3. Human Results

Reliability Verification. Before conducting any statistical evaluations, we conduct thorough checks on human responses to confirm validity. We focus primarily on instances where participants indicate that "*Yes, I can*". Human responses are denoted as t_i^l, where l refers to the stage (where $l = 1$ for stage one and $l = 2$ for stage two). We use the binary decision-making abilities of GPT-4 to determine if the responses are reliable. The template prompt reads: "*From a semantic perspective, please select the option that best matches [t_i^l]: Option 1: [o_{ti}], Option 2: [o_{tj}]*", where o_{ti} corresponds to the correct description of the relationship between the oracle glyph o_{gi} and its semantic character o_{si}, and o_{tj} is a random incorrect option selected from the remaining 363 alternatives in $\{o_{tj}\}_{j=1,\cdots,364}^{j\neq i}$. We present this question 5 times with different incorrect choices. t_i^l is considered reliable when the ground-truth option o_{ti} is chosen 4 or more times. After this reliability verification. After validating process, we remain 713 "Yes" respones and 9434 "Sorry" in stage 1 and similarly, 2897 "Yes" responses and 6537 "Sorry" in stage 2.

Quantitative Analysis. During stage 1, nearly 87% (specifically 316/364) of o_is had more than 23[2] responses of "*Sorry, I cannot*" from different participants. This outcome aligns with our hypothesis, highlighting the formidable challenge that humans face when attempting to discern the connection between oracle glyphs and their associated semantic meanings. Out of the 9434 "*Sorry, I cannot*" responses in stage 1, 30.7% (2897/9434) changed their answer to "*Yes, I can*" after the visual guide was introduced in stage 2. This shift led to a dramatic drop in the proportion of incomprehensible oracles, from 87% (316/364) to 28% (101/364). Examples of these oracles can be seen in Fig.2(c). This is indeed an encouraging result, as it suggests that with carefully designed visual guides, we can significantly enhance the comprehensibility of oracle glyphs. It propels us to further investigate the creation of such aids, potentially unlocking new ways for individuals to grasp these ancient characters.

4. Methodology

In this section, we first propose *TransOV*, a novel metric designed to quantitatively assess the efficacy of visual guides for oracles. Subsequently, we introduce *GenOV*, a framework based on Vision-Language Models (VLMs) tailored for the generation of customised visual guides for oracle glyphs.

4.1. TransOV

To ascertain a more appropriate visual guide for each oracle, a quantifiable metric for evaluating its efficacy is essential. As detailed in Sec. 3.3, negative responses ("Sorry, I cannot") in stage 1 from participants signal the need for additional support. For these trials, a follow-up positive ("Yes,

[2]A large majority (23/28, more than 80%) of the participants struggled to understand the link between o_{gi} and o_{si}, which supports that the general population may also find it challenging to understand o_i.

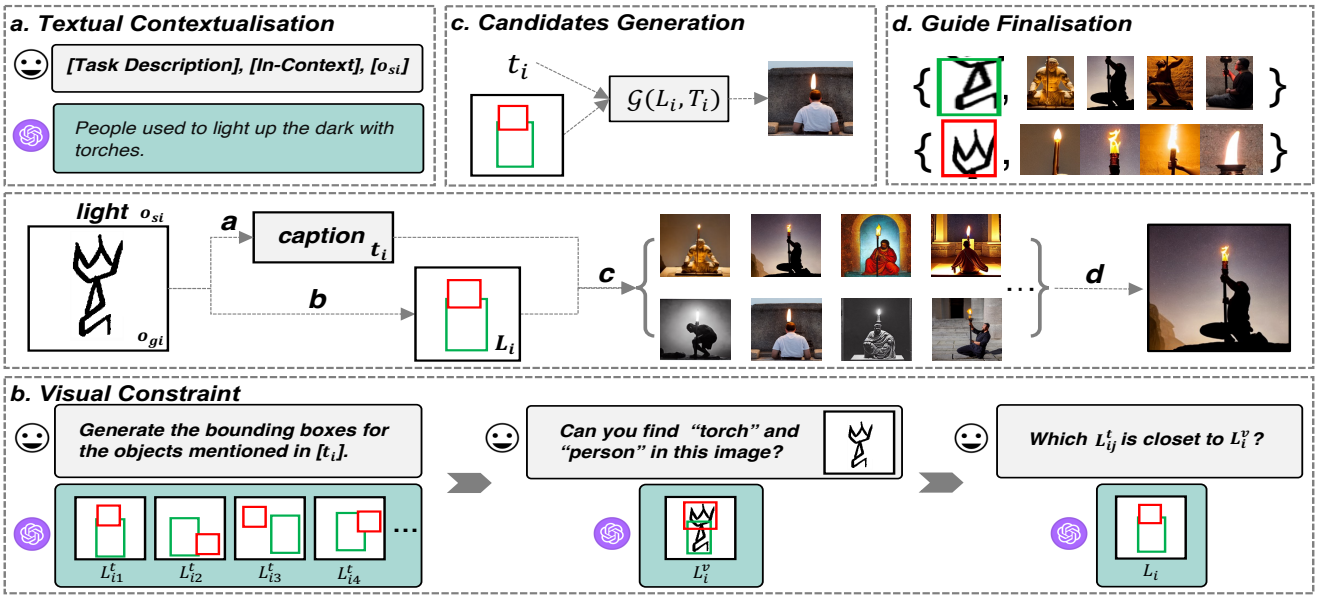

Figure 3. Overview of we proposed *GenOV*: *GenOV* begins with an oracle glyph o_{gi} and its associated semantic character o_{si} as inputs, culminating in the generation of a visual guide. The process encompasses four phases: (a) Textual Contextualisation, where the semantic context of o_{si} is expanded; (b) Visual Constraint, focusing on creating a spatial layout for the visual guide that aligns with the appearance of o_{gi}; (c) Candidate Generation, involving the production of multiple visual guide candidates for o_i using varying random seeds; d) Guide Finalisation, employing an oracle-specific selection method to identify the most effective visual guide that harmonises both the appearance and semantics of o_i. Grey rectangles represent the prompts provided to Ψ, while green rectangles indicate Ψ's outputs.

I can") or continued negative response after viewing the visual guide can be directly used as a binary label of the effectiveness of the given visual guide. Leveraging these labels, it becomes feasible to train a binary classification network aimed at predicting the effectiveness of the provided visual guides in enhancing the understanding of oracles.

Each trial is recorded as $\{o_{gi}, o_{si}, v_i^k, y_i^k\}$, where $i \in [1, 364]$ indicates the index of the oracle bone scripts, $k \in [1, 28]$ denotes the index of the visual guide associated with an oracle, $y_i^k = 1$ signifies that the visual guide v_i^k facilitates the comprehension of the connection between o_{gi} and o_{si}, and conversely. The human labelled data is partitioned into train, valid, and test sets with a ratio of $8:1:1$ according to o_{gi}. Specifically, the training set includes 291 oracles across 7834 trials, valid set contains 36 oracles with 800 trials, and test set consists of 37 oracles, also with 800 trials. Importantly, there is no overlap of oracles among the different data splits, guaranteeing a thorough evaluation of the model's ability to generalise.

The *TransOV* (Transferable Oracle Visuals) metric has been devised to assess the effectiveness of visual guides in linking the visual information of oracle glyphs with their semantic interpretations. So we formulate *TransOV* by considering two primary aspects: Appearance and Semantic.

$$TransOV(o_i, v_i^k) = \Phi(\phi_v(o_{gi}, v_i^k), \phi_s(o_{si}, v_i^k)) \quad (1)$$

where ϕ_v denotes the appearance discrepancy between o_{gi} and v_i^k measuring by DINO [4] feature distance, ϕ_s represents the semantic gap between o_{si} and v_i^k determined through the CLIP [24] score, and Φ is defined as a two-layer MLP (details in the supplementary). The parameters of Φ are updated through training on the train set, with the optimal parameters being determined based on accuracy on the validation set.

The application of the trained *TransOV* metric to the test set yielded a high accuracy rate of 84.75%. This outcome verifies the alignment of the *TransOV* metric with human judgment in evaluating the impact of a visual guide. Consequently, we will employ the trained *TransOV* as the standard of evaluation in the subsequent sections.

4.2. GenOV

Grounded on the premise that an effective visual guide should ease the task of understanding the association between an oracle glyph (visual) and its corresponding semantic (textual) meaning, We propose *GenOV*, a framework that aims to optimally identify a visual guide v_i^* for a given oracle o_i. This task is treated as a dual-conditional image generation problem, balancing both visual and semantic aspects. The *GenOV* process is divided into four phases: Textual Contextualisation, Visual Constraint, Candidates Generation, and Guide Finalisation, as detailed in Fig. 3.

Textual Contextualisation. For an oracle o_i, its associated semantic, represented as o_{si}, is merely a single character. Although this appears to be a succinct representation, it actually holds a wealth of semantic detail, as oracle has been historically used in documenting divination scenes and outcomes. To derive richer semantic information from o_{si}, we utilise the extensive knowledge encapsulated in a large vision-language model, denoted as Ψ to expand its semantic. Typically, given a task description D_{cap}, a few examples C_{cap}, and the oracle semantic o_{si}, we can activate Ψ to generate an informative caption t_i for o_{si} by:

$$t_i = \Psi(o_{si}; \theta_{cap}) \tag{2}$$

where θ denotes the text prompt that guides Ψ, which comprises of D_{cap}, C_{cap} and o_{si}. t_i significantly enhances the semantic context of o_{si}, incorporating aspects like key entities within the scene and their interactions (Fig. 3(a)).

Visual Constraint. The design of the oracle glyph o_{gi} is inherently abstract, a characteristic inherent to the oracle bone script's symbolic nature. This abstraction poses challenges in directly employing o_{gi} for visual conditioning.

Drawing inspiration from the methodology presented in [18, 23], we leverage the spatial imagination and analytical abilities of the VLM Ψ for layout planning. Our approach involves prompting Ψ to produce a set of plausible, coarse-grained layouts L_i^ts based on the natural scene descriptions provided in the caption t_i. These generated layouts naturally align with contemporary social contexts and modern perceptions, thereby making them suitable for use as visual conditions in a pre-trained conditional diffusion model $\mathcal{G}$.

To ensure that the final layout is acceptable in both ancient and modern contexts, we utilise Ψ's visual grounding ability for object localisation. We extract pivotal objects from t_i and guide Ψ to mark their bounding box on o_{gi}, resulting in a visually-informed layout L_i^v. Given the high level of abstraction in o_{gi}, the produced bounding boxes may not be entirely accurate. Therefore, we incorporate an additional verification step, wherein Ψ selects the layout most similar to L_i^v from among the L_i^ts, denoted as L_i. The process of translating these abstract visual elements into a comprehensible layout is mathematically formulated as:

$$L_i = \Psi\{\Psi(o_{gi}, t_i; \theta_{loc}), \Psi(t_i; \theta_{coarse}); \theta_{fine}\} \tag{3}$$

Candidates Generation. Leveraging the textual condition t_i and the visual layout L_i, we engage an off-the-shelf conditional diffusion model, denoted as $\mathcal{G}$, to create diverse visual guide candidates, v_i^ks, for the given oracle o_i. To further increase the variety of v_i^ks, we employ Ψ to reframe t_i into various visual scene descriptions before inputting it into $\mathcal{G}$. The process of conditional generation is mathematically articulated as:

$$v_i^k = \mathcal{G}(\Psi(t_i; \theta_{ref}), L_i; \Theta) \tag{4}$$

Here, Θ signifies the pre-trained parameters within $\mathcal{G}$. This approach not only leverages the textual insights but also aligns with the visual context, enabling $\mathcal{G}$ to produce varied and contextually relevant visual guides for o_i.

Guide Finalisation. The final and critical stage of the *GenOV* process involves selecting the most appropriate visual guide from the set of generated candidates, v_i^ks, for a specific oracle o_i. To achieve this, we have designed a selection method by maximising visual consistency through variability analysis, as depicted in Fig. 3(d).

The selection procedure is bifurcated into two stages: initially identifying objects that have the greatest variability which are crucial for an effective visual guide and then determining the most visually consistent guide for these objects. In "visual constraint", we obtain the bounding boxes of the n key objects extracted from t_i, denoted as $b_1, b_2, \cdots, b_n$. With each bounding box b_i, we can obtain its corresponding visual patch from o_{gi} and v_i^k respectively, facilitating the computation of visual similarity, R_i^k, for each object:

$$R_i^k = Sim(p_{oi}, p_{vi}^k) \tag{5}$$

This step enables the identification of the object i^* with the highest variability in visual appearance across the different visual guides R_i^ks, by:

$$i^* = \underset{i}{argmax}[\underset{k}{max}(R_i^k) - \underset{k}{min}(R_i^k)] \tag{6}$$

Then, the optimal visual guide v_i^* is chosen by maximising visual consistency for the key object i^*:

$$v_i^* = \underset{k}{argmax}(R_{i^*}^k) \tag{7}$$

This approach ensures that the selected guide v_i^* not only aligns with the appearance of the oracle glyph but also maintains semantic consistency across historical and contemporary contexts (see supplementary for details).

5. Experiments

5.1. Settings

Implementation. In Sec. 4.1, the MLP Φ is constructed using a two-layer architecture with dimensions set to $[2 - 64 - 2]$ including a softmax for output normalisation. For the Large Vision-Language Model (VLM) Ψ, utilised in the *GenOV* process, we employ GPT-4V(ision) as detailed in [21], allowing us to utilise different prompts to activate its varied capabilities for accomplishing distinct tasks. It's important to note that for the task of object localisation in the visual constraint component, GPT-4V lacks the requisite capability. Consequently, we adopt QWEN-VL [3] as

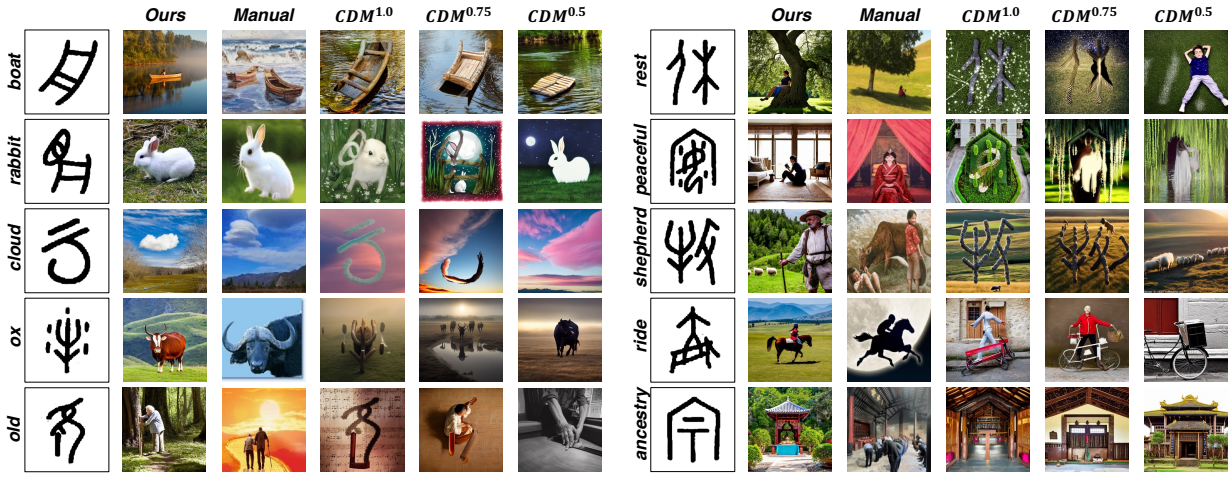

Figure 4. The qualitative results of visual guides generated by different methods.

$CDM^{0.5}$	$CDM^{0.75}$	$CDM^{1.0}$	*Manual*
62.16%	56.76%	43.24%	91.89%
$LDM^{o_{si}}$	$LDM^{o_{ti}}$	$LDM^{\{o_{si},o_{ti}\}}$	*Ours*
54.05%	72.97%	78.38%	89.19%

Table 1. *TransOV* results of different methods on test set.

Ψ for this specific purpose. Furthermore, the conditional image generation model $\mathcal{G}$, referenced in Sec. 4.2, is implemented using GLIGEN [17]. The details regarding the task description, in-context examples, and the specific text prompts used in *GenOV* to harness various capabilities of Ψ are comprehensively provided in the Supplementary.

Baselines. In this paper, we pioneer the use of T2I models to create visual guides aimed at enhancing human comprehension of the relationship between oracle glyphs and their semantic meanings. To our knowledge, this is the first instance of such an application, marking a novel direction in the field. Consequently, there is an absence of existing baseline models tailored to this specific task. To provide a basis for evaluation, we have chosen expert-curated visual guides as a benchmark, which we refer to as the *Manual*. Additionally, for a more technologically advanced comparison, we utilise the state-of-the-art foundation image generation model, ControlNetV1.1 [38]. This model is set to be evaluated against our methods under various condition_scale parameter, denoted as CDM^{*}, where $*$ is the value of $condition\_scale$.

5.2. Main Results

Qualitative Comparison. The qualitative results are presented in Fig. 4, where several key observations are evident. Firstly, for oracles with straightforward semantics like "rabbit" and "ox", our method (*Ours*), the expert-curated approach (*Manual*), and $CDM^{0.5}$ are able to generate images that are visually coherent. In contrast, the $CDM^{1.0}$ and $CDM^{0.75}$ models tend to alter the structural integrity of the objects to align with the oracle glyphs, resulting in distorted, unconventional images. Secondly, for oracles encapsulating more complex meanings, such as "rest" and "shepherd", the effectiveness of the approaches diverges. Our method (*Ours*) shows a higher fidelity to the oracle glyph's layout compared to *Manual* and $CDM^{0.5}$. For instance, in the "rest" oracle, which features people and a tree to convey resting, the original glyph positions the people to the left of the tree. In contrast, the *Manual* solution incorrectly places the people on the right, not aligning with the intended scene. Similarly, in the "shepherd" oracle, which depicts a person herding sheep, the $CDM^{0.5}$ model produces an image with sheep but omits human, failing to capture the complete essence of the glyph. These findings underscore the effectiveness of our method in maintaining semantic integrity while ensuring visual fidelity, paving the way for more nuanced and accurate interpretations of historical symbols and scripts.

Quantitative Comparison. We further explore the influence of text-only conditioning in T2I model and the crucial role of semantic expansion. For this, we compute TransOV using Stable Diffusion under differing text inputs, denoted as $LDM^{o_{si}}$, $LDM^{o_{ti}}$ and $LDM^{\{o_{si},o_{ti}\}}$. The quantitative analysis of *TransOV* for various methods, as detailed in Tab. 1, offers insightful observations: (i) The *GenOV* exhibits a similar *TransOV* performance on the test set with *Manual* (89.19% *v.s.* 91.89%), indicating that the effectiveness of visual guides generated by *GenOV* are nearly on par with those curated by experts. This outcome is particularly promising, suggesting that for oracles lacking manually curated visual guides, *GenOV* can be effectively ap-

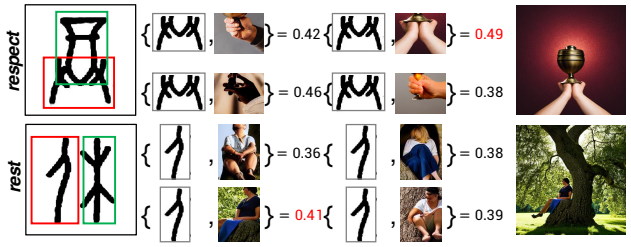

Figure 5. The results of "Guide Finalisation". Each o_{gi} is marked with a red bounding box to identify the key object while others denote additional objects mentioned in t_i.

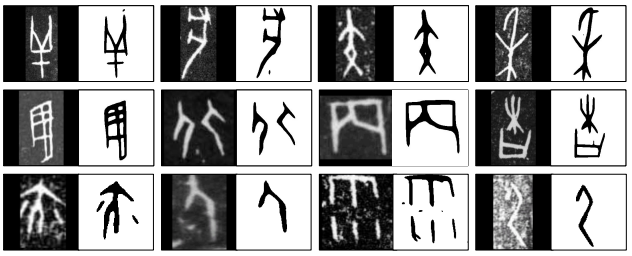

Figure 6. Noisy Oracle Denoiser. For each oracle pairs, the left one is raw, noisy oracle and the right one is the denoising results of our denoiser.

plied to produce high-quality visual guides. This capability is significant as it extends the potential for creating accessible and comprehensible visual interpretations for a wider range of oracle glyphs, especially those not covered by existing expert-curated resources. (ii) $LDM^{o_{si}}$ outperforms both $CDM^{1.0}$ and $CDM^{0.75}$, indicating that adhering too rigidly to the structure of the oracle glyph results in a notable decrease in performance. The underlying reason is the high level of abstraction in the appearance of oracle glyphs and the significant evolution of the concepts they represent over millennia, which suggests that a balance must be struck between structural fidelity and conceptual clarity. The rationale behind incorporating the visual abstraction component in *GenOV* is to address this specific challenge. (iii) The superior performance of $LDM^{o_{ti}}$ and $LDM^{\{o_{si},o_{ti}\}}$ over $LDM^{o_{si}}$ suggests that relying solely on a single character for text conditioning in the generation model is insufficient. It's essential to broaden the semantic scope of the oracle character and unearth the visual narrative behind the glyph. This underscores the critical role of the semantic expansion component in *GenOV*, proving its importance in enhancing the interpretability of the generated visual guides.

Visual Guide Finalisation. In Fig. 5, we display the outcomes of our tailored oracle-specific visual guide selection method. For each oracle glyph, the process begins by identifying the visually sensitive object, which is highlighted in each o_{gi}. Subsequently, we illustrate the visual similarity between the oracle patch and the corresponding image patch for this particular object across various visual guide candidates, culminating in the selection of the final visual guide, v_i^*. Fig. 5 shows the superior visual consistency of the final selected guide compared to other candidates. For instance, in the case of "respect", our method identifies "hand" as the key object. The final guide chosen aligns remarkably well with the semantic of "two hands holding" in the oracle, demonstrating the effectiveness of our approach in maintaining semantic integrity while ensuring visual coherence.

Noisy Oracle Denoising. To address a more complex setting where only raw, noisy oracle bone scripts are available, our experiment sought to answer the question: how can a visual guide be effectively generated from such challenging inputs? We utilised the architecture of DiT [22], initially pre-training the model on a large-scale dataset of free-hand sketches [8]. This step was crucial for establishing a robust foundational understanding of abstract and iconic visual data. Subsequently, we fine-tuned this model using a limited set of scan-clean oracle pair data [34], thereby customising it to function as a specialised oracle denoiser. The effectiveness of this denoiser is shown in Fig. 6, which not only highlights the denoiser's ability to handle raw oracle scripts but also demonstrates its impressive generalisation capabilities. This enhanced denoising functionality significantly broadens the potential applications of our *GenOV* framework, enabling it to operate effectively even with less-than-ideal script conditions.

6. Conclusion

In conclusion, we successfully demonstrated the potential of AI in enhancing the public's understanding and appreciation of historical visual arts, specifically oracle bone scripts. Our key contribution lies in the creation of the *TransOV* metric, a novel tool for quantitatively measuring the effectiveness of visual guides in aiding comprehension of oracle bone scripts, which has been crucial in our empirical analysis, allowing us to validate the quality of our AI-generated guides against human judgement. And we proposed *GenOV* adeptly handles the complex task of interpreting oracle with limited and abstract information, converting them into detailed, understandable visual guides. Our experiments have shown that the guides produced by *GenOV* are not only comparable to those curated by experts but also provide an invaluable resource for scripts lacking expert interpretation. Moreover, *GenOV's* ability to denoise and interpret raw, noisy glyphs broadens its applicability, making it a versatile tool for historians and curators. We anticipate further contributions to both the field of AI and the broader domain of cultural heritage preservation and interpretation.

Acknowledgements We thank the reviewers for their valuable comments. This work was funded by National Natural Science Foundation of China under grant # 62076034

References

[1] Dictionary. http://jiaguwen.shufami.com. 3

[2] Dictionary. https://www.vividict.com. 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5

[5] Xiang Chang, Fei Chao, Changjing Shang, and Qiang Shen. Sundial-gan: A cascade generative adversarial networks framework for deciphering oracle bone inscriptions. In *ACM MM*, 2022. 2

[6] Yoshiyuki Fujikawa, Hengyi Li, Xuebin Yue, CV Aravinda, G Amar Prabhu, and Lin Meng. Recognition of oracle bone inscriptions by using two deep learning models. *International Journal of Digital Humanities*, 2022. 2

[7] Jun Guo, Changhu Wang, Edgar Roman-Rangel, Hongyang Chao, and Yong Rui. Building hierarchical representations for oracle character and sketch recognition. *IEEE Transactions on Image Processing*, 2015. 2

[8] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR*, 2018. 8

[9] Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyuan Liu, and Maosong Sun. Isobs: An information system for oracle bone script. In *EMNLP*, 2020. 2

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[11] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 3

[12] Shuangping Huang, Haobin Wang, Yongge Liu, Xiaosong Shi, and Lianwen Jin. Obc306: A large-scale oracle bone character recognition dataset. In *ICDAR*, 2019. 2

[13] Runhua Jiang, Yongge Liu, Boyuan Zhang, Xu Chen, Deng Li, and Yahong Han. Oraclepoints: A hybrid neural representation for oracle character. In *ACM MM*, 2023. 2

[14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 3

[15] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 3

[16] Jing Li, Qiu-Feng Wang, Kaizhu Huang, Xi Yang, Rui Zhang, and John Y Goulermas. Towards better long-tailed oracle character recognition with adversarial data augmentation. *Pattern Recognition*, 2023. 2

[17] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 7

[18] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. 3, 6

[19] Lin Meng, Bing Lyu, Zhiyu Zhang, CV Aravinda, Naoto Kamitoku, and Katsuhiro Yamazaki. Oracle bone inscription detector based on ssd. In *ICIAP*, 2019. 2

[20] OpenAI. Gpt-4 technical report. 2023. 3

[21] OpenAI. Gpt-4v(ision) system card. *openai.com*, 2023. 6

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 8

[23] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *ACM MM*, 2023. 6

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[25] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *CVPR*, 2023. 3

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3

[29] Daqian Shi, Xiaolei Diao, Lida Shi, Hao Tang, Yang Chi, Chuntao Li, and Hao Xu. Charformer: A glyph fusion based attentive framework for high-precision character image denoising. In *ACM MM*, 2022. 2

[30] Daqian Shi, Xiaolei Diao, Hao Tang, Xiaomin Li, Hao Xing, and Hao Xu. Rcrn: Real-world character image restoration network via skeleton extraction. In *ACM MM*, 2022. 2

[31] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3

[32] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *CVPR*, 2023. 3

[33] Mei Wang and Weihong Deng. Oracle-mnist: a realistic image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:2205.09442*, 2022. 2

[34] Mei Wang, Weihong Deng, and Cheng-Lin Liu. Unsupervised structure-texture separation network for oracle character recognition. *IEEE Transactions on Image Processing*, 2022. 2, 8

[35] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff:

Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023. 3

[36] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *ICCV*, 2023. 3

[37] Xuebin Yue, Hengyi Li, Yoshiyuki Fujikawa, and Lin Meng. Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition. *ACM Journal on Computing and Cultural Heritage*, 2022. 2

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 7