Supplementary: Making Visual Sense of Oracle Bones for You and Me

1. Task-Specific Prompts

Fig. 1 displays the prompts utilised in "Textual Contextualisation" with GPT-4V [5]. This includes a detailed task description, in-context examples, and an input oracle character. The purpose of the task description is to aid the Ψ model in accurately grasping the objective of the task and establishing the format for inputs and outputs. Additionally, the in-context examples are instrumental in reinforcing task-specific linguistic styles, appropriate use of terminology, and conventional reasoning patterns.

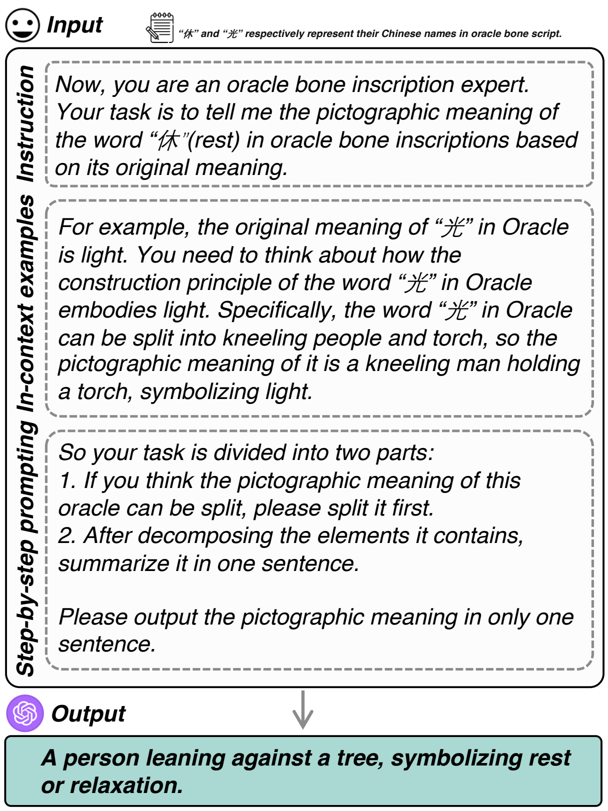

Figure 1. The prompt utilised in "Textual Contextualisation"

Fig. 2 displays the prompts employed in "Visual Constraints" for L_i^v generation with QWEN-VL [2], which aims to activate the object localisation capability of Ψ.

Fig. 3 illustrates the prompt used to acquire L_i^ts with [5],

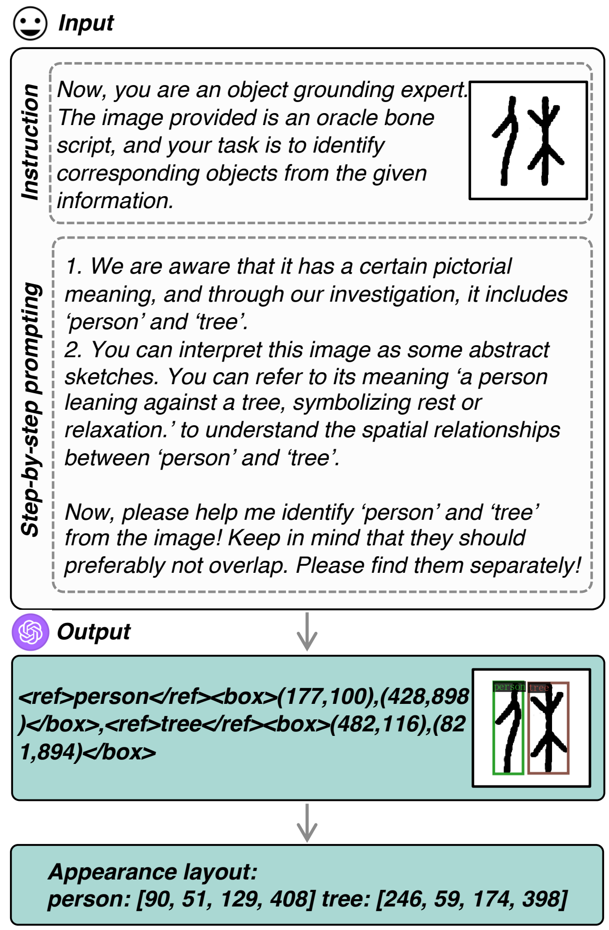

Figure 2. The prompt employed in "Visual Constraints"

aiming to trigger the spatial analysis from textual inputs in the Ψ inspired by [4, 6].

Similarly, Fig. 4 displays the prompt employed to derive L_i with [5], designed to stimulate the multimodal spatial analysis and visual understanding capabilities of Ψ when processing multiple visual inputs.

2. Visual guides of Human Study

Fig. 5 illustrates examples of 28 visual guides of o_i used in human study. The diversity in these guides results from the collected data from the website [1], varying input con-

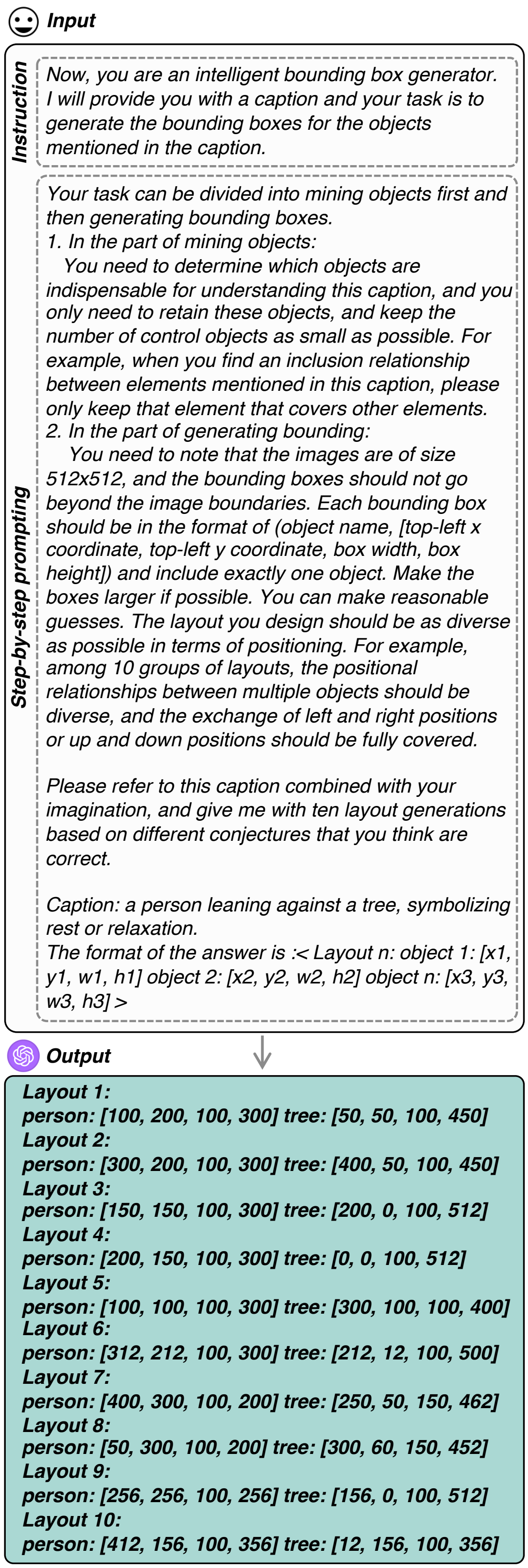

Figure 3. The prompt used to acquire L_i^ts

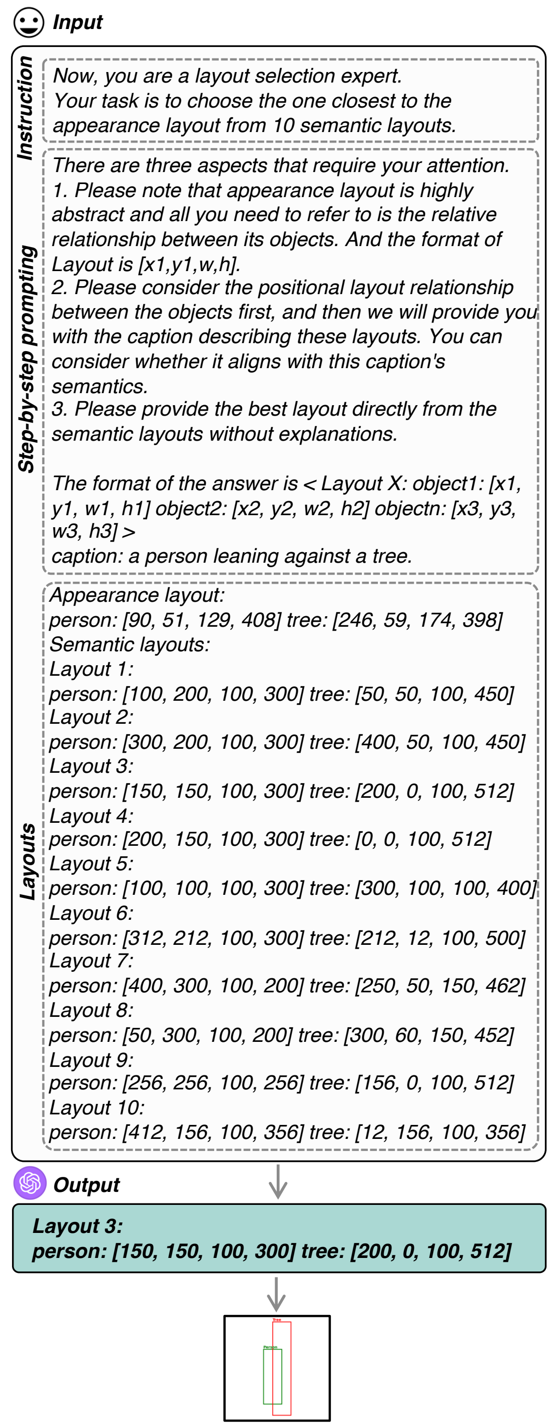

Figure 4. The prompt employed to derive L_i

ditions, and model[7] parameters.

3. The reliability of TransOV

To ensure the reliability of the model, we divide it into two steps: (i) We measure the model calibration by Ex-

Dep.	Wid.	BN	Acc.(↑)	Spec.(↑)	ECE(↓)	MCE(↓)	Dep.	Wid.	BN	Acc.(↑)	Spec.(↑)	ECE(↓)	MCE(↓)
2	64	✓	86.00%	86.06%	0.24	0.27	2	128	✓	81.50%	79.21%	0.25	0.28
3	64	✓	84.25%	85.93%	0.25	0.27	3	128	✓	83.25%	79.90%	0.28	0.29
2	64		54.00%	55.00%	0.46	0.48	3	64		52.25%	53.37%	0.45	0.48

Table 1. Ablation of network structure of *TransOV*.

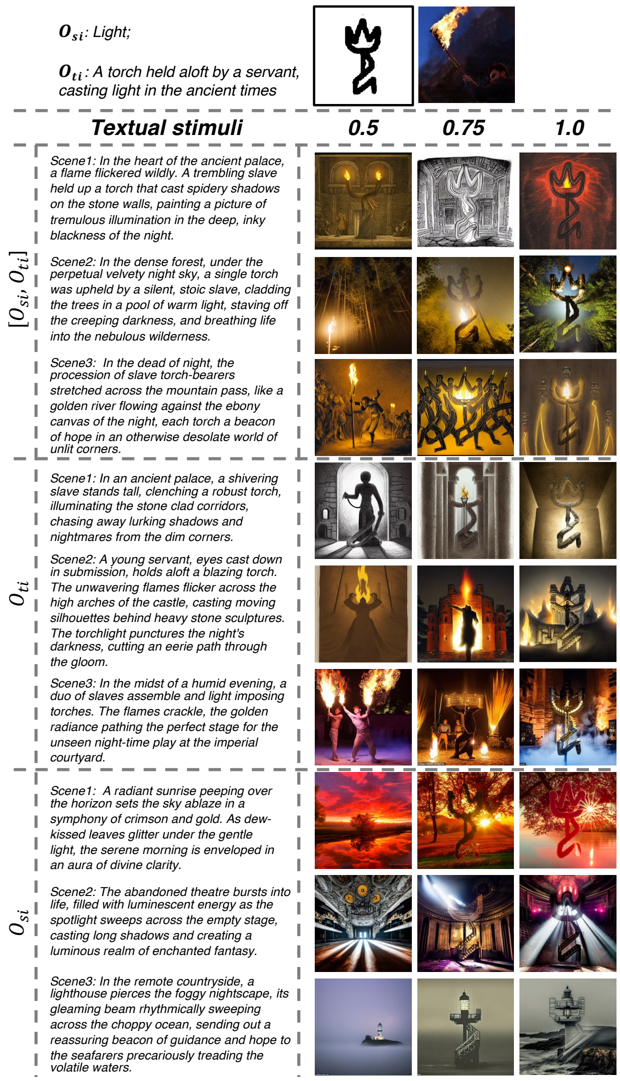

Figure 5. Examples of 28 visual guides

pected Calibration Error (ECE) and Maximum Calibration Error (MCE) with 10 bins [3]. We conducted a series of experiments about three factors of network: depth, width, and batch normalisation. This helped us identify a network structure with comparable reliability, as detailed in Tab.1. (ii) To further remedy the miscalibration, we utilise Platt-scaling, an effective parametric approach to calibration. This approach notably decreased ECE ($0.24 \rightarrow 0.19$) and MCE ($0.27 \rightarrow 0.21$), albeit at the cost of a slight decrease of accuracy ($86.00\% \rightarrow 84.75\%$) and specificity ($86.06\% \rightarrow 85.71\%$). We hold that the reliability of the model is paramount, making this trade-off justifiable (Fig.6).

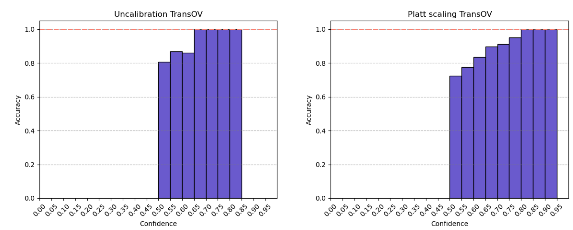

Figure 6. Reliability diagrams of TransOV before and after calibration

4. An example of Guide Finalisation

Fig. 7 illustrates an example in "Guide Finalisation". For oracle "light", the objects extracted from t_i are "person" and "torch". We locate the visual patches corresponding to "person" and "torch" in o_{gi} and v_i^ks respectively, and calculate R_i^k via Eq.1. The x-axis in the figure corresponds to R_i^k for "person", and the y-axis for "torch". Notably, the range of variation for "person" is much larger than for "torch". Consequently, "person" is selected as the key object for the final selection. The image highlighted with a red frame is identified as the final visual guide.

$$R_i^k = Sim(p_{oi}, p_{vi}^k) \tag{1}$$

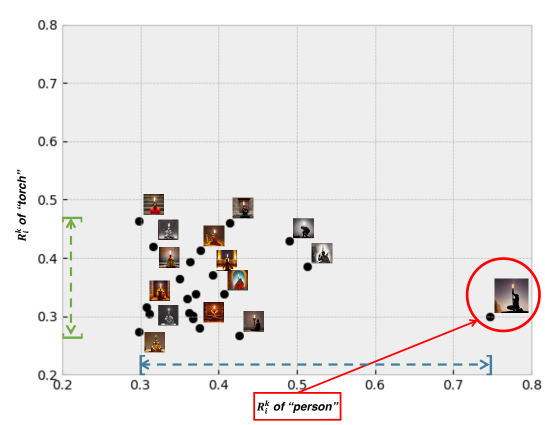

Figure 7. An example in "Guide Finalisation"

References

[1] Dictionary. https://www.vividict.com. 1

[2] 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. 1

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 3

[4] 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. 1

[5] OpenAI. Gpt-4v(ision) system card. *openai.com*, 2023. 1

[6] 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. 1

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