# A MLMM Evaluation Details

## A.1 Comparison of Judgement Templates.

In our study, we explored two distinct methodologies for evaluating the output of MLLMs. The first approach consolidates all generated images into a single composite image, facilitating an evaluative process from a human-centric perspective. This approach is delineated in the judgment template shown in Figure 1. Alternatively, the second method involves separately encoding each generated image, which is illustrated in the judgment template of Figure 2.

Figure 1: Judgment Template from a Human Perspective.

Figure 2: Judgment template for encoding each image.

1

8  The results of these evaluations are presented in Figure 3 and 4, respectively. When employing
9  the first methodology, wherein multiple images are assembled within a single frame, the visual
10 comparisons and subsequent justifications of ratings demonstrate a high degree of consistency.
11 Conversely, the approach of encoding and evaluating each image individually reveals variations in the
12 visual assessments and the rationales provided for the ratings, indicating a divergence in consistency
   compared to the composite image evaluation method.

### Input Image:

Ref     1     2     3     4

### GPT-4V Judgement:

Score of image 1: 2
•Justification of your rating: The color in image 1 is a gradient of orange and brown, somewhat closer to the wooden tones in reference image but not an exact match. The texture is minimal and does not capture the detailed wooden texture observed in reference image. Moreover, the font style is significantly thinner, lacking the boldness and intricate designs present in the font of reference image.

Score of image 2: 2
•Justification of your rating: Image 2 features metallic orange and gray tones, differing greatly from the wooden color of reference image. Its texture is metallic and industrial, contrasting sharply with the natural wooden texture of reference image. Additionally, the font style in image 2 is more futuristic and mechanical, lacking the traditional wooden style's thickness and sharpness seen in reference image.

Score of image 3: 3
•Justification of your rating: The colors in image 3 consist of a gradient of orange and yellow, slightly resembling the wooden hues of reference image but not perfectly matching. The texture is smoother and less detailed than the wooden texture of reference image. The font style shares a similar thickness and sharpness to that of reference image, but it is simpler and lacks the intricate details.

Score of image 4: 1
•Justification of your rating: Image 4 has a green color, completely different from the wooden color of reference image. Its texture appears pixelated and far less detailed compared to the wooden texture in reference image. Furthermore, the font style is significantly different, displaying a different thickness and a design that diverges from the font style of reference image.

### Claude-3-Opus Judgement:

Score of image 1: 4, The color of the typography in image 1 closely matches the orange-brown hue of the reference image. The texture appears slightly smoother and more uniform compared to the reference, but still maintains a good degree of similarity. The font style, thickness, and sharpness are well-preserved, making the overall typography quite similar to the reference.

Score of image 2: 3, Image 2's typography has a more colorful and vibrant appearance compared to the reference, incorporating blue and yellow tones. The texture seems to have more of a metallic or glossy quality that differs from the reference. However, the font style and general design are still fairly well-matched.

Score of image 3: 4, The typography in image 3 exhibits colors that are quite similar to the reference, with a predominantly orange hue. The texture has a slightly more weathered or distressed look compared to the reference, but this difference is relatively minor. The font maintains good fidelity to the reference in terms of style, thickness, and sharpness.

Score of image 4: 3, Image 4's typography incorporates a green color scheme that deviates from the orange tones of the reference image. The texture appears somewhat smoother and more uniform than the reference. Despite these differences in color and texture, the font style and overall design remain largely consistent with the reference typography.

Figure 3: Judgment about Human Perspective template from GPT-4V and Claude-3-Opus.

13

2

| ### Input Image 1 | ### Input Image 2 | ### Input Image 3 | ### Input Image 4 | ### Input Image 5 |

### GPT-4V Judgement:

[Score between image 5 and target image 1]: 2
Justification:
•Color Consistency: Image 1 features a plain brown tone, differing significantly from the gradient orange and yellow hues in image 5.
•Texture Quality: Image 1 has a flat and smooth surface, lacking the textured, almost metallic sheen seen in image 5.
•Font Fidelity: While the shape of the typography in image 1 is somewhat similar, the lack of texture and depth in color makes it feel quite different from the intricate details of image 5.

[Score between image 5 and target image 2]: 4
Justification:
•Color Consistency: The colors in image 2, which include blue, black, and gold, incorporate a gradient effect similar to the reference, but the specific hues differ.
•Texture Quality: The texture in image 2 is rich and detailed, closely matching the intricate, layered look of image 5.
•Font Fidelity: Image 2 shows high similarity in font style, ornate details, and thickness, very well preserved compared to image 5.

[Score between image 5 and target image 3]: 3
Justification:
•Color Consistency: The gradient colors in image 3 (red to yellow) are somewhat similar but not identical to the orange and yellow gradient in image 5.
•Texture Quality: The glossy texture in image 3 reflects some sheen of image 5 but lacks the same depth and metallic quality.
•Font Fidelity: The overall shape and contour of the font in image 3 is good, but the different texture and color scheme alter the overall feel compared to image 5.

[Score between image 5 and target image 4]: 1
Justification:
•Color Consistency: Image 4's flat green color is vastly different from the gradient orange and yellow in image 5.
•Texture Quality: The texture in image 4 is very flat and lacks the detailed, textured appearance seen in image 5.
•Font Fidelity: The font style in image 4 is simplified and does not capture the ornate thickness and curves of image 5, resulting in low fidelity across all aspects: color, texture, and font style.

### Claude-3-Opus Judgement:

Here is my evaluation of the similarity between the typography in images 1-4 compared to the original typography in image 5:

Score about between image 5 and target image 1: 2, While image 3 maintains the general shape of a "T", it deviates significantly from the original typography. The wooden texture and flat brown color are completely different from the metallic orange and red of image 5. The font is noticeably thicker and lacks the sharp, mechanical edges.

Score about between image 5 and target image 2: 4, Image 2 does an excellent job replicating the intricate mechanical design and shiny metallic texture of the original "T" typography. The colors lean more heavily into orange and yellow tones compared to image 5, but overall it achieves a high degree of fidelity to the source material in terms of style and detail.

Score about between image 5 and target image 3: 3, The color scheme of the "T" in image 1 captures the orange and red gradient of the original, but the texture appears more rough and distressed compared to the smooth metallic surface in image 5. The font style and thickness are fairly well preserved.

Score about between image 5 and target image 4: 3, Image 4 preserves the color scheme quite well, utilizing a similar orange to green gradient as the original. However, the font is slightly thicker and fuzzier compared to the crisp lines in image 5. The felt-like texture is also a departure from the smooth metallic surface of the original.

Figure 4: Judgment about encoding each image template from GPT-4V and Claude-3-Opus.

3

A.2   Safety Misclassification of GPT-4V

In Figure 5, we present examples that, despite containing content deemed safe, elicited a 'bad request' error from GPT-4V.
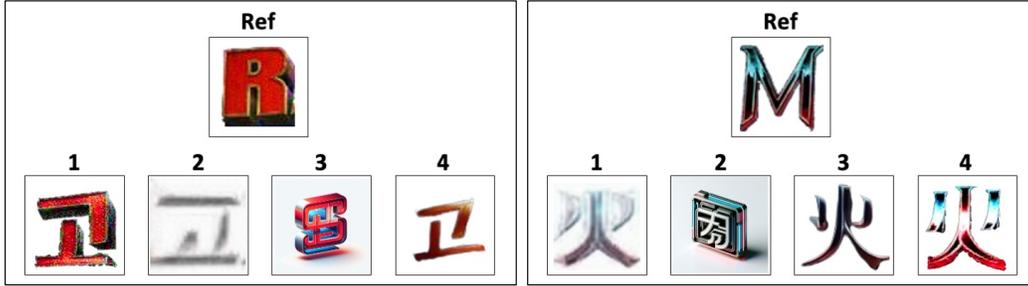
Figure 5: Samples of GPT-4V safety misclassification.

# B   More Ablation Studies

**Visual Text Generation Prompts.**   The prompt for visual text generation can be expressed in various ways, as shown in Figure 6. In this ablation study, the seed was fixed at 100, and the prompt was changed for experimentation. While the initial images generated for each prompt vary greatly, the final generated images at the end of training maintain a variety of style fidelity without losing legibility.
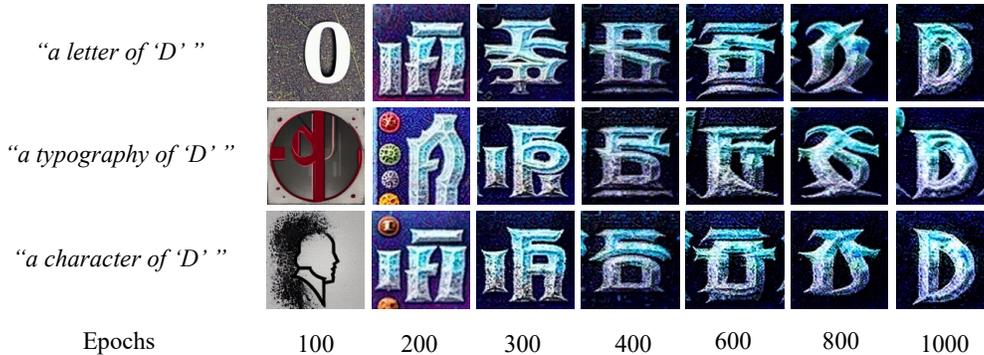
Figure 6: Various prompts for generating visual text 'D' and the corresponding generation results.

**Effectiveness of Latent Space Distances.**   In the generator component of SIGIL, if glyph images are used directly without glyph latent guidance, the generator training results are as shown in Figure 7 (a). In comparison, by using the glyph latent guidance we propose, the generation results can be obtained as shown in Figure 7 (b).

# C   Dataset Curation Details

**Collecting Multilingual Film Poster Images.**   We collected multilingual posters for movie titles from a movie database website. Then, we kept only those titles that had pairs in different languages, creating pairs of multilingual posters.

4

| | *a typography of 'K'* | | | | | *a typography of 'K'* | | | |
| Epochs | 100 | 400 | 700 | 1000 | Epochs | 100 | 400 | 700 | 1000 |
| | | (a) Glyph Image | | | | | (b) Glyph Latent $z_g$ | | | |
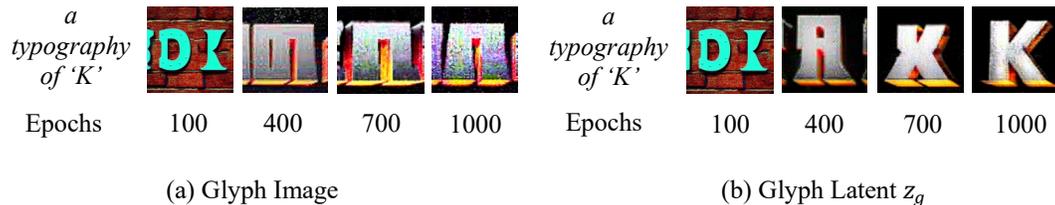
Figure 7: Generation results for the prompt "a typography of 'K'".

**Filtering by Style Similarity.** We hired three AI researchers to perform filtering tasks by determining the style similarity of the text in two poster images displayed on the screen. A screenshot of the full text of instructions for this task can be found in Figure 8. Figure 8 also shows the annotation tool we developed specifically for this task.

**Character-level Bounding Box Annotation.** The three human annotators who participated in the prior filtering task volunteered for this task as well. They also carried out bounding box labeling on the filtered images. The instructions for this task are introduced in Figure 8. Each annotator receives a wage ranging from 12-15 dollars per hour, the total amount spent on participant compensation is 648 dollars.

## C.1 Ethical Considerations

In this paper, we present the MuST-Bench, which incorporates copyrighted film posters designed by human experts. To address the issue of copyright, instead of distributing raw data, we provide download links for each poster image along with bounding box annotations. Furthermore, accompanying code is made available, enabling easy transformation of the posters into the format proposed in the paper. This approach ensures compliance with copyright laws while maintaining the utility of the dataset for research purposes.

## D Implementation and Training Details

SIGIL comprises two main components, the generator and the corrector. The implementation and training details for each are as follows:

**Generator.** For the pre-training configuration, we employed the runwayml/stable-diffusion-v1-5 model publicly available weights from the Huggingface Hub (https://huggingface.co/models). The training dataset utilized was derived from the MuST-Bench style subject, where each style subject allowed for fine-tuning on two glyph combinations. Training was conducted for approximately 1,000 epochs. The total epochs were adjusted based on the dataset volume; for instance, a dataset containing 15 training samples warranted an extension to 1,005 epochs to ensure thorough model training. The learning rate was maintained at 1e-4 throughout the training process. Regarding the framework and computational resources, we extended the LoRA-based implementation of Dreambooth to process multi-subject inputs, allowing for concurrent fine-tuning on two glyph subjects. Training was performed on a single NVIDIA A100 40GB GPU, with each session completing in about 20 minutes.

**Corrector.** The corrector was initiated from the last checkpoint obtained after the preliminary fine-tuning phase (generator's last checkpoint). The training dataset comprised images sampled during the generator's operation, serving as the primary data for further training. Training was conducted with an emphasis on efficiency, incorporating an early stopping mechanism to curtail the training as soon as the model reached a satisfactory level of performance. On average, the training duration

5

Figure 8: A screenshot of the human annotation interface for MuST-Bench dataset curation.

was approximately 30 minutes using the same GPU as the generator. The learning rate was set to 3e-4, which was determined to be optimal for achieving convergence while maintaining training stability. Additionally, the EasyOCR tool was employed to facilitate multilanguage text recognition. While utilizing this off-the-shelf OCR model, we adapted its output mechanism to provide confidence scores for characters presented in the prompt instead of predicted characters.

# E   Image Input to DALL-E3

SIGIL can accept style images as input. However, DALL-E 3 only accepts text inputs. Therefore, to ensure fairness in evaluation, GPT-4V is used as a bridge to enable DALL-E 3 to "see" images. As shown in Figure 9, GPT-4V views the style image and creates a description of that style. This

6

description is then combined with the textual prompt and sent to DALL-E 3. The importance of using GPT-4V for style description can be seen by comparing Figure 10 parts (a) and (b).

"The typography style in the image is metallic, sleek, and has a futuristic appeal with sharp angles and a glossy finish. The use of a contrasting red accent enhances its modern and dynamic look."

(a) Style image            (b) GPT-4V descriptions for style image
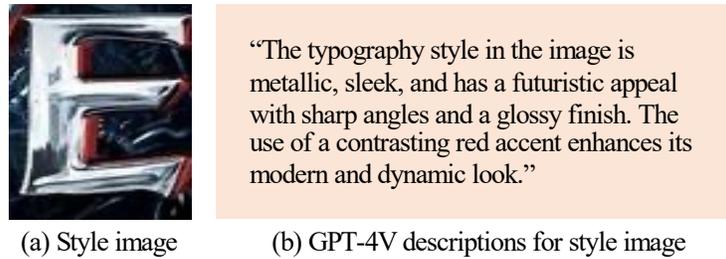
Figure 9: GPT-4V description about style image.

(a) DALL-E3 generated images for "a typography of 'Y'".

(b) DALL-E3 generated images for GPT-4V's style image descriptions + "a typography of 'Y'".
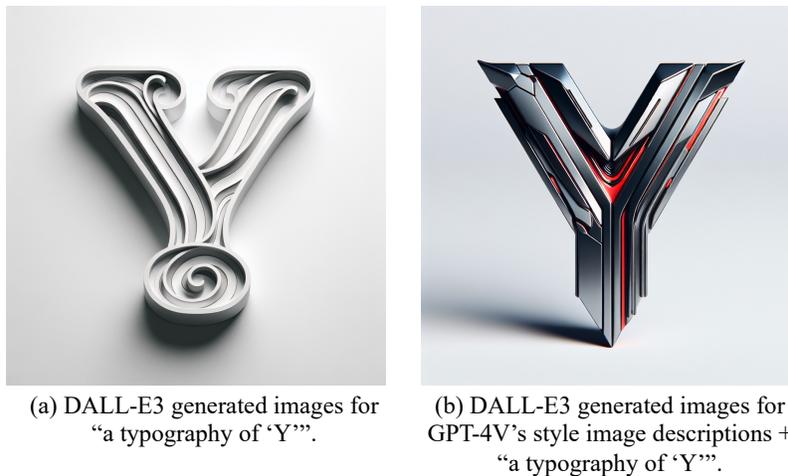
Figure 10: Comparison of DALL-E 3 generated outputs based on the presence of GPT-4V descriptions.

## F   User Studies

In the user study, assessments were conducted on two distinct parameters, style fidelity and legibility. 60 participants were provided with detailed guidelines for each instruction, as outlined in Figure 11, before initiating the study. Figure 12 is a sample of a user study. Figure 13 exemplifies a scoring sheet used for ranking the outcomes of comparison methods based on style fidelity, and illustrates a scoring sheet designed for the evaluation of legibility.

## G   Discussion and Limitations

Our method can optimize generated images to match the input style image even with a small amount of data, without relying on extensive image-caption datasets. In this study, we have successfully combined two specific glyphs per style during training. In future work, we plan to explore methods that enable the combination of a greater number of glyphs.

In the real world, there is abundant high-resolution typography data. However, for collecting multi-language pairs with the same style, movie poster data has proven to be the most effective. When extracting typography from movie posters, the resulting size is relatively small. If the input style


7


Multi-language Visual Text Generation User Study

[English]

In this task, two types of questions will alternate.
# **Question 1** (Style Similarity): Look at the reference image and rank images A, B, C, D in order of similarity to the reference image's style.
- Note: Rank them based on style similarity, not personal preference.
- Style includes color, texture, font, etc.

# **Question 2** (Legibility of Text): Based on the given prompt, view the text in images A, B, C, D and rank them in order of accuracy and legibility.

Figure 11: User study instruction and user interface.
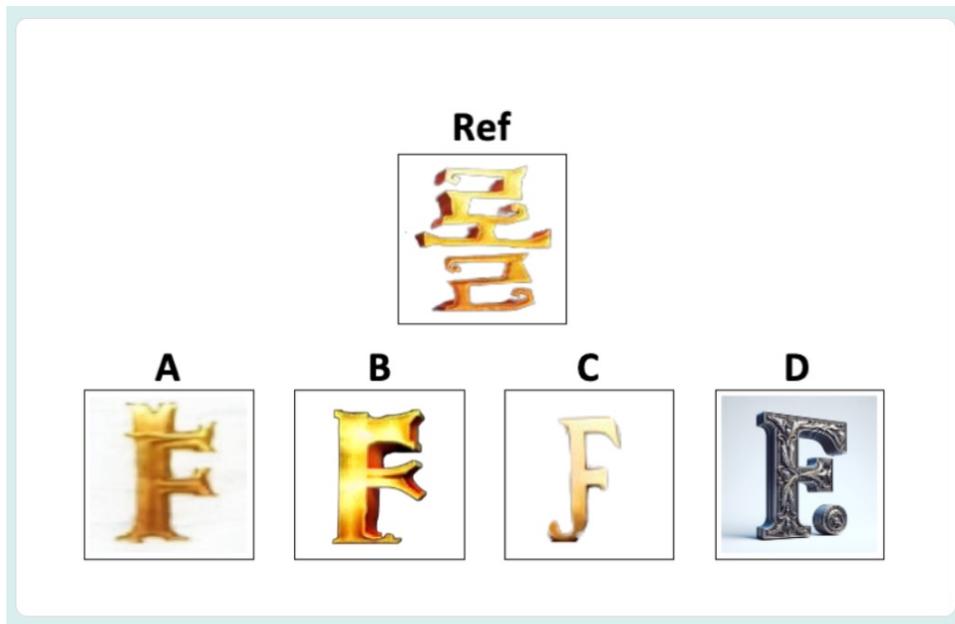
Figure 12: Evaluation sample of the user study.

image is low resolution, it can complicate VAE-based pixel-level encoding, potentially affecting the generation results. In future research, we aim to explore methods to generate typography in different languages from single-language style images, thereby utilizing more real-world data.

Unlike English, some glyphs with complex strokes in Chinese and Korean present challenges in generation. Applying fine-grained image generation methods could enable the accurate creation of all glyphs.

Lastly, we observed that the EasyOCR model used as a reward model in the Reinforcement Learning process sometimes exhibits False Negative issues with generated images. Additionally, if the model

8

Figure 13: Questionnaire examples of style fidelity and legibility.

generates words not in EasyOCR's vocabulary, it can complicate the process. Future work will attempt
to use more robust OCR models with larger vocabularies and higher prediction accuracy.

9