Conditional Text-to-Image Generation with Reference Guidance

Supplementary Material

A. Additional Technical Details

A.1. Auxiliary Network Architecture

We provide details of the auxiliary recognition network architecture. The auxiliary network is composed of the encoder and the recognition head. The encoders for the text image generation and logo image generation share the same encoder structure. The shared encoder network consists of multiple encoding blocks, each sharing the structure of the blocks in the VAE network of the diffusion model. More specifically, each encoder block comprises a GroupNorm layer, two Residual blocks, and another GroupNorm layer, where the number of groups is set to 16. We use four encoding blocks for both text recognition and log recognition networks. Except for the last encoder block, the spatial dimension is down-sampled by the factor of 2 using the convolution layer with stride=2.

Text recognition head. The text recognition head is composed of four cross-attention layers, where the cross-attention is computed between the sequence of input image tokens and the sequence of positional embeddings corresponding to the characters of the output text word. The cross-attention we compute here can be formulated as:

$$\text{Attention}(\mathbf{r}, \tilde{\mathbf{z}}, \tilde{\mathbf{z}}) = \text{softmax}(\frac{\mathbf{r}\tilde{\mathbf{z}}^T}{\sqrt{d_k}})\tilde{\mathbf{z}}, \tag{1}$$

where $\mathbf{r}$ is the sequence of positional embeddings corresponding to i^{th} character of the word, and $\tilde{\mathbf{z}}$ is the encoding from the auxiliary encoder.

Logo recognition head. The recognition head for the auxiliary logo recognition network is composed of a single fully connected layer, where the output dimension corresponds to the size of the logo set.

Text Recognition Loss. The text recognition loss that we use is defined as below,

$$\mathbf{r}_k = \text{RoIAlign}(\hat{\mathbf{z}}^0, \mathbf{B}_k) \tag{2}$$

$$\mathbf{o}_k = \psi_{\text{recog}}(\mathbf{r}_k) \tag{3}$$

$$\mathcal{L}_{\text{recog}} = \frac{-1}{L}\sum_{j=1}^{L} y_j \log \mathbf{o}_{k,j}, \tag{4}$$

where $\mathbf{B}_k$ is the bounding box label for k-th region in the image, $\mathbf{o}_k$ is the network output, and L denotes the length of the word in the k-th region. j is an enumerator for each character in the word. We predict the probability distribution over the character set at each character position j denoted as $\mathbf{o}_k, j \in \mathbb{R}^{|C|}$. For each position, we apply cross-entropy loss between the one-hot label $y_j \in \mathbb{R}^{|C|}$. For MLT loss, we use the same loss design, but the output dimension is extended to 847 to cover all the characters of the target languages.

A.2. Additional Details of Augmented CFG

Classifier-free guidance [21] considers sharpened posterior distribution $P_\theta(\mathbf{z}|\mathbf{c}) \propto P_\theta(\mathbf{z})P_\theta(\mathbf{c}|\mathbf{z})^\omega$. Using Bayes' rule for some timestep t,

$$\begin{aligned}\nabla_{\mathbf{z}} \log P_\theta(\mathbf{z}_t|\mathbf{c}) = \nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t) \\ + \omega(\nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t|\mathbf{c}) - \nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t)).\end{aligned} \tag{5}$$

Since the score function is parameterized with ϵ_θ, we have

$$\hat{\epsilon_\theta}^t := \epsilon_\theta(\mathbf{z}^t, \mathbf{c}, t) + \omega(\epsilon_\theta(\mathbf{z}^t, \mathbf{c}, t) - \epsilon_\theta(\mathbf{z}^t, \varnothing, t)), \tag{6}$$

where $\epsilon_\theta(\mathbf{z}^t, \varnothing, t)$ denotes unconditional ϵ prediction with an empty text prompt.

Similarly we model the posterior distribution that can be represented as,

$$\begin{aligned}P_\theta(\mathbf{z}|\mathbf{s}, \mathbf{m}, \mathbf{c}) \propto P_\theta(\mathbf{z})P_\theta(\mathbf{s}, \mathbf{m}, \mathbf{c}|\mathbf{z})^{\omega_{\text{all}}} \\ \cdot P_\theta(\mathbf{z})P_\theta(\mathbf{s}, \mathbf{m}|\mathbf{z})^{\omega_{\text{prompt}}} P_\theta(\mathbf{c}|\mathbf{z})^{\omega_{\text{ref}}}.\end{aligned} \tag{7}$$

Similarly, for some timestep t,

$$\begin{aligned}\nabla_{\mathbf{z}} \log P_\theta(\mathbf{z}_t|\mathbf{s}, \mathbf{m}, \mathbf{c}) = \nabla_{\mathbf{z}_t} \log P(\mathbf{z}_t) \\ + \omega_{\text{all}} \left(\nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t|\mathbf{s}, \mathbf{m}, \mathbf{c}) - \nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t)\right) \\ + \omega_{\text{ref}} \left(\nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t|\mathbf{s}, \mathbf{m}, \varnothing) - \nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t)\right) \\ + \omega_{\text{prompt}} \left(\nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t|\varnothing, \varnothing, \mathbf{c}) - \nabla_{\mathbf{z}_t} \log P_\theta(\mathbf{z}_t)\right).\end{aligned} \tag{8}$$

From this, we derive,

$$\begin{aligned}\hat{\epsilon}^t = \underbrace{\epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)}_{\text{Unconditional}} \\ \underbrace{+\omega_{\text{prompt}}\Big[\epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \mathbf{c}, t) - \epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]}_{\text{Caption Guidance}} \\ \underbrace{+\omega_{\text{ref}}\Big[\epsilon_\theta(\text{concat}(\mathbf{z}^t, \mathbf{s}, \mathbf{m}), \varnothing, t) - \epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]}_{\text{Reference Guidance}} \\ \underbrace{+\omega_{\text{all}}\Big[\epsilon(\text{concat}(\mathbf{z}^t, \mathbf{s}, \mathbf{m}), \mathbf{c}, t) - \epsilon(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]}_{\text{All Guidance}},\end{aligned} \tag{9}$$

A.3. Additional Details of Datasets

MARIO-10M [12] is a compilation of data from various publicly accessible sources, including the LAION-400M

[53], TMDB [6], and Open Library datasets. The images within this dataset are filtered by a text detector and annotated by an OCR model. A training set containing 10 million images with English-only annotations are utilized for training.

ICDAR2019 [14] stands as a multi-lingual text (MLT) dataset initially curated for text detection and recognition. This dataset contains 10,000 training images and provides coordinates of each text word with corresponding character labels. We use a subset with six languages, including English, French, German, and Italian to train our MLT image generation from this dataset.

FlickLogos-32 [49] contains 8,240 images with 32 logo brands. This dataset was originally collected for the logo retrieval and logo detection/recognition tasks. We use a subset of 2,240 images that contain at least one logo as the training set for the logo image generation task.

Logos in the wild (LITW) [61] comprises 11,054 images obtained from the Google image search engine, featuring 871 unique brands. We filter the dataset by excluding images with logos smaller than 45 pixels high and images with no logo. After filtering, a total of 4,206 images are obtained for training.

LAION-Aesthetics [4]. LAION-5B [54] is a large-scale dataset of image-caption pairs. LAION-Aesthetics is the subset of LAION-5B, filtered by a trained aesthetic score predictor, and contains only high aesthetic score (above 6.5) images. This dataset does not provide any annotations of text or logo, and is included in our training to prevent overfitting and catastrophic forgetting.

Aesthetic text dataset (AText). We collect an additional dataset for English text generation with high-aesthetic and high-artistic images from the Internet [5]. We measure the aesthetic scores of the images using a pretrained model from [54] and exclude those with low aesthetic scores. We run an OCR model to detect text regions and annotate the text within the regions. Example images are shown in Figure 5.

Synthetic logo/text dataset. Due to the limited numbers of MLT and logo training samples from public datasets, we augment the training set with synthesis. We use high-resolution images from public datasets as the background and render the target text or logo. For the multi-lingual experiment, we use an E-book text corpus [1] and translate the text into the target languages of interest. For the logo image experiment, we use publicly available icon images [3] as the target objects to be rendered on the background images (Figure 4). In total, we obtain 23,892 synthetic logo images and 184,912 synthetic MLT images.

A.4. Additional Details of Benchmarks

Benchmark for English generation. We adopt the MARIO-Eval [12] benchmark to assess the model performance in generating English scene-text images. The MARIO-Eval benchmark contains 5,414 prompts. This benchmark is composed of six different subsets, each a subset of multiple benchmarks, including DrawBenchText [39], DrawTextCreative [39], ChineseDrawText [41], and Mario-10M test set [12]. Following the protocol [12], we exclude images generated from ChineseDrawText and DrawBenchText, hence 5,000 images are used for measuring FID.

Benchmark for multi-lingual generation. Due to the lack of pre-existing benchmarks for MLT image generation, we developed a new benchmark for model evaluation on nine different languages. We create 25 prompt templates, such as "A raccoon holding a paper saying words.". We exclude the non-English target words from the prompts, as they cannot be processed by the text encoder. Each prompt is designed to include up to three words in the image. Except for English, we allocate the rest of eight languages a set of 200 prompts, and 400 for English, yielding 2,500 prompts in total. For computing FID, we construct a set of images by randomly sampling a subset of the ICDAR2019 dataset.

Benchmark for logo generation. We additionally construct a benchmark for logo image generation. Specifically, we create 25 template prompts, e.g., "A building displaying a [KEYWORD] sign.", where [KEYWORD] here is the name of the logo. We obtain 150 logos in total, which is the union of the class of FlickrLogos-32 dataset [26] and LITW dataset, yielding 1,500 prompts in total. For computing FID, we construct an image set sampled from the merged set of FlickrLogos-32 [49] and LITW [61] dataset.

A.5. Text Image Editing Details

We elaborate on the extension of our model for text image editing. In line with [12], we augment the input channel by incorporating additional encoded latent obtained from masked images. Specifically, we use VAE to encode an image containing masked regions and concatenate the encoded latent with the original model input. We use the same VAE used for reference image encoding to encode the masked image, and noise is not added to the resulting latent. Consequently, we obtain an augmented input $\tilde{\mathbf{x}}_i^t \in \mathcal{R}^{(3c+1)\times h\times w}$ as depicted below:

$$\tilde{\mathbf{x}}_i^t = \text{concat}(\mathbf{z}_i^t, \mathbf{s}_i, \mathbf{m}_i, \mathbf{z}_i^{\text{masked}}), \tag{10}$$

where $\mathbf{z}_i^{\text{masked}} \in \mathcal{R}^{c\times h\times w}$ denotes VAE encoded latent of the masked image. During training, along with the text region, we randomly select an additional 1-3 random regions to be masked. During inference, we mask the region to be

modified. We provide additional qualitative results in Appendix C.

B. Additional Ablation Studies

In this section, we discuss additional ablation studies.

B.1. Ablation Study on Languages

Per language comparison. We provide OCR Accuracy for each language and compare it with the existing methods in Table A. Remarkably, our model is shown to be the most accurate, as it surpasses existing methods across all average accuracy metrics, including the average of Latin, non-Latin, and the entire language set by a significant margin. Moreover, our method achieves high English accuracy comparable to that of TextDiffuser, which was tailored for English image generation. Although TextDiffuser shows high accuracy in some languages, such as English and German, it shows incapability in the rest of the language set, as it achieves 0 accuracy. In contrast to TextDiffuser, our method can generate text in all languages, and this demonstrates the multi-lingual capability of our model.

Latin *v.s.* non-Latin. Our model exhibits higher OCR Accuracy scores in Latin languages such as English, Italian, German, and French (Tables A and B). It is noteworthy that for Russian, Thai, and Greek, our model is exclusively trained on synthetic images, yet it achieves comparable or superior FID and CLIP scores compared to the results for Latin languages. Given that the model was pretrained on a large-scale English dataset [12], we speculate that this contributes as one factor to its superior performance in Latin languages.

Different levels of complexity among languages. We note that different languages present different levels of complexity, with some being more challenging to generate accurately (Tables A and B). For instance, the size of the alphabet in English is 26, whereas in Thai, the size is 59. In addition, in languages such as Thai and Greek, diacritics are combined with other letters, and extra complexity is added. Some of the predictions that our model makes look similar to the actual word, but are evaluated to be incorrect due to the mis-generation of diacritics. Besides, the shapes of characters in non-Latin languages (e.g., Bengali), are notably more intricate than Latin characters, whose alphabet typically requires a greater number of strokes to form.

B.2. Ablation Study on CFG scheduling

We provide an ablation study on the proposed Augmented Classifier-Free Guidance with scheduling detailed in Appendix A.2 in Appendix A.2. For better reference, we repeat the Equation (9) as below:

$$\begin{aligned}\hat{\epsilon}^t = & \underbrace{\epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)}_{\text{Unconditional}} \\ & \underbrace{+\omega_{\text{prompt}}\Big[\epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \mathbf{c}, t) - \epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]}_{\text{Caption Guidance}} \\ & \underbrace{+\,\omega_{\text{ref}}\Big[\epsilon_\theta(\text{concat}(\mathbf{z}^t, \mathbf{s}, \mathbf{m}), \varnothing, t) - \epsilon_\theta(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]}_{\text{Reference Guidance}} \\ & \underbrace{+\,\omega_{\text{all}}\Big[\epsilon(\text{concat}(\mathbf{z}^t, \mathbf{s}, \mathbf{m}), \mathbf{c}, t) - \epsilon(\text{concat}(\mathbf{z}^t, \varnothing, \varnothing), \varnothing, t)\Big]}_{\text{All Guidance}},\end{aligned}$$

where ω^t_{ref} and ω^t_{prompt} are the values of guidance scale for reference and prompt condition at timestep t, and ω_{all} is a guidance scale for both conditions which is set to be a constant. We dynamically schedule the guidance scales as follows:

$$\omega^t_{\text{ref}} = \gamma \frac{t}{\overline{T}}^{\rho_{\text{speed}}}$$
$$\omega^t_{\text{prompt}} = \gamma(1 - \frac{t}{\overline{T}}^{\rho_{\text{speed}}}),$$

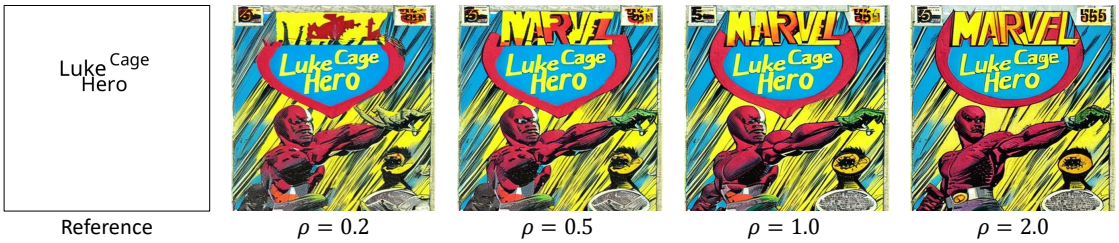
where t is counted in decreasing order, hence ω_{ref} is scheduled to be decreasing, and ω_{prompt} to be increasing. Here, γ is a constant value corresponding to the max/min value of $\{\omega^t_{\text{r}}\}$ and $\{\omega^t_{\text{c}}\}$, and ρ_{speed} denotes the speed of each guidance schedule.

Impact of the scheduling speed. We first study the impact of CFG scheduling by varying ρ_{speed}, the increase/decrease rate of the guidance scale, in Table C. We observe a rapid decrease in ω_{ref} with a rapid increase in ω_{ref} result in a decrease in the OCR Accuracy and improvements in FID and CLIP scores. As ω_{ref} decreases rapidly, the impact of the prompt gets higher and the reference condition becomes less influential in the generation process. As a consequence, non-target words within the prompt gain increased influence and affect the generation, resulting in the inclusion of non-target words in the output (Figure A).

Impact of different portions of scheduled guidance. We examine the effect of different portions of scheduled guidance γ within the total amount of guidance (Table D). The result of CFG without scheduling is also provided for comparison. We observe the general trend of a decrease in CLIP scores and an increase in OCR Accuracy as γ increases. We speculate that the initial guidance scale of the reference condition plays an important role in establishing the initial layout of the text to be generated which, in turn, influences the effect of the prompt condition and consequently affects CLIP and FID.

Method	English	Italian	German	French	Hindi	Bengali	Russian	Thai	Greek	Latin	Non-Latin	Mean	FID
ControNet	15.00	12.00	08.50	07.00	08.50	11.00	15.00	6.00	19.50	10.63	12.00	11.75	119.996
SD	0.00	0.50	0.0100	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.15	**93.198**
TextDiffuser	**84.00**	**59.00**	**62.50**	27.50	0.00	0.00	0.00	0.00	0.00	58.25	0.00	31.70	96.826
Ours	70.00	48.00	56.00	51.50	21.00	**11.50**	38.00	15.50	33.00	56.38	23.80	41.45	97.523
Ours†	81.75	58.50	60.50	**57.50**	**22.00**	10.00	**48.00**	**18.00**	**48.00**	**64.56**	**29.20**	**48.60**	97.156

Table A. **OCR Accuracy comparison result for MLT image generation, presented individually for each language.** Results marked with † indicate the utilization of CFG scheduling. We set $\rho_{\text{speed}} = 0.2$, $\gamma = 3.5$ and $\omega_{\text{all}} = 4$ for CFG scheduling.



"Marvel Comics Retro 'Luke Cage Hero' for Hire Comic Book Cover No15 in Chains aged"

Figure A. **Text-to-image generation with different CFG scheduling speeds.** The utilized reference image is denoted on the left where the target words to be generated are rendered. The text prompt with text words to be generated enclosed with punctuations is denoted at the bottom. Note the non-target word 'MARVEL' starts to appear at the top of each image as ρ_{speed} increases. We use identical random seeds for all the generations.

Script	Language	FID↓	CLIP↑	Acc↑	F-1↑
Latin	English	157.954	0.2977	70.00	87.80
	Italian	195.653	0.2946	48.00	70.43
	German	195.230	0.2918	56.00	78.47
	French	190.083	0.2942	51.50	70.01
Non-Latin	Hindi	193.788	0.3125	21.00	37.54
	Bengali	192.716	0.3135	11.50	17.60
	Russian	188.567	0.2998	38.00	56.89
	Thai	203.457	0.3015	15.50	27.59
	Greek	192.215	0.3023	33.00	55.20
	Average	97.523	0.3006	41.45	58.94

Table B. **Evaluation results for MLT image generation, presented individually for each language.**

B.3. Ablation Study on Hyper Parameters

Impact of synthetic loss weight. We first analyze the impact of the synthetic loss weight in Table E. We note that the synthetic loss weight mainly impacts the **CLIP** as the text prompts of the real dataset better describe the content of an image. We observe the model trained with a synthetic loss weight of 0.5 achieves a lower **CLIP** than the one trained with 1.0, which validates the effectiveness of lowering the weight of reconstruction loss when synthetic images are provided.

Impact of recognition loss weight. We analyze the impact of the recognition loss weight on accuracy and the quality of the generated results in Table F. We note that the OCR accuracy improves as the model is trained with increased recognition loss weight. In general, models with higher loss weight achieve better OCR Accuracy and a worse score on FID. We choose the recognition weight as 0.025 for a proper trade-off.

Impact of LoRA rank. We examine the influence of various rank configurations for LoRA [23] and provide justification for the selected rank. We assess models trained with different rank values trained with the MARIO-10M dataset (Table G). We observe a consistent pattern of performance improvement with increasing rank values, then saturates at $r = 32$, thus we opt for a rank setting of $r = 32$ as our final choice.

B.4. Analysis on Reference Image.

Impact of reference image during inference. To study the impact of the appearance of the reference image, we compare generation results using the reference image with and without subject rendering *i.e.,* blank reference in Tables H and I. The results show that the generation result is greatly influenced by the appearance of the image. We also provide qualitative results in Figure B.

B.5. Ablation Study on SD Versions

To show generalization over different versions, we provide results on SD-v1.5 in Table J. Our method effectively improves on both versions.

ρ_{speed}	γ	ω_{all}	FID↓	CLIP↑	Acc↑
0.2	3.5	4	45.355	0.3406	**64.65**
0.5	3.5	4	44.680	0.3419	63.11
1.0	3.5	4	**43.884**	**0.3428**	60.53

Table C. **Evaluation results on MARIO-EVAL with varying CFG scheduling speeds ρ_{speed}.**

ρ_{speed}	γ	ω_{all}	FID↓	CLIP↑	Acc↑
0.5	0.5	7	**38.104**	**0.3454**	59.84
0.5	1.5	6	39.932	0.3448	59.25
0.5	2.5	5	42.192	0.3434	61.73
0.5	3.5	4	44.680	0.3419	62.51
0.5	4.5	3	49.827	0.3317	**63.85**
-	0	7.5	38.593	**0.3454**	58.26

Table D. **Evaluation results on MARIO-EVAL with varying portions of scheduled guidance γ.** $\gamma = 0$ denotes the result of the CFG without the guidance scheduling.

Weight	CLIP↑
0.5	0.3011
1.0	0.2997

Table E. **Ablation study on synthetic loss weight for multi-lingual image generation on Russian subset.**

Weight	Acc↑	FID↑
0.025	43.47	89.97
0.05	46.13	92.82

Table F. **Ablation study on recognition loss weight for logo image generation.** Results are denoted in [%].

Rank	F-1 ↑	CLIP↑	# Lora Params (M)↓
1	51.58	0.3359	0.89
4	69.44	0.3372	3.57
32	**70.52**	**0.3445**	**28.55**
128	70.19	0.3436	114.21

Table G. **Ablation study for different rank settings of LoRA [23].** The model is trained on MARIO-LAION [12]. The number of parameters indicated on a million scale. We denote with **bold** for the selected rank configuration.

Type	Acc ↑	F1 ↑	CLIP
Blank Reference	20.27	27.18	0.3695
Ours	**42.87**	**48.86**	**0.3759**

Table H. **Impact of subject existence in logo generation.**

Type	Acc ↑	F1 ↑	CLIP
Blank Reference	1.88	2.71	0.2855
Ours	**55.19**	**79,24**	**0.3468**

Table I. **Impact of subject existence in English generation.**

Model	Acc↑	CLIP↑
SD-v1.5	0.03	0.3015
Ours-v1.5	**51.40**	**0.3647**
SD-v2.1	0.02	0.3221
Ours-v2.1	**49.20**	**0.3685**

Table J. **Evaluation of different SD versions on the OpenLibraryEval500 subset.**

B.6. Ablation Study on Simultaneous Task Training

Our method employs separate plugin modules, each acting as an expert for a specific task. When a task is prompted, its corresponding expert plugin is loaded. This modular design reduces interference between tasks, which often occurs when training with diverse objectives. Experimental results confirm this advantage: simultaneous task training leads to degraded performance (Table K).

Training Task	Acc↑
English	**58.26**
English+Logo	51.64

Table K. **Simultaneous training result for English generation.**

Impact of subject location in reference. We present a visual analysis of how the spatial position of the subject in the reference image affects the generated outputs. As illustrated in Figure D, the subject's location in the reference image is consistently reflected in its position within the generated results.

Joint influence of prompt and reference. We observe that *both* the input prompt and the reference image influence the generation results. To show this, we visualize the generation result with the same reference image, but with a different input prompt. As shown in Figure C, the generation results not only depend on the appearance of the subject in the reference image, but also on the input prompt.

C. Additional Qualitative Results

In this section, we provide additional qualitative results.

Additional qualitative results for English texts. We provide additional qualitative results for English text image generation (Figure E). The model fine-tuned on AText is adopted for the English image generation. We use the MARIO-EVAL benchmark prompts [12] for the generation.

Additional qualitative results for multi-lingual texts. We provide additional qualitative results for the multilingual, Latin text image generation (Figure F), and multilingual, non-Latin text image generation (Figure G). We use

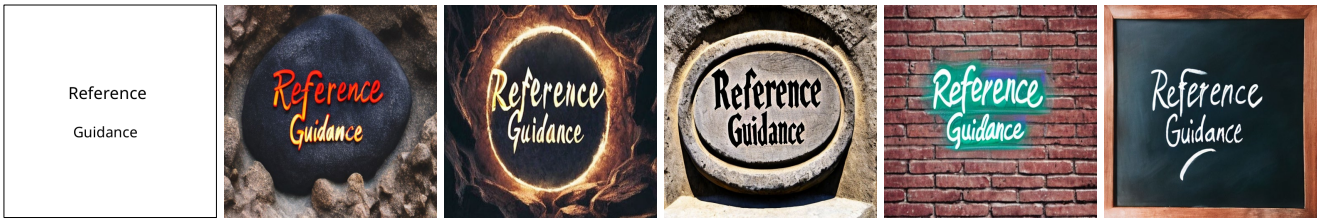
Figure B. **Generation result with blank reference image.**

the model trained on the merged set of ICDAR2019 [14], and synthetic images of all the languages. Despite the deficiency of real samples of Russian, Greek, and Thai, our model shows the capability of generalizing to these languages.

Additional qualitative results for logos. We present additional qualitative results of logo image generation. We first present the generation results of the logos that are included in the training set (Figure H). The results are generated using the logo benchmark prompts. The generated results confirm the model's capability of generating logos in the desired location of the corresponding reference logo image, and that this reference extends beyond text renderings. Moreover, we present the generation results of the logos that are never seen during training (Figure I). The results further validate the effectiveness of the proposed method and the model's ability to generalize across novel instances.

Comparison results with personalization methods. We provide comparison results with personalization methods in Figure J.

Additional text image editing results. We provide additional text image Editing results in Figure K. The model successfully edits the specified region to include the intended words without modifying the rest of the region.



Reference Image

‘Reference Guidance’ in fiery lava font on volcanic rock

‘Reference Guidance’ glowing in a fiery runic over a dark cave

‘Reference Guidance’ engraved in an ancient carved stone font on a temple wall

‘Reference Guidance’ in neon graffiti on a brick wall

‘Reference Guidance’ written in a chalk drawn font on a blackboard

Reference Image

‘Take Easy’ in fiery lava font on volcanic rock

‘Take Easy’ glowing in a fiery runic over a dark cave

‘Take Easy’ engraved in an ancient carved stone font on a temple wall

‘Take Easy’ in neon graffiti on a brick wall

‘Take Easy’ written in a chalk drawn font on a blackboard

Reference Image

‘Dream Big’ in fiery lava font on volcanic rock

‘Dream Big’ glowing in a fiery runic over a dark cave

‘Dream Big’ engraved in an ancient carved stone font on a temple wall

‘Dream Big’ in neon graffiti on a brick wall

‘Dream Big’ written in a chalk drawn font on a blackboard

Reference Image

‘Stay Strong’ in fiery lava font on volcanic rock

‘Stay Strong’ glowing in a fiery runic over a dark cave

‘Stay Strong’ engraved in an ancient carved stone font on a temple wall

‘Stay Strong’ in neon graffiti on a brick wall

‘Stay Strong’ written in a chalk drawn font on a blackboard

Figure C. **Generation result with the same reference image.** The images in the same row are generated with the same reference, but with different input prompts.

‘Welcome Home’ spelled in a marquee lights

‘Happy Days’ printed on a stop sign

‘Dream Team’ in neon icing on a glazed cookie

Figure D. **Generation result with varying position.**

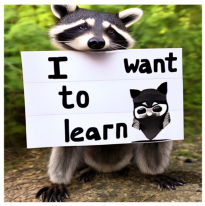
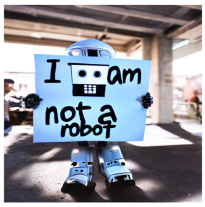
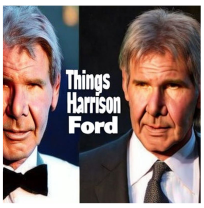
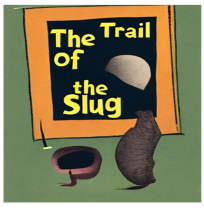
A volunteer is carrying a backpack with the words 'Love' printed on it		A globe with the words 'Planet Earth' written in bold letters, with continents in bright colors		Little raccoon holding a sign that reads 'I want to learn'	
Studio shot of sculpture of text 'cheese' made from cheese, with cheese frame.		A cranky sunflower with a 'No Solar Panels' sign		dslr portrait of a robot is holding a sign with text 'i am not a robot'	
NATURE S 'EARTHLY CHOICE'		studio shot, word 'wow' in script made from rainbow colored fur, in a furry frame, white background, centered		A little girl is holding a book with the words 'Fairy Tales' in her hands	
'Lake Tahoe 2020'		'Flight of the Conchords'		'Keep Calm and Trust a Press Officer' Posters	
15 'Things' You Didn t Know About 'Harrison Ford'		'THE' ESSENTIAL 'HOWARD SHORE'		'Cherry Land' Door 'County Wisconsin' Post Cards	
A book with a title text of 'Bahrain'		A cover titled 'The Trail of the Slug'		A book cover with logo 'The Latest Thing And Other Things' on it	
A TV show poster titled 'Underworld Blood Wars'		A poster with a title text of 'Trouble in Paradise'		A poster design with a title text of 'Tokyo Halloween Night'	

Figure E. **Additional results for English generation.** The odd columns are the prompts, and the even columns are the generated images. Words enclosed with punctuation are the target keywords to be generated.

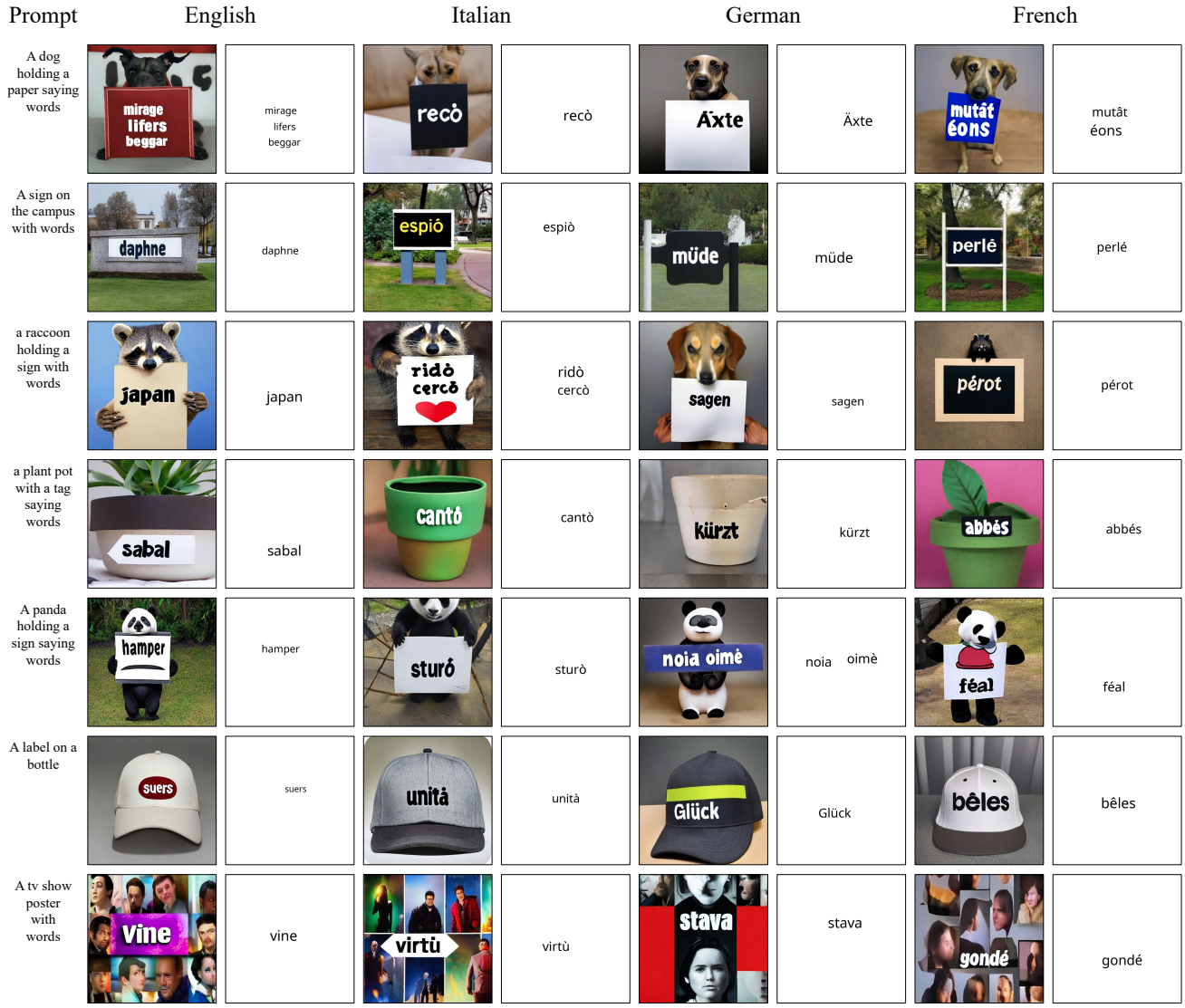


Figure F. **Additional qualitative results for multi-lingual, Latin text image generation.** Images in the same row correspond to the same prompt templates. The first column denotes the prompts. The even column denotes the generated results, and the odd column denotes the reference image of text renderings.

Figure G. **Additional qualitative results for multi-lingual, non-Latin text image generation.** Images in the same row correspond to the same prompt templates. The first column denotes the prompt template. The even column denotes the generated results, and the odd column denotes the reference image of text renderings.

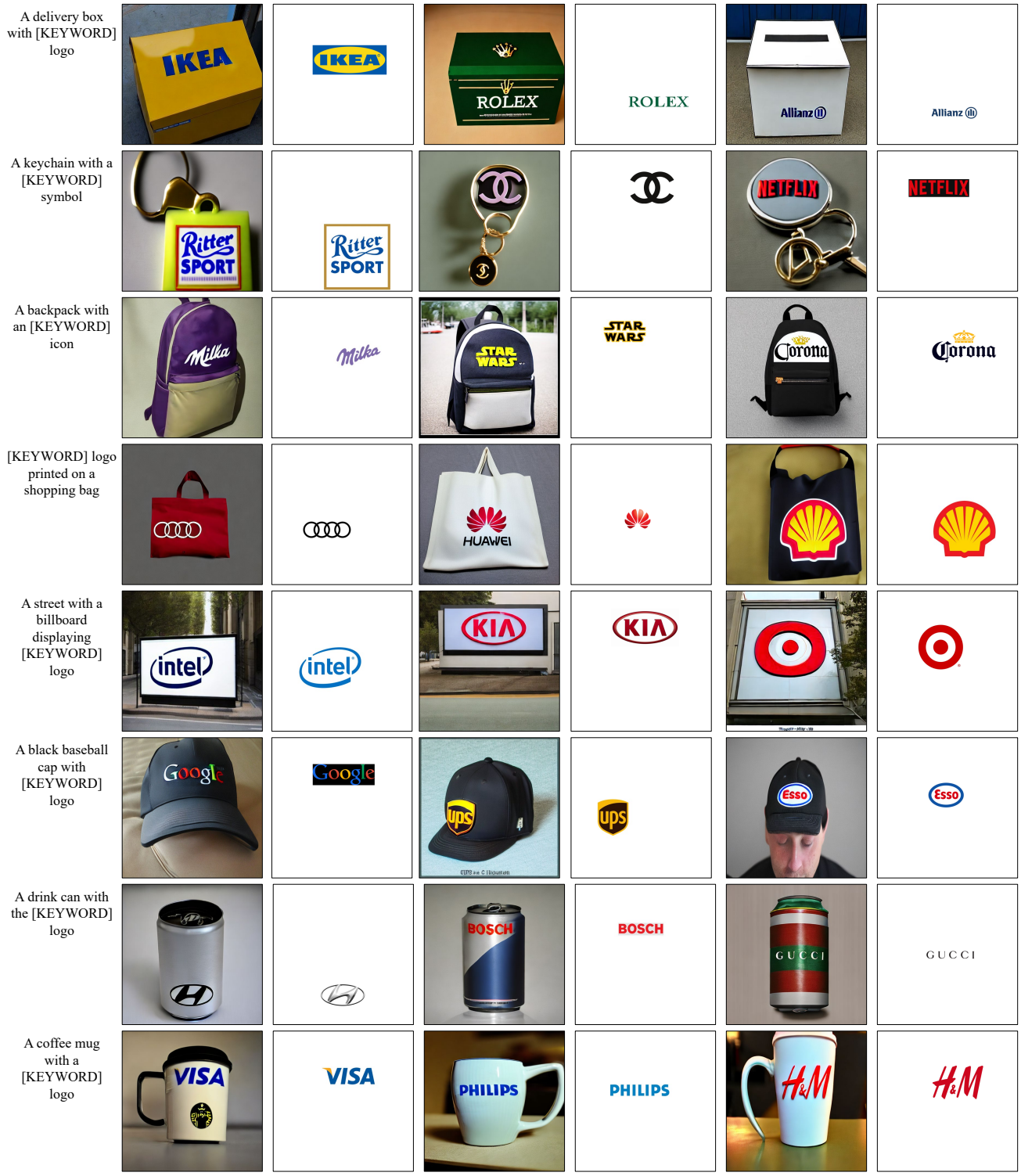


Figure H. **Additional qualitative results for logo image generation.** The reference logos utilized here are part of the training logo set. Images in the same row correspond to the same prompt templates. The first column denotes the prompt template. We replace "[KEYWORD]" with the name of the logo during the inference. The even columns denote the generated results, and the odd columns denote the reference logo images.

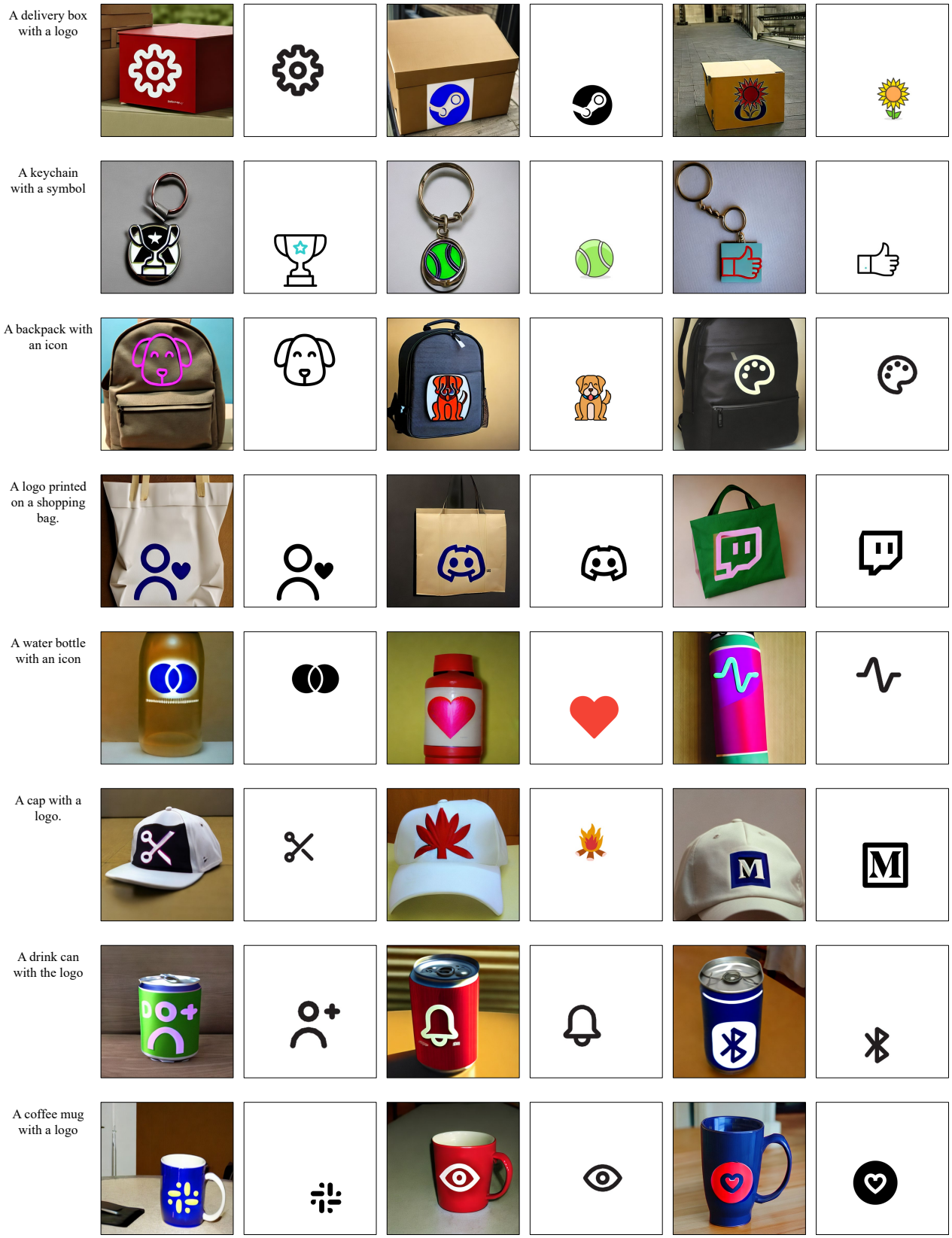


Figure I. **Additional qualitative results for logo image generation.** The reference logos utilized here are not included in the training logo set and hence, are unseen during the model training. Images in the same row correspond to the same prompt templates. The first column denotes the prompt template. The even columns denote the generated results, and the odd columns denote the reference logo images.

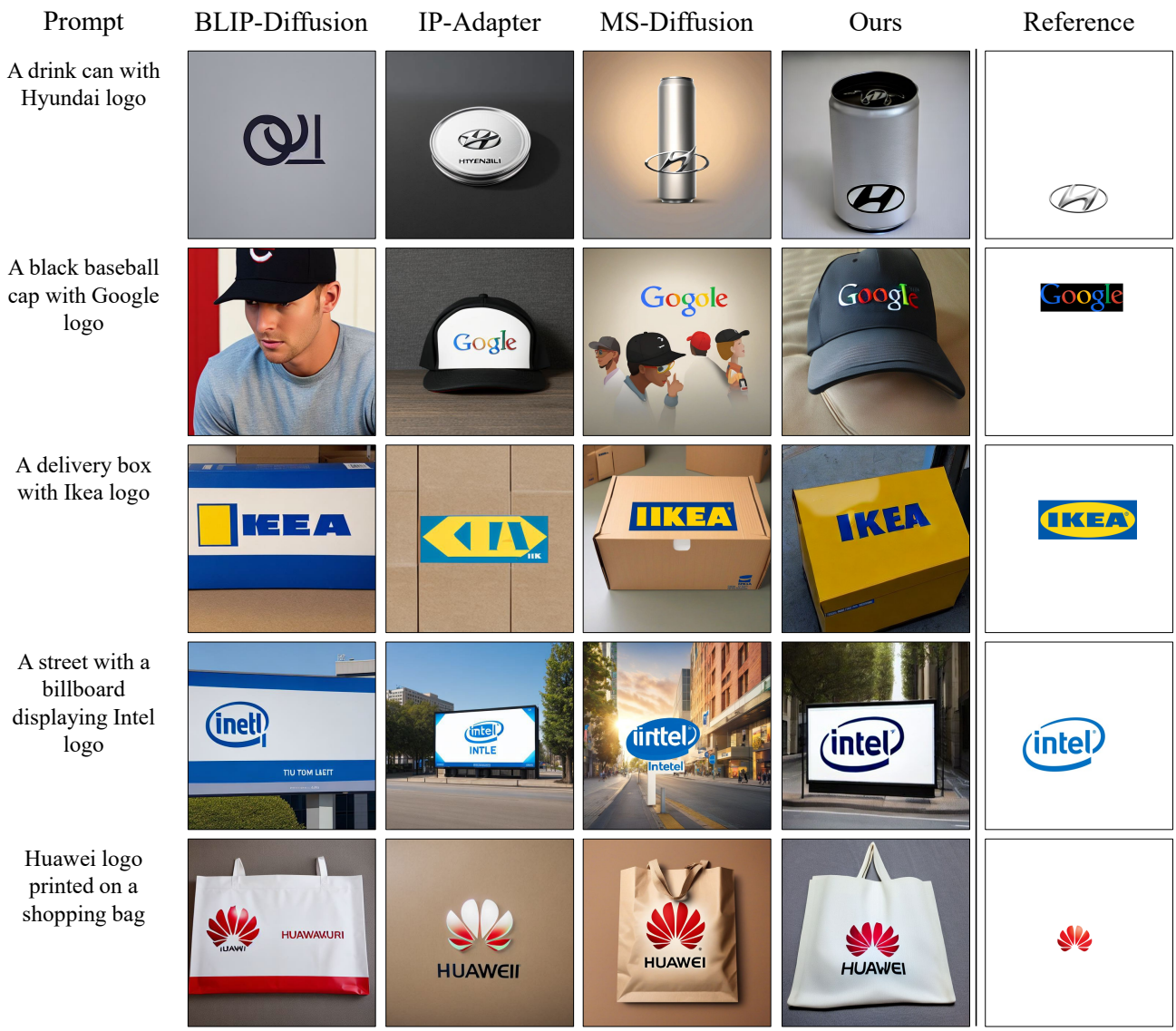


Figure J. **Comparison results with personalization methods.**

“DOLPHINS”

“REFDIFFUSER”

“Together”

“Batman”

“Good”, “Morning”

“Feed”, “Me”

“Leaving”, “Coming”

“NarrowWay”

“ECCV”

“Burgerking”

Figure K. **Additional qualitative results for Text Image Editing.** Regions denoted in yellow are masked during the inference. We denote the added text below each of the results.

References

[1] Project gutenberg's adventures of sherlock holmes. https://www.gutenberg.org/files/48320/48320-h/48320-h.htm. 2

[2] Deepfloyd. https://www.gutenberg.org/files/48320/48320-h/48320-h.htm. 1, 2, 7

[3] Foundation icon fonts 2. https://zurb.com/playground/foundation-icons. 2

[4] LAION-ASTHETIC. https://laion.ai/blog/laion-aesthetics/. 2

[5] Pexels. https://www.pexels.com/. 4, 2

[6] Tmdb movie metadata. https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata. 2

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. 2017. 2

[8] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. 2023. 2

[9] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2

[10] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, 2022. 2

[11] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[12] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. 2023. 1, 2, 3, 4, 5, 6, 7

[13] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. 2024. 2

[14] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. ICDAR2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *International Conference on Document Analysis and Recognition*, 2019. 4, 5, 2, 6

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. 7

[16] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022. 2

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. 2014. 2

[19] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4, 6

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 5, 1

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. 2, 4

[23] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 4, 5

[24] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In *CVPR*, 2022. 2, 3

[25] Mingxin Huang, Jiaxin Zhang, Dezhi Peng, Hao Lu, Can Huang, Yuliang Liu, Xiang Bai, and Lianwen Jin. Estextspotter: Towards better scene text spotting with explicit synergy in transformer. In *CVPR*, 2023. 2

[26] Yannis Kalantidis, Lluis Garcia Pueyo, Michele Trevisiol, Roelof van Zwol, and Yannis Avrithis. Scalable triangulation-based logo recognition. In *Proceedings of ACM International Conference on Multimedia Retrieval*, 2011. 6, 2

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[28] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. 2022. 2

[29] Taeho Kil, Seonghyeon Kim, Sukmin Seo, Yoonsik Kim, and Daehee Kim. Towards unified scene text spotting based on sequence generation. In *CVPR*, 2023. 2

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2

[32] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2023. 7

[33] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. 2016. 2

[34] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *NeurIPS*, 2023. 7

[35] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[36] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet ++:

Improving conditional controls with efficient consistency feedback. In *ECCV*, 2024. 2

[37] 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. 2

[38] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *ECCV*, 2020. 2

[39] Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. In *arXiv preprint arXiv:2212.10562*, 2022. 2

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[41] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *arXiv preprint arXiv:2303.17870*, 2023. 1, 2

[42] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. 2022. 2

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 7

[44] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arxiv 2015. *ICLR*, 2016. 2

[45] 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 *ICLR*, 2021. 2, 3

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 21(1):5485–5551, 2020. 1, 2

[47] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. 2015. 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 6, 7

[49] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. Scalable logo recognition in real-world images. In *Proceedings of ACM International Conference on Multimedia Retrieval*, 2011. 6, 2

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. 3

[51] 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. 2

[52] 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. 2022. 1, 2

[53] C Schuhmann, R Vencu, R Beaumont, R Kaczmarczyk, C Mullis, A Katta, T Coombes, J Jitsev, and A LAION Komatsuzaki. 400m: Open dataset of clip-filtered 400 million image-text pairs. arxiv 2021. *arXiv preprint arXiv:2111.02114*. 2

[54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. 2022. 5, 6, 2

[55] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 2

[56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015. 2

[57] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. 2015. 2

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 2, 6

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 2

[60] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *ICLR*, 2023. 2, 7

[61] Andras Tüzkö, Christian Herrmann, Daniel Manger, and Jürgen Beyerer. Open Set Logo Detection and Retrieval. In *Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications: VISAPP*, 2018. 6, 2

[62] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. 30, 2017. 2

[63] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *ICLR*, 2025. 7

[64] Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Binary latent diffusion. In *CVPR*, 2023. 2

[65] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 2

[66] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. 2024. 2

[67] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. 2024. 2, 7

[68] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael

Zeng, et al. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023. 2

[69] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 7

[70] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. Deepsolo: Let transformer decoder with explicit points solo for text spotting. In *CVPR*, 2023. 2

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