Appendix

A. Additional Results

A.1. Qualitative Samples

Fig. S1 and Fig. S2 present additional qualitative samples generated by DCText. As shown in the results, our method consistently produces accurate text and high-quality images across a wide range of themes, including illustrations, real scenes, posters, artistic styles, and animations. Notably, these results are achieved without any additional training, relying solely on scheduled attention masking at inference time. In particular, Fig. S2 demonstrates that the textual regions faithfully reflect the target sentences, even when multiple texts need to be rendered. At the same time, the overall image remains thematically aligned with the global prompt. This is enabled by DCText's attention masks (M_{focus}, M_{expn}), which effectively regulate the information flow between the target regions and the background.

A.2. Quantitative Results

Tab. S1 provides a comprehensive comparison across all datasets and baselines. To support broader evaluation, we additionally report baseline results obtained using the same number of denoising steps as our method, indicated by †. For text accuracy metrics (Acc. and NED), DCText consistently outperforms all baselines across all datasets. This underscores the effectiveness of our divide-and-conquer strategy, which generates text segments rather than entire sentences at once, resulting in more reliable text generation. In terms of overall image quality (Qual. and Aesth.), our method also achieves high scores on most datasets, indicating that improvements in text accuracy do not come at the expense of image quality. In addition, when baselines are evaluated under the same reduced number of denoising steps, their performance typically declines across most metrics. Although this setting reduces inference latency, our method still maintains the lowest latency, with strong performance.

A.3. Comparison with Training-based Methods

In the main paper, we compare DCText with other approaches that, like ours, leverage a pre-trained text-to-image model without additional training. We further compare our method against training-based approaches, including AnyText [18], GlyphControl [20], TextDiffuser2 [4], and EasyText [13]. Tab. S2 presents a quantitative comparison on the single-sentence datasets. In terms of text accuracy (Acc. and NED), training-based baselines—trained on large-scale text-centric datasets (AnyWord-3M [18], LAION-Glyph [20], MARIO-10M [3], and EasyText-1M [13]) with glyph-level conditioning—generally achieve higher scores. However, this comes at the cost of overall image quality. As reflected by their low aesthetic scores (Aesth.)—and as visually confirmed in Fig. S3—training-based methods lack stylistic diversity and fail to produce artistic text. For instance, in columns 3 and 4, where the prompts specify rendering text with fur and vines, these methods generate plain, generic text instead of following the intended styles. In the more challenging multi-sentence setting, training-based approaches also struggle. As shown in Tab. S3, their text accuracy degrades significantly as the number of sentences (n) increases. In contrast, DCText maintains consistent performance and outperforms all baselines across different sentence counts.

A.4. Comparison to Regional-Prompting

To highlight DCText's performance in visual text generation, we compare it with Regional-Prompting [2], a method that relies solely on the Region-Isolation Attention Mask (M_{isol}) for inference-time attention control. As discussed in Sec. 3.2, the exclusive use of M_{isol} often results in redundant text rendering and unnatural regional artifacts (see Fig. 6, left). Regional-Prompting addresses this by replacing the global prompt with a background-only prompt, removing all content information, and additionally performs a separate denoising process with the original global prompt, spatially blending the two resulting latents. However, this approach not only doubles the number of function evaluations (NFEs), but also remains less effective for visual text generation. As shown in Fig. S4, Regional-Prompting often produces illegible or semantically meaningless text. This occurs because the fine-detailed visual features required for faithful text rendering are diluted during latent blending. In contrast, DCText generates text that is both accurate and natural, while also requiring only 15.66 seconds per image generation compared to 27.79 seconds for Regional-Prompting. This demonstrates that our two novel masks—Text-Focus Attention Mask (M_{focus}) and Context-Expansion Attention Mask (M_{expn})—enable effective and efficient attention control for visual text generation.

B. Additional Ablation Studies

B.1. Text-Focus Attention Mask Design

To construct M_{focus}, we take the union of four partial masks ($M_{r^c \to \{r_i\}}, M_{\{p_i\} \to r^c}, M_{p_g \to \{r_i\}}, M_{\{p_i\} \to p_g}$), each of which enables a specific directional attention flow. To evaluate the contribution of each component, we conduct an ablation study by selectively excluding individual partial masks. The results are shown in Fig. S5 and Tab. S4. When the attention flow from the background region to the textual regions is disabled (*i.e.* w/o $M_{r^c \to \{r_i\}}$), duplicated text tends to appear in the background, and the transition between background and target regions becomes unnatural. The awkward bright areas of the first row in Fig. S5 illus-

Dataset	n	Method	Acc.	NED	CLIP	Qual.	Aesth.	Steps	Latency (sec.)
CreativeDrawText	1	Flux	0.257	0.544	0.339	4.695	3.718	24	13.89
		AMO	0.261	0.559	0.339	4.691	3.693	28	25.93
		AMO†	0.193	0.524	0.338	4.697	3.707	24	21.71
		TextCrafter	0.330	0.764	**0.346**	4.726	**3.767**	30	36.89
		TextCrafter†	0.330	0.750	**0.346**	4.727	3.722	24	28.61
		DCText (Ours)	**0.427**	**0.809**	0.345	**4.775**	3.761	24	16.60
DrawTextCreative	1	Flux	0.223	0.525	0.351	4.657	3.774	24	13.81
		AMO	0.234	0.499	0.351	4.670	3.831	28	25.99
		AMO†	0.234	0.516	0.350	4.624	3.755	24	21.76
		TextCrafter	0.286	0.645	0.351	4.698	3.848	30	36.92
		TextCrafter†	0.286	0.664	**0.353**	4.691	3.855	24	28.61
		DCText (Ours)	**0.337**	**0.675**	0.350	**4.765**	**3.934**	24	16.61
TMDBEval500	1	Flux	0.318	0.667	0.340	4.618	4.061	24	13.88
		AMO	0.326	0.648	0.336	4.614	4.064	28	25.89
		AMO†	0.340	0.665	0.338	4.607	4.041	24	21.74
		TextCrafter	0.372	0.758	**0.352**	**4.687**	**4.119**	30	36.91
		TextCrafter†	0.358	0.744	0.351	4.661	4.074	24	28.63
		DCText (Ours)	**0.396**	**0.768**	0.351	4.670	4.018	24	16.56
CVTG-Style	2	Flux	0.608	0.809	0.344	4.675	3.471	24	13.88
		AMO	0.642	0.826	0.341	4.654	3.584	28	25.91
		AMO†	0.622	0.820	0.341	4.664	3.580	24	21.98
		TextCrafter	0.758	0.919	0.346	4.697	3.616	30	38.22
		TextCrafter†	0.745	0.923	0.348	4.688	3.584	24	28.97
		DCText (Ours)	**0.792**	**0.923**	**0.347**	**4.791**	**3.726**	24	15.66
	3	Flux	0.508	0.715	0.345	4.662	3.377	24	13.88
		AMO	0.575	0.750	0.343	4.689	3.489	28	25.99
		AMO†	0.540	0.741	0.342	4.679	3.480	24	21.86
		TextCrafter	0.722	0.880	**0.351**	4.659	3.571	30	39.01
		TextCrafter†	0.710	0.882	0.350	4.670	3.595	24	29.47
		DCText (Ours)	**0.768**	**0.906**	**0.351**	**4.735**	**3.709**	24	16.96
	4	Flux	0.389	0.628	0.338	4.707	3.421	24	13.93
		AMO	0.488	0.690	0.337	4.709	3.518	28	25.97
		AMO†	0.469	0.689	0.337	4.709	3.515	24	21.97
		TextCrafter	0.693	0.867	0.352	4.665	3.488	30	39.60
		TextCrafter†	0.722	0.877	0.354	4.697	3.530	24	30.07
		DCText (Ours)	**0.760**	**0.892**	**0.353**	**4.745**	**3.659**	24	18.14
	5	Flux	0.366	0.608	0.336	4.591	3.165	24	13.91
		AMO	0.432	0.660	0.335	4.652	3.218	28	26.02
		AMO†	0.402	0.636	0.336	4.618	3.190	24	21.94
		TextCrafter	0.685	0.859	**0.349**	4.659	3.506	30	40.53
		TextCrafter†	0.661	0.848	0.343	4.562	3.396	24	30.78
		DCText (Ours)	**0.693**	**0.860**	0.343	**4.697**	**3.569**	24	19.26
Average	-	Flux	0.381	0.642	0.342	4.658	3.570	24	13.88
		AMO	0.423	0.662	0.340	4.668	3.628	28	25.96
		AMO†	0.400	0.656	0.340	4.657	3.610	24	21.84
		TextCrafter	0.549	0.813	**0.350**	4.684	3.702	30	38.30
		TextCrafter†	0.545	0.813	0.349	4.671	3.679	24	29.31
		DCText (Ours)	**0.596**	**0.833**	0.349	**4.740**	**3.768**	24	17.11

Table S1. **Full quantitative comparison.** Comparison results with baselines on four datasets: ChineseDrawText [14], DrawTextCreative [12], TMDBEval500 [3], and CVTG-Style [5]. † indicates methods that use the same number of denoising steps as DCText.

Method	Acc.	NED	Qual.	Aesth.
AnyText	0.096	0.442	4.128	2.990
GlyphControl	0.630	0.901	3.935	2.884
TextDiffuser2	0.552	0.860	3.463	2.488
EasyText	0.159	0.484	4.380	3.361
DCText (Ours)	0.387	0.751	4.737	3.904

Table S2. **Quantitative comparison between training-based baselines.** Results are averaged over three single-sentence datasets (ChineseDrawText, DrawTextCreative, TMDBEval500).

trate this issue. Disabling the attention flow from textual prompts to the background region (w/o $M_{\{p_i\}\to r^c}$) often causes incorrect text to be generated, significantly lowering text accuracy. When the attention from the global prompt to the textual regions is blocked (w/o $M_{p_g\to\{r_i\}}$), irrelevant text tends to appear. On the other hand, removing the attention from textual prompts to the global prompt (w/o $M_{\{p_i\}\to p_g}$) produces text that is less stylistically aligned with the overall image. Overall, incorporating all four directional attention flows results in the highest text accuracy and consistently high image quality.

B.2. Text-Focus Denoising Steps

Fig. S6 and Tab. S5 present the results of an ablation study on varying the number of denoising steps T_{focus} during which the text-focus attention mask M_{focus} is applied, while keeping T_{init} and T_{expn} fixed. As shown in the figure, when $T_{\text{focus}} = 0$, that is, when M_{focus} is not applied, the model fails to focus on the designated region, often producing region-irrelevant or entirely missing text. As T_{focus} increases, alignment between the generated text and the target region improves. However, excessive values of T_{focus} prevent regions from attending to the background for extended periods, leading to more noticeable boundaries between regions and their surroundings, and, as shown in the table, even causing a decline in text accuracy.

B.3. Context-Expansion Denoising Steps

To assess the contribution of the context-expansion attention mask M_{expn}, we conduct an ablation study varying the number of denoising steps allocated to the text-focus phase (T_{focus}) and the context-expansion phase (T_{expn}), keeping the total number of $T_{\text{focus}}+T_{\text{expn}}$ steps fixed (*i.e.* gradually substituting M_{focus} with M_{expn}). Fig. S7 and Tab. S6 show qualitative and quantitative results under different allocations of these steps. When $T_{\text{expn}} = 0$ (leftmost column), the target text is accurately aligned within the designated region, but the lack of attention to surrounding context results in sharp and unnatural boundaries between the region and background. As more steps are allocated to context-expansion, this boundary effect is gradually alleviated, leading to more visually natural results. However, when T_{expn} becomes too dominant, information within the region starts to leak outward, leading to text generation that is no longer confined to the intended region (similar to the failure cases shown in the leftmost examples of Fig. S6). These results highlight the importance of a balanced sequential application of T_{focus} and T_{expn}—where T_{focus} helps localize the text within the target region, and T_{expn} promotes natural integration into the full image. Tab. S6 further supports this finding: both overly strong text-focus attention (first row) and excessive context-expansion (last row) lead to performance drops, while a balanced allocation yields the balanced high performance across all metrics.

n	Method	Acc.	NED	Qual.	Aesth.
5	AnyText	0.065	0.275	4.160	2.647
	GlyphControl	0.490	0.722	3.679	2.369
	TextDiffuser2	0.028	0.241	3.847	2.565
	EasyText	0.405	0.744	4.049	2.502
	DCText (Ours)	**0.693**	**0.860**	**4.697**	**3.569**
4	AnyText	0.052	0.269	4.324	2.815
	GlyphControl	0.507	0.729	3.867	2.512
	TextDiffuser2	0.081	0.321	3.869	2.493
	EasyText	0.454	0.759	4.235	2.717
	DCText (Ours)	**0.760**	**0.892**	**4.745**	**3.659**
3	AnyText	0.054	0.261	4.281	2.781
	GlyphControl	0.610	0.795	3.853	2.521
	TextDiffuser2	0.252	0.508	3.785	2.439
	EasyText	0.433	0.762	4.266	2.747
	DCText (Ours)	**0.768**	**0.906**	**4.735**	**3.709**
2	AnyText	0.052	0.275	4.386	2.857
	GlyphControl	0.692	0.862	4.009	2.624
	TextDiffuser2	0.528	0.729	3.755	2.438
	EasyText	0.460	0.791	4.363	2.870
	DCText (Ours)	**0.792**	**0.923**	**4.791**	**3.726**

Table S3. **Quantitative comparison between training-based baselines.** Comparison results on the CVTG-Style dataset across different numbers of sentences (n).

B.4. Localized Noise Initialization Steps

As shown in Fig. S8 and Tab. S7, increasing T_{init} improves text alignment within textual regions and enhances text accuracy. However, since this approach creates an initial latent with uneven noise levels between region and background, we observe that setting $T_{\text{init}} > 2$ leads to image collapse under our experimental setup with 24 denoising steps.

C. Broader Applications of DCText

C.1. General Object

While DCText is primarily designed to address the challenging task of rendering long or multiple texts, its core strategy generalizes effectively to broader visual generation tasks. To evaluate this generality, we apply DCText to the GenEval [7] benchmark, which focuses on the compo-

Mask	Acc.	NED	CLIP	Qual.	Aesth.
w/o $M_{r^c \to \{r_i\}}$	0.275	0.683	0.347	4.721	3.878
w/o $M_{\{p_i\} \to r^c}$	0.266	0.666	0.345	4.735	3.885
w/o $M_{p_g \to \{r_i\}}$	0.330	0.721	**0.350**	4.733	3.890
w/o $M_{\{p_i\} \to p_g}$	0.347	0.732	0.349	**4.746**	**3.921**
$\boldsymbol{M_{\text{focus}}}$	**0.387**	**0.751**	0.349	4.737	3.904

Table S4. **Ablation study for M_{focus} design.** Each row reports the result when one of the partial masks (defined in Sec. 3.2) is removed, evaluated on the single-sentence datasets.

T_{focus}	Acc.	NED	CLIP	Qual.	Aesth.
0	0.273	0.626	0.344	4.739	3.888
1	0.316	0.701	0.348	**4.740**	**3.905**
2	**0.387**	0.751	0.349	4.737	3.904
3	0.372	**0.757**	**0.350**	4.728	3.894
4	0.337	0.746	0.349	4.717	3.875

Table S5. **Ablation study for T_{focus} steps.** Quantitative results for different values of T_{focus}, evaluated on the single-sentence datasets.

sitional generation of general objects. In this experiment, we follow the original DCText pipeline as-is, but modify the GPT-4o [9] instructions for constructing textual prompts and regions. For textual prompts, we extract the target object from the global prompt and generate an object-centric prompt that includes a description aligned with the global context. For textual regions, we revise the original text-based instructions into object-based ones. We use the same denoising schedule as in the text generation setup: $(T_{\text{init}}, T_{\text{focus}}, T_{\text{expn}}) = (1, 2, 2)$ for single-object generation and $(2, 3, 2)$ for multi-object generation.

Fig. S9 shows qualitative results, and Tab. S8 summarizes quantitative comparisons. For evaluation, we generate four samples per prompt across all 553 prompts in the benchmark. As shown, DCText significantly outperforms the base model Flux, improving the overall GenEval score from 0.66 to 0.78. These results demonstrate the strong generalization capability of DCText beyond text rendering.

C.2. Stable Diffusion 3.5

We further evaluate the performance of DCText on another Multi-Modal Diffusion Transformer model, Stable Diffusion 3.5 Large (SD3.5-L) [6]. Following the same experimental setup as in the main paper, we also compare DCText against the same three baselines: SD3.5-L (the base model), AMO Sampler [8], and TextCrafter [5]. For fair comparison, we use a fixed number of 28 denoising steps across all methods, while keeping all other configurations at their respective defaults. As the AMO Sampler does not provide an official implementation for SD3.5, we re-implement it ourselves.

$T_{\text{expn}}(T_{\text{focus}})$	Acc.	NED	CLIP	Qual.	Aesth.
0 (4)	0.289	0.681	0.347	4.716	3.829
1 (3)	0.339	0.723	0.349	**4.739**	3.895
2 (2)	**0.387**	**0.751**	0.349	4.737	3.904
3 (1)	0.340	0.750	**0.350**	4.731	**3.905**
4 (0)	0.336	0.746	0.349	4.725	3.889

Table S6. **Ablation study for T_{expn} steps.** Quantitative results under different allocations of T_{expn} and T_{focus} (values in parentheses). As T_{expn} increases, T_{focus} is reduced accordingly, evaluated on the single-sentence datasets.

T_{init}	Acc.	NED	CLIP	Qual.	Aesth.
0	0.600	0.798	0.347	4.640	3.538
1	0.717	0.878	0.348	4.702	3.653
2	**0.753**	**0.895**	**0.349**	**4.742**	**3.666**

Table S7. **Ablation study for T_{init} steps.** Quantitative results for different values of T_{init}, evaluated on the multi-sentence dataset.

As shown in Fig. S10, DCText consistently produces accurate and coherent text aligned with the overall image context. In contrast, SD3.5-L and AMO Sampler often fail to render any text or generate inaccurate content, while TextCrafter tends to produce duplicated text and unnatural region boundaries. Tab. S9 presents the quantitative results on three single-sentence datasets, where all baselines generate three samples per prompt using the same random seed. Consistent with the Flux-based results in the main paper, DCText achieves the best performance across most metrics, including text accuracy and image quality.

D. Limitation

Our method relies on Flux's reliable short-text generation capability. If the textual prompt fails to generate the target text from the noise corresponding to the textual region, our method may not render the text correctly in that area. In addition, for our method to operate effectively, glyph-level features are expected to emerge before the Text-Focus denoising phase. This is because, after Text-Focus denoising, attention expands to the background region, followed by global denoising with full attention. While Flux typically forms coarse glyph structures during the early denoising steps, it occasionally fails to produce recognizable glyph features during this phase.

Fig. S11a illustrates such a case. The left images show intermediate results obtained by independently denoising each region for T_{focus} steps (with T_{init} set to 0 for simplicity). In the image for p_1, features resembling the word *sale* begin to emerge, whereas in the image for p_2, the model generates features related to the object light rather than the text *light*. In such cases, our method often fails to render the

Model	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding
Diffusion Models							
LDM [17]	0.37	0.92	0.29	0.23	0.70	0.02	0.05
SD1.5 [17]	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD2.1 [17]	0.50	0.98	0.51	0.44	0.85	0.07	0.17
SD-XL [15]	0.55	0.98	0.74	0.39	0.85	0.15	0.23
DALLE-2 [16]	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALLE-3 [1]	0.67	0.96	0.87	0.47	0.83	0.43	0.45
Flow Matching Models							
FLUX.1 Dev [10]	0.66	0.98	0.81	0.74	0.79	0.22	0.45
SD3.5-M [6]	0.63	0.98	0.78	0.50	0.81	0.24	0.52
SD3.5-L [6]	0.71	0.98	0.89	0.73	0.83	0.34	0.47
SANA-1.5 4.8B [19]	0.81	0.99	0.93	0.86	0.84	0.59	0.65
DCText (Ours)	0.78	1.00	0.90	0.51	0.84	0.82	0.61

Table S8. **Quantitative comparison on the GenEval benchmark.** We highlight the best scores in blue and second-best in green. Results for all baseline models are adopted from Flow-GRPO [11]. Obj.: Object; Attr.: Attribution.

Method	Acc.	NED	CLIP	Qual.	Aesth.
SD3.5-L	0.264	0.654	0.362	4.448	3.643
AMO-SD3.5	0.351	0.685	0.360	4.496	3.715
TextCrafter-SD3.5	0.241	0.707	**0.366**	4.326	3.507
DCText-SD3.5 (ours)	**0.359**	**0.742**	0.360	**4.618**	**3.728**

Table S9. **Quantitative comparison between SD3.5-based baselines.** Results are averaged over three single-sentence datasets. Since the official implementation of AMO-SD3.5 is not available, we implemented it ourselves.

target text in the corresponding region, as shown in the final output on the right.

However, such failures are often compensated for during global denoising. As in Fig. S11a, Fig. S11b also shows that no glyph-like features appear in the image for p_2, leading to missing text in that region of the final image. Nevertheless, since the global prompt includes the phrase corresponding to p_2, the final image still successfully generates the text *Meeting Room*.

E. Experimental Details

E.1. Implementation

For the ChineseDrawText [14], DrawTextCreative [12], and TMDBEval500 [3] datasets, we generate both textual prompts and textual regions using GPT-4o [9]. Textual prompts are constructed following the instruction in Tab. S10. For each sentence contained in the prompt, we produce a description and format the result as: *'Rendering word: "{sentence}"\n Description: {description}'*. Textual regions are constructed according to the bounding box generation instructions outlined in Tab. S11. In the Localized Noise Initialization process, we set the weighting factor $\alpha = 0.7$. During denoising, we use a guidance scale of 5.0. Ours attention masks are applied to all MM-DiT blocks, including both double- and single-stream variants. For pooled textual embeddings, we average the embeddings obtained from all textual prompts, including the global prompt. The text accuracy for multiple sentences is evaluated using GPT-based recognition, following the instructions provided in Tab. S12.

E.2. Human Evaluation

We conduct our human evaluation using a pairwise comparison (A/B test) protocol. In each test, participants are shown two images: one from our proposed model, DCText, and one from a randomly selected baseline (Flux [10], AMO Sampler [8], or TextCrafter [5]). To mitigate bias, the display order of the images is randomized. The participants are then asked to choose the superior image based on the following three criteria:

- **Text Accuracy:** *Which image renders the text more accurately (i.e., correct spelling, legibility, and completeness of the intended words)?*
- **Prompt Alignment:** *Which image better reflects the content and intent of the given prompt, including both the visual elements and the embedded text?*
- **Image Quality:** *Which image has higher overall quality in terms of visual naturalness, aesthetic appeal, and artis-*

tic style?

The evaluation interface is illustrated in Fig. S12.

To assess the significance of user preferences, we perform one-sided binomial tests for each pairwise comparison, excluding ties. DCText shows statistically significant improvements in text accuracy over all baselines ($p < 0.0001$), and in prompt alignment over AMO Sampler and TextCrafter ($p < 0.001$). For overall image quality, the improvement over TextCrafter is also significant ($p = 0.002$), while those over AMO and FLUX do not reach significance.

E.3. Abbreviated Prompts

Due to space constraints in Fig. 1 and 4 of the main paper, we present only abbreviated examples of the global prompts. The complete set of prompts is provided in Tab. S13, where the target rendering text is highlighted.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 5

[2] Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024. 1, 11

[3] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023. 1, 2, 5

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer, 2024. 1

[5] Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025. 2, 4, 5

[6] 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. In *Forty-first International Conference on Machine Learning*, 2024. 4, 5

[7] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 3

[8] Xixi Hu, Keyang Xu, Bo Liu, Qiang Liu, and Hongliang Fei. Amo sampler: Enhancing text rendering with overshooting, 2025. 4, 5

[9] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4, 5

[10] Black Forest Labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. 5

[11] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 5

[12] 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. *arXiv preprint arXiv:2212.10562*, 2022. 2, 5

[13] Runnan Lu, Yuxuan Zhang, Jiaming Liu, Haofan Wang, and Yiren Song. Easytext: Controllable diffusion transformer for multilingual text rendering. *arXiv preprint arXiv:2505.24417*, 2025. 1

[14] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870*, 2023. 2, 5

[15] 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. 5

[16] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 5

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[18] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023. 1

[19] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025. 5

[20] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36:44050–44066, 2023. 1

You are given a text-to-image generation prompt that includes quoted text.

Your task is to extract each quoted sentence and generate a visual style description for the text inside the quotation marks.

- Describe how the text visually appears in the image, including font style, color, texture, effects, etc.
- For each extracted sentence, write a concise and context-aware visual style description.
- Do not describe the sentence's position or relative order.
- Do not mention any rendering words. Avoid using quotation marks or referring to specific text.

Example:

```
[
  {
    "sentence": "diamonds",
    "description": "A sleek, modern sans-serif font in metallic silver."
  }
]
```

Table S10. GPT-4o instruction for generating sentence descriptions within textual prompts.

You are given a text-to-image prompt with quoted text.
Your task is to extract the quoted text and generate a bounding box.

Step-by-step Instructions

1. Quoted Text Isolation

- Extract the text inside quotation marks only.
- Example:
  Prompt: A sign that says `"Do not reserve a seat"` → Use: `'Do not reserve a seat'`

2. Bounding Box Layout Rules

- The bounding box must be placed in regions where the text is likely to appear, as implied by the prompt.
- Bounding boxes must not overlap.

3. Bounding Box Calculation

- Output each bounding box as normalized coordinates, meaning all values (x1, y1, x2, y2) are between 0 and 1, representing a fraction of the image width and height.
- Consider the number of characters, including spaces and punctuation, for the size of the box.
- The height of every bounding box must be {`height`}.
- The width of every bounding box must be at least {`min_width`}.
- Final format: `[x1, y1, x2, y2]`.

Table S11. GPT-4o instruction for generating bounding boxes of textual regions.

Recognize all textual elements in the image as they would be perceived by a human and organize them into accurate, sentence-level units.

- Split the text based on meaningful sentence boundaries.
- Each sentence must come from a single region in the image.
- Do not correct or modify awkward words or phrases.
- Include a score indicating the visual recognition confidence of each sentence.

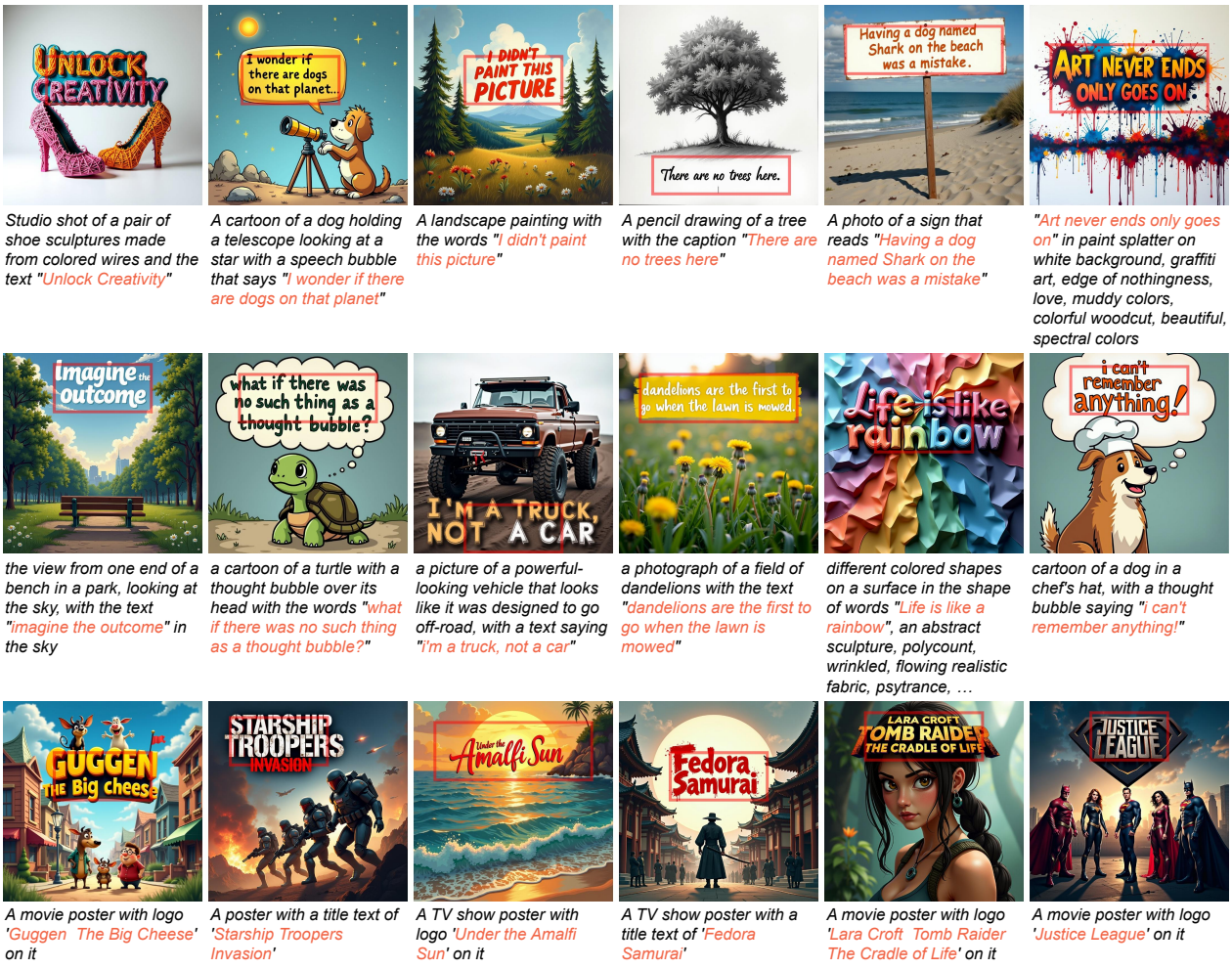
Example Output Format (JSON):

```
[
   {"sentence": "New Specials Every Week", "score": 0.96},
   {"sentence": "We are OPEN EVERY DAY", "score": 0.91}
]
```

Table S12. GPT-4o instruction for text recognition in accuracy evaluation.

Figure	Prompt
Figure 1	• A sprawling financial district at dusk, where the text "DCText: Scheduled Attention Masking for Visual Text Generation via Divide-and-Conquer Strategy" is projected across the mirrored glass surface of a skyscraper. The characters are bold, futuristic sans-serif with a subtle neon blue glow, appearing as if etched into the building façade, reflecting surrounding city lights faintly. • A quaint street corner during dusk, with a classic 1950s-style diner sign. The text "DCText: Scheduled Attention Masking for Visual Text Generation via Divide-and-Conquer Strategy" is displayed in a retro script font with glowing red and cream-colored bulbs along the letters, evoking a warm, nostalgic roadside ambiance. • A vintage-style parchment sheet with burned edges, where the text "DCText: Scheduled Attention Masking for Visual Text Generation via Divide-and-Conquer Strategy" is hand-painted in an ornate calligraphic style using dark ink with faint gold leaf accents. The imperfect, organic strokes give the title an ancient manuscript appearance. • A futuristic night sky above a modern metropolis, where hundreds of synchronized drones form the text "DCText: Scheduled Attention Masking for Visual Text Generation via Divide-and-Conquer Strategy" in the air. Each character is composed of tiny glowing blue lights, forming a perfectly sharp sans-serif display that glimmers softly against the starry sky, while the city below remains dim and distant.
Figure 4	• On a sunny beach scene, a lifeguard tower sign proclaiming 'Swimming Area Open' in large red letters, a beach umbrella with 'Relax and Enjoy the Sun' in colorful cursive, a kiosk displaying 'Beach Gear Rentals' in medium blue, and a signpost pointing to 'Surf Lessons Starting at 10 AM' in bold orange. • Charming restaurant exterior showcasing a sign with 'Family Meals Available' in large red letters, a window display reading 'Daily Fresh Catch' in bold blue letters, a door plaque labeled 'Welcome Diners' in green cursive, a patio banner saying 'Outdoor Seating Open' in italic large letters, and a menu board displaying 'Chef Specials Tonight' in medium bold letters.

Table S13. The full text-to-image prompts for Figure 1 and 4, ordered left to right (Figure 1) and top to bottom (Figure 4).



Studio shot of a pair of shoe sculptures made from colored wires and the text "Unlock Creativity"

A cartoon of a dog holding a telescope looking at a star with a speech bubble that says "I wonder if there are dogs on that planet"

A landscape painting with the words "I didn't paint this picture"

A pencil drawing of a tree with the caption "There are no trees here"

A photo of a sign that reads "Having a dog named Shark on the beach was a mistake"

"Art never ends only goes on" in paint splatter on white background, graffiti art, edge of nothingness, love, muddy colors, colorful woodcut, beautiful, spectral colors

the view from one end of a bench in a park, looking at the sky, with the text "imagine the outcome" in the sky

a cartoon of a turtle with a thought bubble over its head with the words "what if there was no such thing as a thought bubble?"

a picture of a powerful-looking vehicle that looks like it was designed to go off-road, with a text saying "i'm a truck, not a car"

a photograph of a field of dandelions with the text "dandelions are the first to go when the lawn is mowed"

different colored shapes on a surface in the shape of words "Life is like a rainbow", an abstract sculpture, polycount, wrinkled, flowing realistic fabric, psytrance, …

cartoon of a dog in a chef's hat, with a thought bubble saying "i can't remember anything!"

A movie poster with logo 'Guggen The Big Cheese' on it

A poster with a title text of 'Starship Troopers Invasion'

A TV show poster with logo 'Under the Amalfi Sun' on it

A TV show poster with a title text of 'Fedora Samurai'

A movie poster with logo 'Lara Croft Tomb Raider The Cradle of Life' on it

A movie poster with logo 'Justice League' on it

Figure S1. **Qualitative samples on single sentence.** Prompts, including the sentence to be rendered (highlighted in red), are shown below each image. Corresponding textual regions are indicated with red boxes.

In a train car, a monitor shows 'Next Departure' in medium regular, a poster on the wall says 'Buy Tickets' in large red letters.

On a university campus, a building banner announces 'Welcome New Students!' in large bold letters and a noticeboard shows 'Lecture Schedule for Spring 2025' in neat black font.

A bustling marketplace with a neon sign that reads 'Fresh Produce Daily' in bright green cursive, and a sandwich board outside a shop that says 'Healthy Meals Ready in 10 Minutes' in large letters.

On a library wall, a poster reads 'Read, Explore, Discover' in elegant script, and a nearby table has a sign with 'Quiet Zone, Please Respect Silence' in small blue letters.

A coffee shop window displays 'Brewing Happiness Since 1998' in elegant gold cursive, and a chalkboard outside advertises 'Special Latte of the Day: Cinnamon Bliss' in bold letters.

At a subway station, a sign reads 'Next Train' in large yellow letters, a nearby vending machine displays 'Snack Time' in bold green.

At a beach resort, a flag waves with 'Relax' in large blue cursive letters, a beach chair has 'Sun' written on it in medium yellow, and a drink menu board shows 'Tropical' in bright orange bold letters.

In a café, the menu board says 'Coffee' in large brown cursive, a napkin on the table has 'Enjoy' in small green, and a wall poster shows 'Special' in bold orange.

A shopping mall atrium with an overhead banner reading 'Seasonal Clearance Event' in large red bold, a floor sign showing 'Sale Ends This Weekend' in medium blue, and an interactive screen advertising 'Click for Exclusive Deals' in bright green cursive.

On a park bench, a flyer advertises 'Join Us' in large blue letters, a nearby tree has a plaque reading 'Established 1999' in small italic, and a bench sticker says 'Rest Here' in medium green.

A coffee shop interior with a chalkboard menu saying 'Hot Coffee' in large cursive, a table with 'Special Offer' in small bold on a napkin, and a wall art with 'Stay Cozy' written in medium regular.

In a gym, a motivational poster reads 'Push Harder' in large italic, a digital clock shows '7:15 AM' in bold, and a water dispenser says 'Hydrate Now' in medium blue.

A coffee shop counter with a chalkboard sign that says 'Brew' in large brown cursive, a napkin with 'Sip' written on it in small black regular letters, a table with 'Relax' engraved on the edge in medium green italic, and a coffee cup featuring 'Warm' in bold red.

In a public park, the bench sign reads 'Rest Area' in medium green bold letters, the playground entrance displays 'Kids Zone' in large colorful, the picnic table banner says 'Family Time' in small blue italic, and the trail sign shows '5 Miles' in medium brown.

A coffee shop with a sign that says 'Fresh Brew' in large brown cursive letters, a menu board displays 'Daily Specials' in medium black, a window sticker reads 'Open 24/7' in bold red, and a table placard shows 'Free WiFi' in italic blue.

A lively beach scene with a surfboard saying 'Ride' in bright blue large letters, a beach towel with 'Relax' on it in colorful cursive, a sandcastle labeled 'Fun' in bold yellow, and a cooler with 'Chill' written in italic white.

A downtown street scene, a neon sign flashing '24/7 Convenience Store' in large neon green bold, a parking lot banner saying 'Free Parking for 2 Hours' in medium yellow regular, a subway station sign showing 'Next Train in 5 Minutes' in small red italic, and a bus stop with 'Route 42 - Downtown to Uptown' in large blue regular.

A vintage diner scene with a jukebox featuring 'Hits' in large bright red letters, a menu board showing 'Burgers' in medium yellow bold, a neon sign with 'Shake' in italic pink, and a napkin dispenser with 'Fun' in small blue cursive.

A university campus with a board saying 'Library' in large blue letters, a student union building with 'Cafe' in medium brown, a banner on the quad reading 'Welcome' in large bold green, a bulletin board marked 'Events' in medium red, and a bench with 'Study' in small italic.

A street market scene, a sign above the stall says 'Fresh Produce' in large green letters, a banner on a nearby truck reads 'Discounted Prices' in bold red, a sign on a fruit crate displays 'Organic Apples' in small italic, a street vendor apron says 'Healthy Snacks' in large cursive, and a poster on a wall reads 'Sale Today' in bold blue.

In a café, a chalkboard menu says 'Hot Coffee' in large brown cursive, a napkin has 'Enjoy Your Day' in small green, a table number reads 'Table 4' in medium black bold, a sign by the counter says 'Order Here' in large red italic, and a coffee cup displays 'Fresh Brew' in blue regular.

A book cover with the title 'The Hidden Path' at the top in large elegant cursive, a subtitle saying 'A Journey Begins' in medium italic, a review box showing 'Top Seller' in bold gold, an author name 'John Doe' in medium regular, and a publisher logo 'Epic Reads' in small blue.

A city street scene with a billboard saying 'Visit the New Downtown Mall' in large bold blue letters, a taxi cab with 'Fare Starts at 5' in green, a restaurant window displaying 'Try Our Signature Dish Today' in medium italic, a traffic light with 'Stop for Safety' in red, and a newsstand sign reading 'Catch Up on Daily Headlines' in bold black.

At a lively restaurant, a menu board lists 'Burger Special' in large bold red letters, a table sign reads 'Reserved' in medium italic black letters, a chalkboard displays 'Happy Hour' in large green letters, a wall decoration shows 'Fresh Salad' in medium regular white letters, and a door sticker says 'Welcome' in large cursive blue letters.

Figure S2. **Qualitative samples on multiple sentences.** Prompts, including the sentences to be rendered (highlighted in red), are shown below each image. Corresponding textual regions are indicated with red boxes.

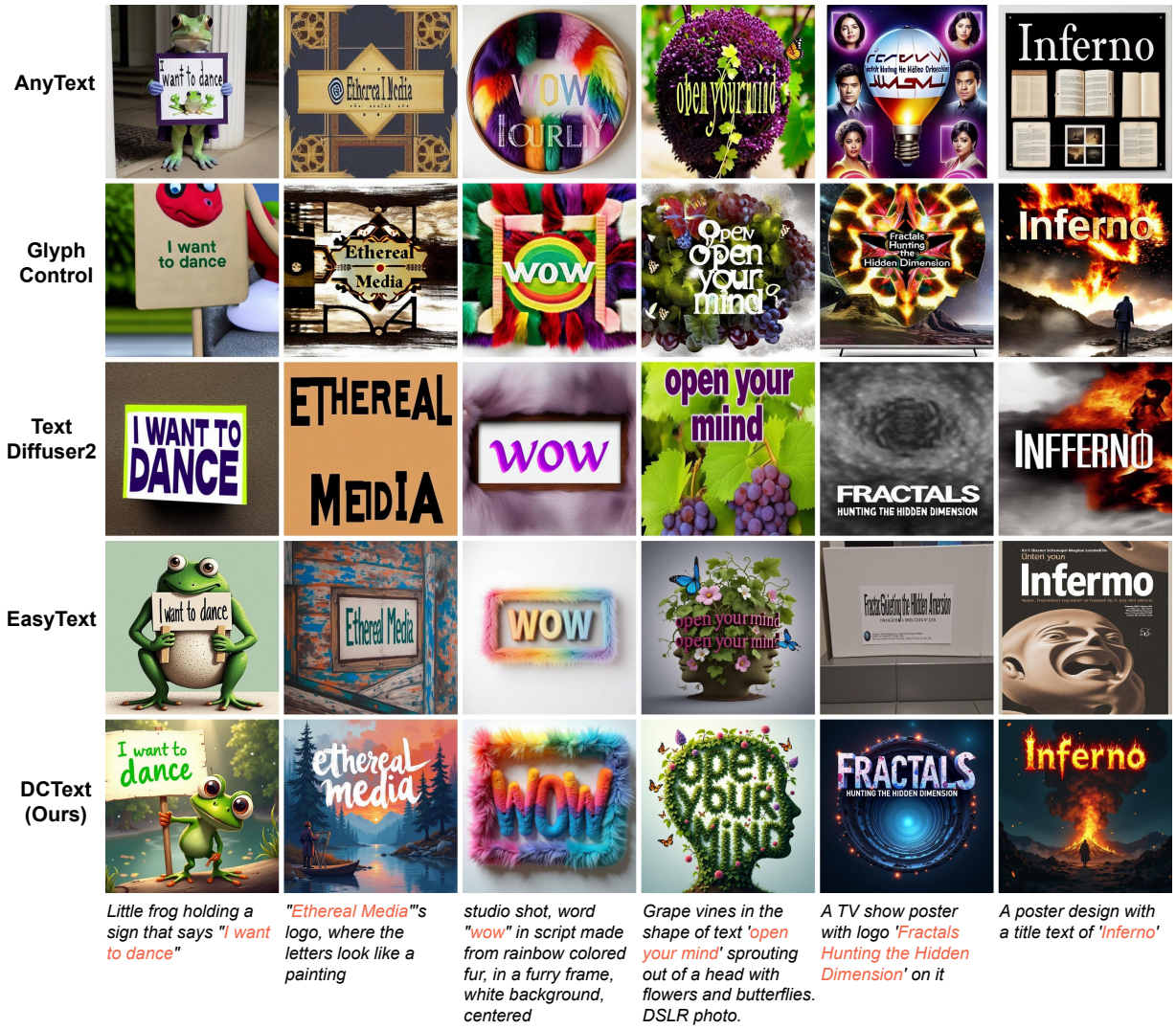


Figure S3. **Qualitative comparison between training-based baselines.**

Figure S4. **Comparison to Regional-Prompting [2]** Comparison of generation results with another attention control method that relies solely on the Region-Isolation Attention Mask (M_{isol}). For a fair comparison, we set $T_{\text{init}} = 0$.

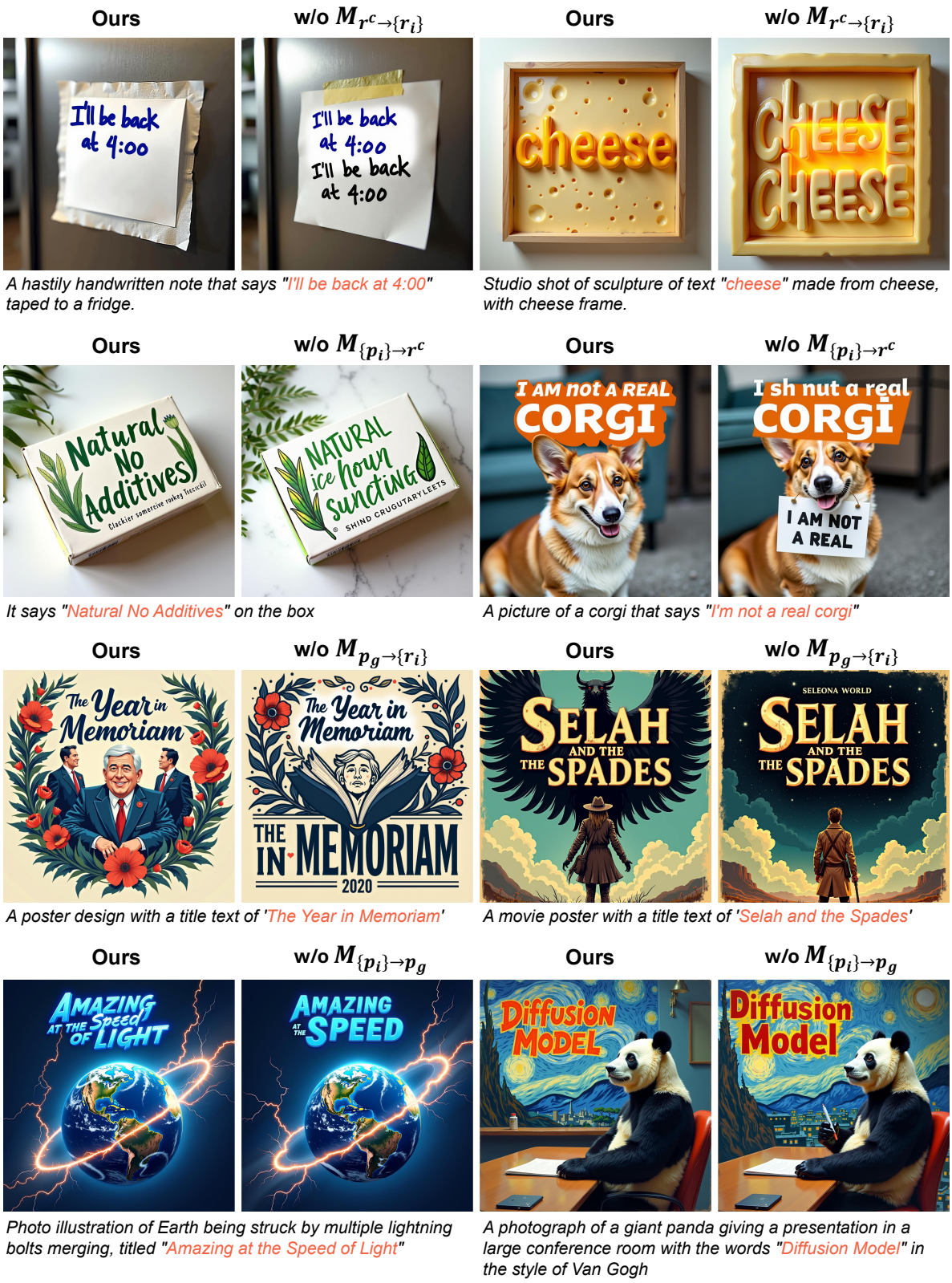


Figure S5. **Ablation study for the text-focus attention mask design.** In each pair, the right image shows the result without the corresponding partial mask, and the left image shows the result with it applied.

Figure S6. **Ablation study for T_{focus} steps.** Qualitative results for varying T_{focus}, with $T_{\text{init}} = 1$ and $T_{\text{expn}} = 2$ fixed.

Figure S7. **Ablation study for T_{expn} steps.** Qualitative results for varying T_{expn}, where T_{focus} is reduced accordingly under a fixed total number of steps, with $T_{\text{init}} = 1$ fixed.

A retro book cover showing a detective holding a magnifying glass with 'Crime Scene' in bold, a title at the top that says 'The Mystery' in large italic, and the author name at the bottom with 'Coming Soon' in small regular letters.

In a library, a label displays 'Mystery Novel' in large italic blue letters, a desk has a notebook with 'Chapter 1' written on it in small regular font, and a shelf has a book titled 'Secrets Unfold' in medium cursive.

Figure S8. **Ablation study for T_{init} steps.** Qualitative results for varying T_{init}, with $T_{\text{focus}} = 3$ and $T_{\text{expn}} = 2$ fixed.

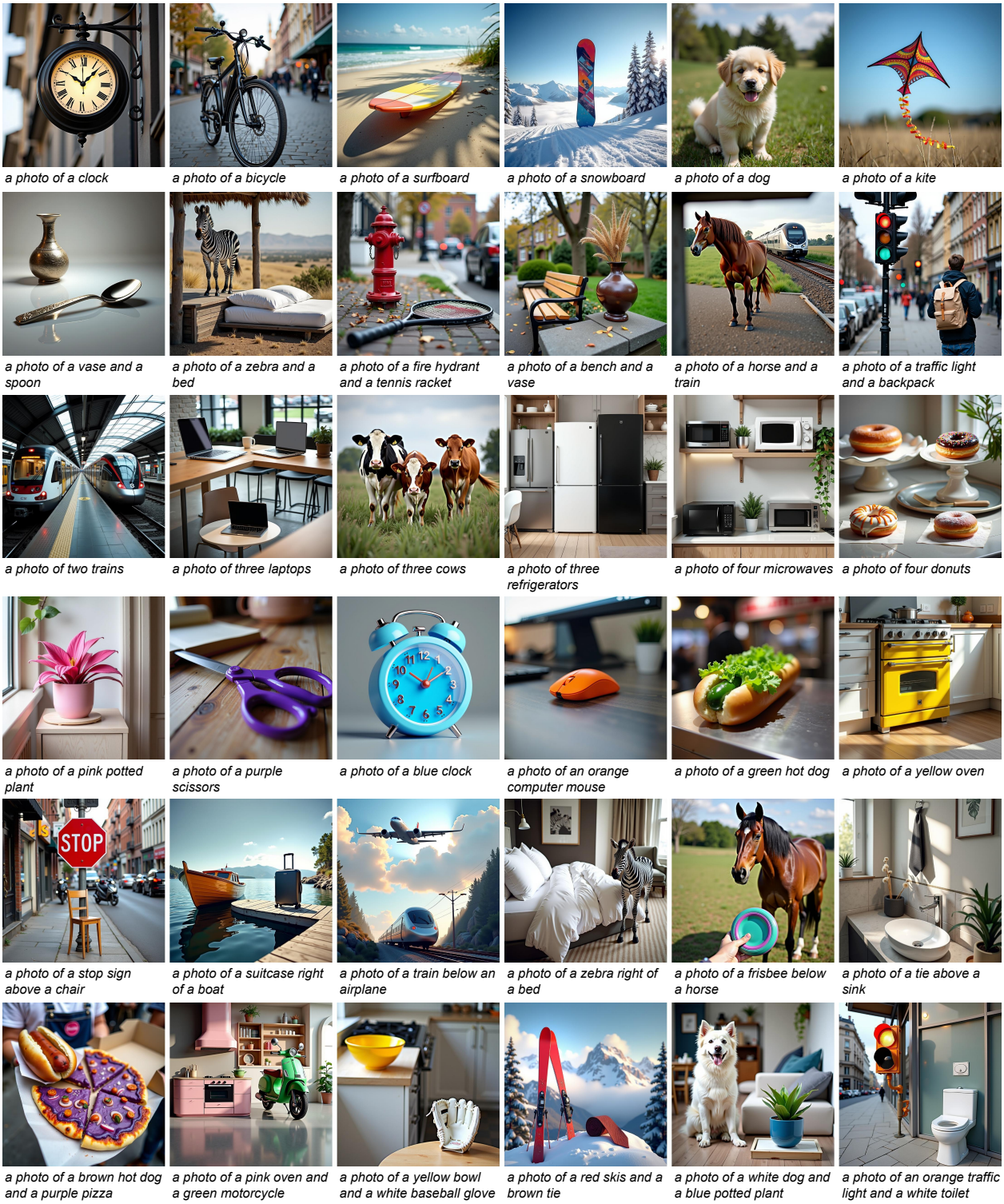


Figure S9. **Qualitative samples on the GenEval benchmark.** DCText-generated samples on the GenEval benchmark. Rows correspond to Single Object, Two Object, Counting, Colors, Position, and Attribution Binding tasks.

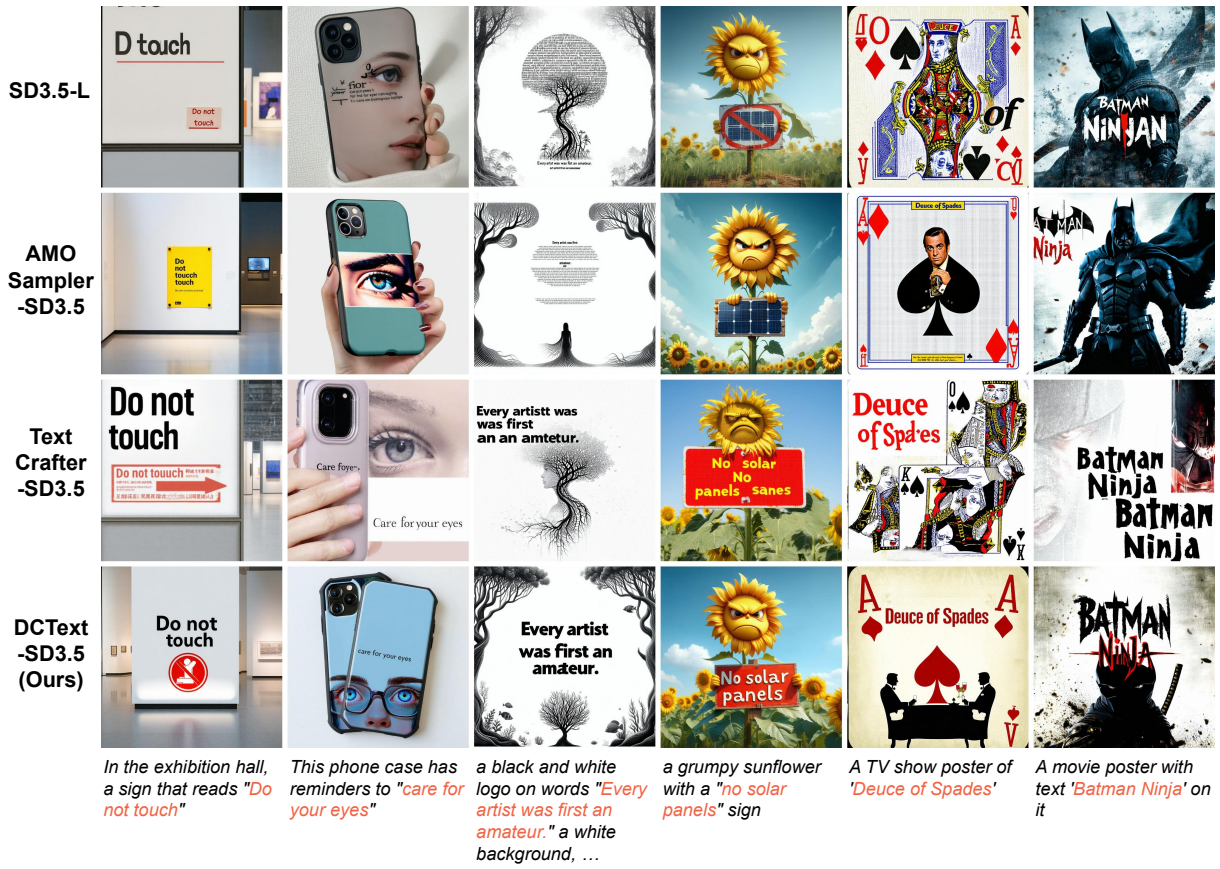


Figure S10. **Qualitative comparison between SD3.5-based baselines.** Samples generated by each method using the SD3.5-L.

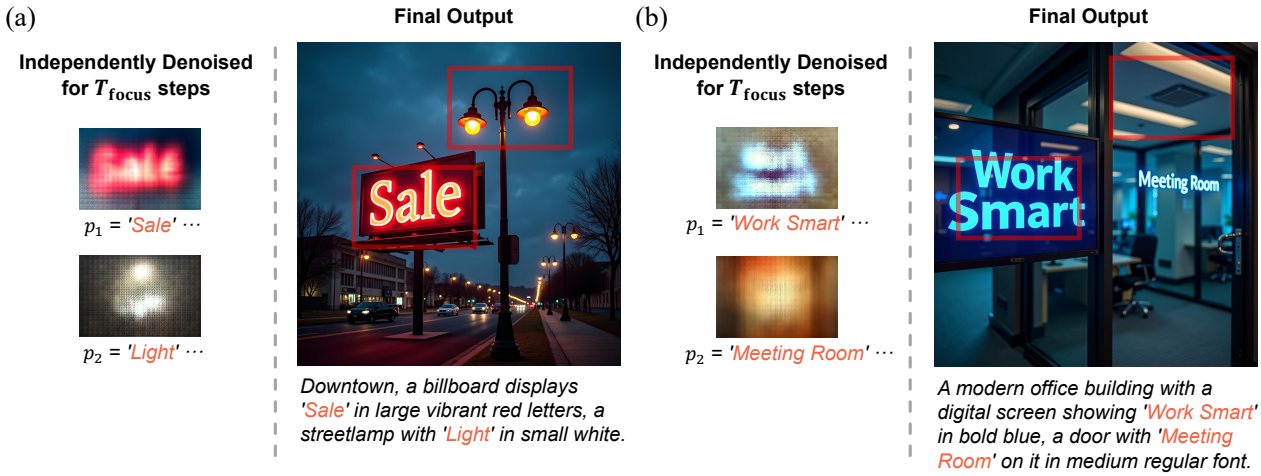


Figure S11. **Failure Cases.** Each region is extracted from the initial noise used to generate the final image (right) and denoised for T_{focus} steps using the corresponding textual prompts (left). (a) The prompt p_1 leads to clear glyph-like features, but not p_2. As a result, only *Sale* appears in the final image. (b) Similar case where the region for p_2 fails to form glyphs early on. Nevertheless, the global prompt allows *Meeting Room* to appear during global denoising.

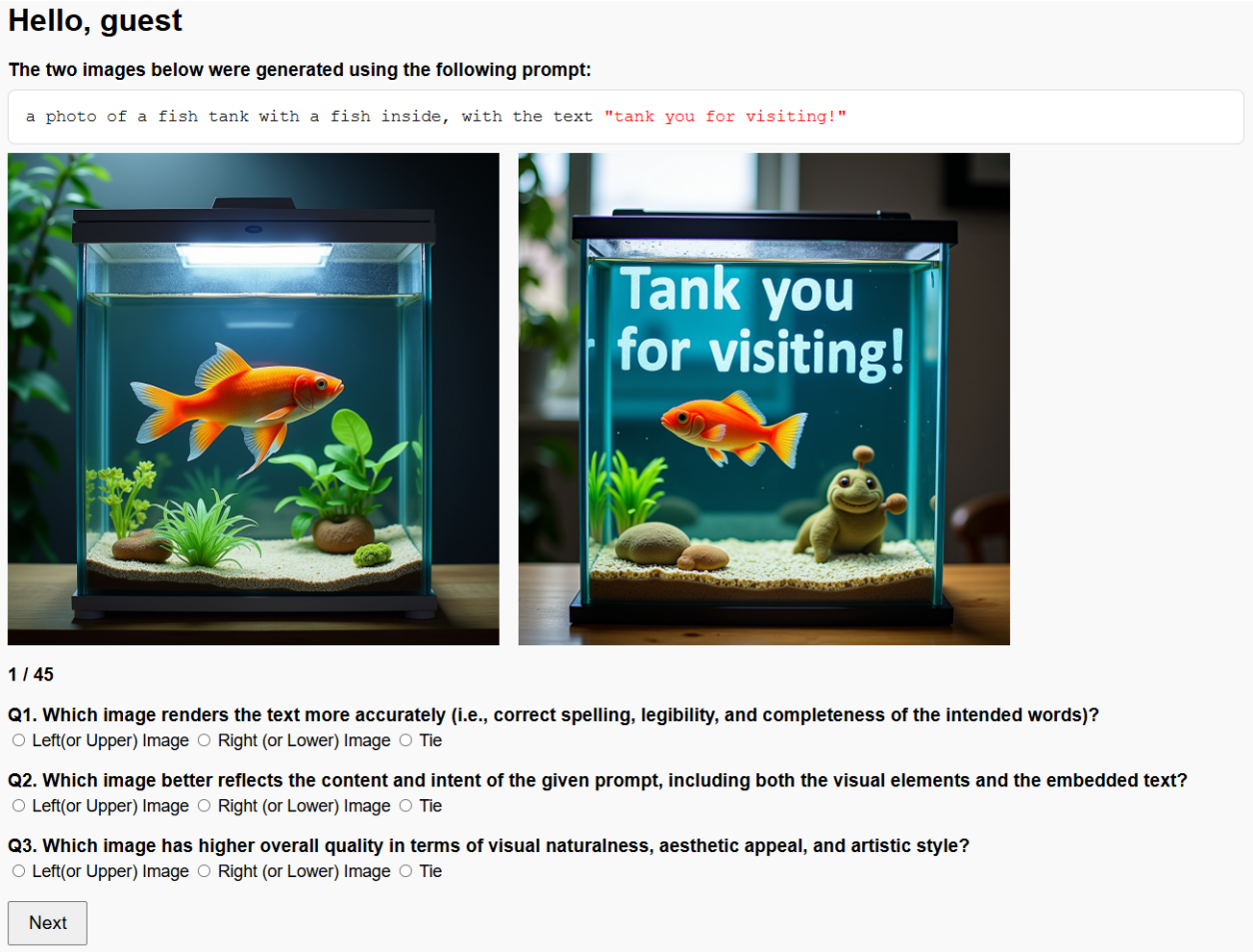


Figure S12. **Human evaluation interface.** For each prompt, evaluators perform a pairwise comparison of two generated images, assessing them on text accuracy, prompt alignment, and image quality.