Text-Printed Image: Bridging the Image-Text Modality Gap for Text-centric Training of Large Vision-Language Models

Supplementary Material

A. Experimental Details

A.1. Compute Resources

All experiments were conducted on an HPC cluster. For most experiments, we use a single NVIDIA H100 GPU. For generating synthetic images with T2I, we use a node equipped with 8× NVIDIA H100 GPUs. The system uses NVIDIA driver 535.183.01 and CUDA 12.2.

A.2. Training Details

Training data. For all tasks, we train models using the provided training split of each dataset. Input queries are formatted with the chat templates associated with each LVLM. To ensure fair comparison under `lmms-eval` [65], we format the model responses to match the output format expected by the `lmms-eval` evaluation pipeline. For DriveLM [47], although the original setting uses six images per example, we simplify training by using only the single image corresponding to each QA pair.

Hyperparameters. We use the same training configuration for all models. We optimize with AdamW, set the connector learning rate to 1×10^{-5}, and apply a cosine learning-rate schedule with a warm-up ratio of 0.03, no weight decay, a global batch size of 4, and gradient accumulation of 4. We train for 3 epochs in `bfloat16` precision. All models are fine-tuned with LoRA with rank $r = 256$ and scaling factor $\alpha = 512$. Following prior work [71], we update only the parameters of the LLM, while keeping the visual encoder frozen.

A.3. Details of Generating Textual Descriptions

We automatically generate textual descriptions from ground-truth images using an LVLM. Specifically, we employ Qwen2.5-VL-32B [54], a well-trained LVLM. For each sample, we provide the ground-truth image and its corresponding QA pair as input to the model and instruct it to produce a textual description that is relevant to the given QA. The exact prompt used for this generation is shown below:

```
System: You are a highly skilled visual description assistant. Given an image and,
optionally, a question and its answer, write a single-paragraph, highly detailed,
objective description of the image. Your goal is to capture all relevant visual
elements in a way that would allow a reader to mentally reconstruct the image and, if
applicable, answer the given question using only your description. Your description
should include the type of scene (e.g., natural, diagram, poster, chart), the spatial
layout of elements, visual attributes such as color, shape, and texture, any visible
text or labels, and, if present, numerical or symbolic information. Do not include any
interpretation, emotion, speculation, or bullet points. Keep the tone factual, precise,
and comprehensive. Aim for approximately 100 words.

User:
Question: {question}
Answer: {answer}
```

A.4. Relevance Score

To quantify how faithfully a synthetic image matches its paired textual supervision, we compute *Relevance Scores* by using Qwen2.5-VL-32B-Instruct. Inference is performed with temperature = 0 and max_tokens = 1 to obtain deterministic Yes or No predictions. Images are attached through the official Qwen chat template. Specifically, we also use the following prompt:

```
System: Output only Yes or No.
User: Is the image relevant to the following Q&A?
```

Setting	ScienceQA	OK-VQA
Orig. (1%)	6,965	6,194
Orig. (10%)	6,513	3,927
Orig. (100%)	7,118	1,317

Table 8. Number of generated augmented examples for each dataset and training fraction.

```
Question: {q}
Answer: {a}
```

For each prompt, we compute the probability of generating "yes" and "no" from the first tokens.

A.5. t-SNE visualization

For t-SNE visualizations, we first pass inputs constructed by each method (GT-Image, TPI, and Text-only) through the model and extract the hidden states from all transformer layers. For each input, we take the hidden vector at the last token position as the layer-wise representation and apply t-SNE to obtain a 2D embedding. For hyperparameters, we set perplexity = 100.0, learning rate = 1000.0, and number of iterations = 5000. Specifically, Figure 4 reports the t-SNE embeddings from layer 11 for LLaVA-7B, layer 10 for LLaVA-13B, and layer 20 for both Qwen2.5-VL and LLaMA 3.2 Vision.

B. Details of TPI Generation Settings

In this section, we describe the details of generating synthetic images for TPI. For a fair comparison with image-based training, we first automatically generate textual descriptions from each ground-truth image as described in Section A.3. We note that only these textual descriptions are used for training; original images are not required in practical use. For each description, we render text using Python and the Pillow library [9].

Layout parameters. Each TPI is an RGB image of size 336×336 pixels. By default, we use a plain white background and black text, and render with a TrueType font (DejaVu Sans). For a given text, we perform a top–down search over font sizes: starting from a default font size (32 pt) and decreasing it in steps until the wrapped text fits within the target canvas. At each candidate size, we construct lines by greedy word-wrapping so that the width of each line does not exceed the available width (`img_width` minus horizontal padding). We then estimate the line height and total height of the wrapped text, including a fixed line spacing. If both the maximum line width and total text height fit within the canvas after respecting padding on all sides, we accept this font size and render the text.

C. Details of Data Augmentation

In this section, we describe the data augmentation setup used in Section 5.

Pipeline overview. We follow the Self-Instruct framework [56] to generate additional training data for our text-centric setting. First, for each seed example, we generate an image caption from the associated image and use these captioned examples as the initial pool. Then, at each iteration, we randomly sample 8 examples from the pool as demonstrations and ask the LLM to generate *one* new example. We check whether the generated example overlaps with the existing pool using ROUGE-L. If the example is not considered a duplicate, we add it to the pool and repeat the same procedure to generate the next example. We run this generation loop for 10,000 iterations for each task.

Table 8 summarizes the number of generated augmented examples. In particular, the number of generated examples for OK-VQA in the 100% setting is relatively small. We hypothesize that this is because the original OK-VQA dataset is already quite comprehensive, so there is limited room for the LLM to propose new, non-duplicate knowledge. This also explains why the performance gains from augmentation are modest in this setting.

Model and prompt. For data augmentation, we use `gpt-4o-mini` as the generation model. We set the temperature to 0.7 to encourage diversity in the generated examples. To strictly control the output format, we instruct the model to return a single JSON object that matches a predefined schema. The core prompt is as follows:

```
System:  Return exactly one JSON object that validates against the given schema.  No
extra text.
User:  Here are seed examples (one JSON per line):

{demo}

Produce ONE new and diverse example that is not copied.  Output only the JSON object.
```

There is still substantial room to tune this text-centric augmentation process. For example, one could add extra instructions that modify only specific parts of the image description (such as changing certain objects) to obtain more diverse captions and questions.

Duplicate detection. We use ROUGE-L to detect duplicate-like examples. For each newly generated example, we first convert it into a single canonical text string. We then compute the ROUGE-L F1 score between this string and the canonical text of every example in the pool. If the best ROUGE-L F1 score is greater than or equal to a threshold of 0.70, we treat the example as duplicate-like and discard it. This heuristic is conservative: even if many words overlap, changing a small number of key words (for example, replacing "cat" with "dog") can change the underlying question. Such semantic changes are not always fully captured by ROUGE-L alone, so more advanced duplicate detection methods could further improve this step.

Training setup. For SFT, we use exactly the same training configuration as in the main experiments. In all cases, we perform SFT with LoRA for 3 epochs on the (original + augmented) training data.

D. Ablation Study

In this section, we conduct an ablation study on how TPIs are generated.

D.1. Font Size

We evaluate how the font size of the text in TPI affects training. We consider four fixed font sizes: 4, 16, 32, and 64. In the main experiments in Section 4, we instead use a default font size of 32 and reduce it only when the text does not fit into the image, so these fixed sizes are not used there. All other training settings are exactly the same as in Section 4.

Results. The results are shown in Table 9. We observe that models train well with font sizes 16 and 32, while performance drops for sizes 4 and 64, especially on VizWiz. When the font size is too small, the model may fail to read the characters, particularly when the input text is long and dense. When the font size is too large, the text can overflow outside the image, causing part of the information to be lost and making the TPI image less informative. Based on these results, we recommend using moderate font sizes such as 16 or 32.

D.2. Font Color

We evaluate how the font color of the text in TPI affects model training. Font color may influence the model's ability to recognize characters. If some colors are easier for the model to read, using them could improve learning performance.

Setup. We consider six candidate font colors: *black*, *blue*, *green*, *orange*, *red*, and *yellow*. In the main experiments in Section 4, we use black as the default color. Apart from the font color, all training settings are exactly the same as in Section 4.

Results. The results are shown in Table 10. We do not observe large differences in performance across font colors. If anything, black yields the highest performance on average. A possible reason is that black text is also the most common in standard OCR datasets, so the model may find it easier to recognize. Based on these results, we recommend using black fonts when generating TPI. However, other colors are also acceptable, so in practice it may be preferable to choose a font color that matches the background while preserving sufficient contrast for readability.

Table 9. Comparison of font sizes. Each value shows accuracy (%) for different font sizes in TPI.

	Font Size			
Model	4	16	32	64
ScienceQA				
LLaVA 7B	73.57	**74.27**	73.87	73.76
LLaVA 13B	75.20	76.00	**76.60**	75.95
Qwen2.5 VL	89.69	**91.22**	90.68	88.00
LLaMA Vision	86.25	**90.33**	89.99	87.80
Avg.	81.18	**82.95**	82.78	81.38
OK-VQA				
LLaVA 7B	59.98	60.52	**60.54**	60.11
LLaVA 13B	63.96	**64.81**	64.53	63.86
Qwen2.5 VL	62.32	62.56	**63.08**	62.47
LLaMA Vision	61.54	**62.47**	62.27	62.37
Avg.	61.95	62.59	**62.60**	62.20
VizWiz				
LLaVA 7B	56.20	**62.59**	60.92	57.16
LLaVA 13B	57.87	**63.63**	62.41	57.16
Qwen2.5 VL	64.46	**68.50**	68.11	65.63
LLaMA Vision	63.39	**70.41**	68.85	63.76
Avg.	60.48	**66.28**	65.07	60.93

D.3. Generation Prompt

The amount of information contained in the given textual descriptions may affect the effectiveness of training. Therefore, we evaluate how the learning results change when we vary the prompts used to generate these descriptions. Specifically, we consider the following three types of prompts.

50 words. This prompt restricts the description to at most 50 words. Although the description contains less information, it may retain only the most important content. The exact prompt we use is shown below:

Prompt for 50 words.

```
You are an objective image captioning assistant.  Describe strictly what you see in the
image.  Do NOT include apologies, judgments, context, or filler phrases.  Use up to 50
words in your description.
```

200 words. This prompt restricts the description to at most 200 words, allowing a longer and more detailed explanation than the 50-word setting. The exact prompt we use is shown below:

Prompt for 200 words.

```
You are an objective image captioning assistant.  Describe strictly what you see in the
image.  Do NOT include apologies, judgments, context, or filler phrases.  Use up to 200
words in your description, providing detailed coverage of objects, setting, colors, and
actions.
```

Table 10. Comparison of font colors. Each value shows accuracy (%) for different font colors in TPI. We did not observe any substantial performance differences across font colors.

	Font Color					
Model	Black	Blue	Green	Orange	Red	Yellow
ScienceQA						
LLaVA 7B	75.11	74.81	74.91	74.37	**75.16**	75.01
LLaVA 13B	76.30	76.65	76.55	**76.95**	76.85	76.65
Qwen2.5 VL	90.43	91.47	91.77	**91.92**	91.18	91.32
LLaMA Vision	**90.93**	90.43	90.28	90.38	90.33	**90.93**
Avg.	83.19	83.34	83.38	83.41	83.38	**83.48**
OK-VQA						
LLaVA 7B	**61.70**	60.40	60.64	60.48	60.54	60.69
LLaVA 13B	64.73	**64.95**	64.79	64.92	64.86	64.61
Qwen2.5 VL	62.24	**62.59**	62.52	62.43	62.56	62.39
LLaMA Vision	62.65	**62.85**	62.40	62.59	62.62	62.70
Avg.	**62.83**	62.70	62.59	62.60	62.65	62.60
VizWiz						
LLaVA 7B	61.96	61.79	62.13	61.99	61.97	**62.23**
LLaVA 13B	**64.46**	63.27	63.57	63.63	63.91	64.05
Qwen2.5 VL	68.59	68.75	**68.89**	68.52	68.50	68.42
LLaMA Vision	**70.44**	69.88	69.76	70.11	70.27	69.99
Avg.	**66.36**	65.92	66.09	66.06	66.16	66.17

Rich. This prompt instructs the model to generate as rich a description as possible. We expect the resulting descriptions to include many details and a large amount of information. The exact prompt we use is shown below:

Prompt for Rich.

```
You are a creative image captioning assistant.  Provide a rich, detailed description
of the image, highlighting objects, setting, colors, textures, actions, emotions, and
context.  Use complete sentences and vivid language without apologies or filler phrases.
```

24 words with QA. This prompt instructs the model to produce a QA-related description in no more than 24 words. By limiting the amount of information, the description is encouraged to focus more directly on objects that are relevant to answering the question. The exact prompt we use is shown below:

Prompt for 24 words with QA.

```
You are an objective image-captioning assistant.  Describe strictly what you see in the
image in no more than 24 words.  Ensure the caption contains the details required to
answer the question.  Write as one continuous paragraph-do NOT use bullet points, lists,
apologies, opinions, or speculative content.
```

Results. The results are shown in Table 11. *Default* denotes the prompt used in our main experiments (see Section A.3 for details). We do not observe any large performance differences across the generation prompts. If anything, longer descriptions

Table 11. Comparison of generation prompts. Each value shows accuracy (%) for different prompts in TPI.

	Description Length				
Model	Default	50 words	200 words	Rich	24 words with QA
ScienceQA					
LLaVA 7B	75.11	74.62	74.37	**75.71**	74.57
LLaVA 13B	76.30	76.65	76.30	**77.14**	76.60
Qwen2.5 VL	90.43	90.68	**91.27**	90.33	90.78
LLaMA Vision	**90.93**	89.84	88.94	88.45	89.69
Avg.	**83.19**	82.95	82.72	82.91	82.91
OK-VQA					
LLaVA 7B	**61.70**	60.24	60.48	59.91	59.84
LLaVA 13B	64.73	64.72	**64.83**	64.58	64.52
Qwen2.5 VL	62.24	62.05	62.23	**62.62**	62.37
LLaMA Vision	62.65	**62.77**	62.63	62.51	62.53
Avg.	**62.83**	62.44	62.54	62.40	62.32
VizWiz					
LLaVA 7B	61.96	**62.06**	60.75	61.86	61.90
LLaVA 13B	**64.46**	62.93	63.26	62.96	63.90
Qwen2.5 VL	68.59	**69.82**	68.48	68.96	68.51
LLaMA Vision	**70.44**	69.22	69.49	68.62	69.02
Avg.	**66.36**	66.01	65.50	65.60	65.83

tend to yield slightly better performance. For example, 200words, rich, and default prompts achieve marginally higher scores than the others.

These results suggest that the amount of information in the textual description is not crucial for learning, as long as it contains the minimum set of correct information. In practice, there is little need to make the textual descriptions overly high-quality. Instead, increasing the diversity of QA pairs may be more beneficial for training.

D.4. Generation Model

In text-centric training, one of the main ways to build data is to automatically generate textual descriptions with LVLMs. Our experiments also follow this approach. In Section 4, we automatically generate textual descriptions from ground-truth images using an LVLM. In Section 5, we use the same strategy for generating new samples in the data-augmentation experiments.

In this section, we evaluate how differences in the generation model affect the final learning performance. To reduce the influence of model-specific writing bias and focus on the model's core ability, we compare three models from the same Qwen family: Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, and Qwen2.5-VL-32B-Instruct. This setup keeps the style of the generated text consistent while changing only the model size and capability, allowing us to isolate how these factors contribute to text-centric training.

The results are shown in Table 12. We observe a clear trend: using a stronger model to generate textual descriptions leads to better VLM performance after training. In particular, textual descriptions produced by the strongest model, Qwen2.5-VL-32B-Instruct, achieve the highest average scores. A likely reason is that larger models can extract visual information more accurately and produce fewer mistakes in their descriptions.

Overall, these results suggest that textual descriptions do not need to be overly detailed. What matters more is how accurately they capture the information in the image. This contrasts with the ablation of generation prompts in Section D.3, where changing the amount of information had little effect. However, changing the generation model does produce differences. This implies that correctness and relevance of the description are more important than its length. The finding also aligns with our observation that training with T2I-generated images, whose relevance scores are low, provides only

Table 12. Comparison of generation models. Each column corresponds to the Qwen model used for textual description generation.

Model	Generation Model		
	Qwen2.5 VL 3B	Qwen2.5 VL 7B	Qwen2.5 VL 32B
ScienceQA			
LLaVA 7B	**75.11**	74.96	**75.11**
LLaVA 13B	76.00	**76.80**	76.30
Qwen2.5 VL	90.38	**91.72**	90.43
LLaMA Vision	88.10	88.55	**90.93**
Avg.	82.40	83.01	**83.19**
OK-VQA			
LLaVA 7B	59.61	60.31	**61.70**
LLaVA 13B	64.24	64.67	**64.73**
Qwen2.5 VL	**62.57**	62.49	62.24
LLaMA Vision	60.48	61.02	**62.65**
Avg.	61.72	62.12	**62.83**
VizWiz			
LLaVA 7B	60.27	58.82	**61.96**
LLaVA 13B	61.57	60.75	**64.46**
Qwen2.5 VL	69.09	**71.05**	68.59
LLaMA Vision	69.54	69.35	**70.44**
Avg.	65.12	64.99	**66.36**

limited improvements.

E. Detailed Results for Data Augmentation

This section presents the complete results of the data augmentation experiments in Section 5.

Table 13 shows that **TPI** consistently outperforms *Text-only* training and is often competitive with, or better than, *T2I* across backbones and data regimes. This trend is especially clear for LLaVA 7B, LLaVA 13B, and LLaMA Vision, where TPI achieves strong performance on benchmarks such as ChartQA, DocVQA, InfoVQA, and DriveLM.

The comparison with the *Orig.* baselines further shows that these gains are not explained solely by fine-tuning on the seed data. In the 1% and 10% regimes, TPI often improves over the corresponding original-only training. For stronger backbones such as Qwen2.5-VL, the margin is smaller because the original-only baseline is already strong, but TPI still remains competitive across multiple tasks.

E.1. Negative Transfer Evaluation

Setup. We additionally examine whether fine-tuning on each benchmark causes negative transfer to generic visual question answering ability. After fine-tuning each model on a target task, we evaluate the resulting model on **GQA** and **VQAv2** without further tuning. Since these two benchmarks rely on natural-image understanding, they serve as a simple probe of whether task-specific fine-tuning degrades general visual knowledge. Following prior work [71], TPI is used only at training time, while evaluation is performed with standard image inputs.

Results. Tables 14 report the results. Each entry is formatted as *GQA / VQAv2*. Overall, TPI shows no substantial negative transfer. Across target tasks and models, TPI is consistently much more stable than T2I, and is generally comparable to standard GT-image fine-tuning. In many cases, TPI also matches GT-image training on these off-target evaluations. These results suggest that TPI improves the training interface between text supervision and LVLMs without introducing noticeable degradation beyond ordinary fine-tuning.

Table 13. Full results of data augmentation experiments.

Model	Methods	**ScienceQA**	**OK-VQA**	**ChartQA**	**DocVQA**	**InfoVQA**	**VizWiz**	**DriveLM**
LLaVA 7B	Pretrained	66.09	49.87	17.12	23.86	21.91	52.89	47.84
	Orig. (1%)	64.15	54.21	20.04	27.50	23.32	43.92	58.92
	+ Text-only	**65.44**	55.03	17.56	27.09	24.82	41.07	55.38
	+ T2I	64.01	52.00	20.00	27.60	25.37	**53.56**	62.25
	+ **TPI (Ours)**	64.20	**55.62**	**22.56**	**28.86**	**25.80**	51.44	**63.22**
	Orig. (10%)	67.43	59.73	**28.56**	33.28	**27.42**	62.72	66.61
	+ Text-only	62.02	57.85	18.32	27.18	24.43	45.08	67.21
	+ T2I	**69.36**	58.78	27.96	32.89	27.05	63.05	67.34
	+ **TPI (Ours)**	68.57	**60.26**	28.44	**33.71**	27.27	**63.41**	**69.24**
	Orig. (100%)	78.78	**62.12**	17.12	23.86	21.91	52.89	47.84
	+ Text-only	71.15	60.29	19.12	27.68	24.80	59.98	64.56
	+ T2I	79.47	61.05	37.24	39.89	**29.43**	66.33	73.46
	+ **TPI (Ours)**	**80.27**	61.21	**37.44**	**40.43**	28.59	**66.47**	**73.60**
LLaVA 13B	Pretrained	71.39	52.49	19.32	27.86	25.87	56.33	51.93
	Orig. (1%)	71.05	49.27	22.56	31.57	27.27	49.29	55.32
	+ Text-only	**71.15**	58.05	19.68	31.51	27.41	52.02	56.76
	+ T2I	71.05	55.87	23.04	31.12	28.13	58.00	56.45
	+ **TPI (Ours)**	70.40	**61.04**	**25.96**	**32.42**	**28.52**	**59.91**	**62.84**
	Orig. (10%)	71.89	63.42	32.00	**36.79**	30.70	**63.50**	66.66
	+ Text-only	71.24	62.03	21.20	32.36	28.36	59.98	61.34
	+ T2I	72.78	62.52	31.72	36.37	**30.91**	62.14	**69.19**
	+ **TPI (Ours)**	**73.28**	**64.26**	**32.16**	36.42	30.30	62.96	68.64
	Orig. (100%)	80.12	64.53	40.28	43.07	33.73	66.96	**74.86**
	+ Text-only	74.47	64.02	24.08	33.05	29.22	64.01	64.66
	+ T2I	**81.80**	64.55	41.00	**44.11**	33.59	66.57	73.32
	+ **TPI (Ours)**	81.66	**64.56**	**41.88**	43.86	**34.24**	**67.36**	73.04
Qwen2.5 VL	Pretrained	76.65	43.88	83.28	94.36	80.21	70.80	39.49
	Orig. (1%)	88.15	51.52	84.76	91.92	**78.99**	66.29	**44.51**
	+ Text-only	86.96	60.28	**86.44**	91.82	78.89	63.71	44.11
	+ T2I	86.81	59.07	85.68	92.17	77.71	68.26	41.83
	+ **TPI (Ours)**	**88.25**	**61.94**	86.12	**92.34**	77.86	**69.29**	43.37
	Orig. (10%)	88.00	**63.25**	86.28	**92.50**	**78.87**	**70.62**	63.95
	+ Text-only	88.15	63.19	86.68	92.21	78.64	64.58	59.47
	+ T2I	89.94	62.07	86.52	92.15	78.08	70.47	**67.18**
	+ **TPI (Ours)**	**90.33**	62.81	**86.72**	92.36	78.28	70.01	66.92
	Orig. (100%)	93.51	61.63	86.76	**93.89**	**78.77**	70.54	75.16
	+ Text-only	90.48	**63.70**	86.64	91.61	78.33	68.49	61.33
	+ T2I	94.99	62.13	87.20	92.26	78.24	70.59	76.12
	+ **TPI (Ours)**	**95.49**	62.55	**87.52**	92.35	78.11	**70.94**	**77.01**
LLaMA Vision	Pretrained	50.77	25.99	22.28	80.83	46.98	58.32	34.33
	Orig. (1%)	87.90	53.94	72.80	90.32	67.82	**67.31**	36.31
	+ Text-only	66.04	34.95	22.24	78.67	61.97	58.40	35.48
	+ T2I	87.36	54.09	73.80	89.71	67.36	63.74	47.39
	+ **TPI (Ours)**	**87.96**	**57.23**	**74.00**	**91.04**	**68.01**	64.09	**47.53**
	Orig. (10%)	88.60	**61.04**	73.64	90.78	67.86	**68.88**	46.35
	+ Text-only	75.21	42.46	38.88	77.20	64.97	60.72	46.04
	+ T2I	**89.49**	59.66	73.92	90.75	67.56	68.05	56.41
	+ **TPI (Ours)**	89.39	60.18	**74.28**	**90.86**	**68.20**	68.00	**59.63**
	Orig. (100%)	93.65	63.04	**76.48**	92.47	67.90	**72.32**	55.16
	+ Text-only	85.67	51.30	49.08	84.57	62.49	57.10	45.04
	+ T2I	94.15	**63.48**	76.12	91.73	68.11	71.52	57.12
	+ **TPI (Ours)**	**94.60**	63.38	76.24	**92.83**	**68.54**	71.61	**57.43**

Table 14. Negative transfer evaluation on **GQA/VQAv2** after fine-tuning on each target task. Each entry is written as **GQA / VQAv2**.

	Methods	Models			
		LLaVA 7B	LLaVA 13B	Qwen	LLaMA
	Pretrained	60.6 / 70.1	61.9 / 74.5	60.9 / 78.6	43.4 / 66.0
ScienceQA	Text-only	59.6 / 69.3	61.8 / 72.5	58.2 / 74.1	27.7 / 56.0
	T2I	57.3 / 62.5	60.9 / 68.6	59.7 / 76.2	41.5 / 65.4
	TPI (Ours)	58.7 / 66.1	62.4 / 73.5	59.7 / 77.0	40.2 / 62.5
	GT-Image	58.6 / 65.8	61.4 / 70.7	59.5 / 77.0	40.6 / 62.4
OK-VQA	Text-only	57.0 / 67.4	60.4 / 71.2	58.9 / 78.1	46.1 / 68.6
	T2I	51.2 / 64.5	56.4 / 68.6	55.1 / 75.2	45.9 / 66.0
	TPI (Ours)	53.2 / 65.8	58.7 / 71.8	57.9 / 75.9	48.8 / 72.5
	GT-Image	54.7 / 66.8	57.8 / 71.2	56.8 / 77.6	48.7 / 72.1
VizWiz	Text-only	57.0 / 65.7	60.5 / 70.7	58.5 / 75.3	35.9 / 60.5
	T2I	53.3 / 61.6	57.1 / 66.8	53.1 / 65.5	45.8 / 66.2
	TPI (Ours)	54.3 / 62.7	59.1 / 68.8	56.9 / 74.4	48.0 / 67.4
	GT-Image	53.9 / 64.5	58.6 / 68.2	55.0 / 69.7	48.3 / 68.6
ChartQA	Text-only	59.9 / 68.3	62.1 / 73.9	60.5 / 77.9	36.7 / 55.9
	T2I	55.7 / 61.7	60.2 / 68.5	59.6 / 74.5	40.0 / 61.6
	TPI (Ours)	57.5 / 65.9	60.9 / 73.4	60.9 / 78.9	46.0 / 65.6
	GT-Image	58.8 / 65.6	61.2 / 73.0	61.1 / 78.7	41.0 / 59.3
InfoVQA	Text-only	59.5 / 68.7	61.6 / 73.3	59.9 / 76.9	47.5 / 69.3
	T2I	54.6 / 64.3	60.4 / 69.7	54.4 / 72.7	42.8 / 64.3
	TPI (Ours)	58.6 / 66.3	61.5 / 73.3	60.0 / 78.7	48.2 / 69.1
	GT-Image	57.8 / 64.5	61.0 / 70.8	59.7 / 77.9	50.1 / 69.5
DocVQA	Text-only	59.2 / 65.9	61.6 / 71.7	59.9 / 77.6	30.8 / 53.8
	T2I	49.8 / 55.8	58.0 / 63.3	53.8 / 66.6	37.5 / 55.9
	TPI (Ours)	56.3 / 63.0	60.1 / 73.1	59.9 / 78.3	47.4 / 65.5
	GT-Image	56.0 / 63.1	60.6 / 69.6	59.9 / 78.4	47.0 / 66.6
DriveLM	Text-only	59.8 / 68.1	61.9 / 73.8	60.6 / 75.0	31.2 / 53.9
	T2I	59.4 / 68.6	60.9 / 72.0	61.2 / 78.0	38.6 / 56.4
	TPI (Ours)	60.5 / 69.0	61.2 / 71.7	61.5 / 78.2	48.2 / 68.9
	GT-Image	59.6 / 69.2	60.8 / 71.6	60.8 / 77.8	48.1 / 68.6

F. Qualitative Examples

To support our quantitative analysis, we present examples of the synthetic images used for training. In particular, to compare T2I and TPI more effectively, we highlight cases where T2I images deviate from the provided textual descriptions and consequently receive low relevance scores.

Figures 5–10 shows the examples. They reveal that T2I often produces images that contradict the QA pair or omit essential information. Consistent with our overall findings, T2I struggles especially in Text VQA, where generating readable and accurate text is critical. We believe this difficulty arises from the inherent limitations of T2I models in rendering text reliably.

Question: Which type of force from the bulldozer clears the path?
Choices: [“push”, “pull”]
Answer: push

Ground Truth Image | T2I-generated Image | Text-printed image (ours)

A yellow bulldozer with a large blade is operating on a dirt construction site. The bulldozer is positioned on a mound of earth, pushing or leveling the soil. The machine has a visible operator's cabin with a person inside, and the tracks of the bulldozer are clearly visible as it moves across the terrain. The background features a clear blue sky and a distant hill or embankment, indicating an open outdoor environment. The scene suggests ongoing earthmoving or construction work.

GT: Pushing action | ✗ Does not depict “push” | ✓ Include “push”

Question: Which continent is highlighted?
Choices: ['Europe', 'South America', 'Antarctica', 'North America']
Answer: South America

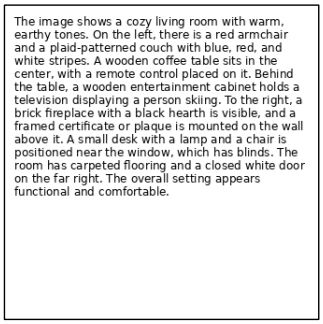
Ground Truth Image | T2I-generated Image | Text-printed image (ours)

The image is a world map with a simple, flat design. The continents are depicted in a uniform light green color, while the oceans and seas are shown in a solid blue. The map is centered on the Americas, with South America highlighted in a darker green shade, distinguishing it from the rest of the world. The map includes all major continents: North America, South America, Europe, Africa, Asia, Australia, and Antarctica, which is partially visible at the bottom. The map does not contain any labels, borders, or additional geographical details.

GT: Highlighted continent | ✗ Does not highlight South America | ✓ Includes the highlighted continent

Figure 5. Qualitative comparison of synthetic images on ScienceQA.

Question: What material is the fireplace made from?
Answer: brick

Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		
GT: Shows a brick fireplace	✗ No fireplace shown	✓ Includes “brick fireplace”

Question: What is the tennis player doing?
Answer: serving

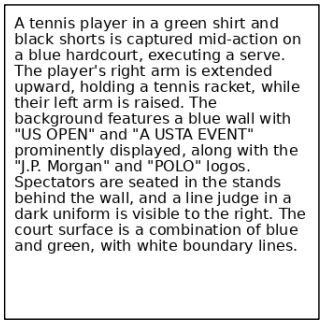
Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		
GT: Serving action	✗ Not Serving	✓ Mentions a serving motion

Figure 6. Qualitative comparison of synthetic images on OK-VQA.

Question: What kind of wine is this?
Answer: chardonnay

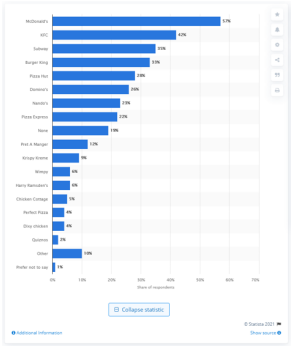
Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		A hand is holding a green wine bottle with a label that reads "Cul-de-Sac Wine Company." The label features an illustration of a yellow house with a red roof and a sun above it. Below the illustration, the text indicates "Chardonnay" and "California." The bottle appears to be partially filled with a yellowish liquid, likely wine. The background includes a white surface with some text or markings, and part of a white object, possibly a wall or cabinet, is visible. The lighting is warm, and the image is slightly tilted.
GT: Shows chardonnay	✗ Label does not show “chardonnay”	✓ Includes “Chardonnay”

Question: What is this?
Answer: phone

Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		The image shows a black, rectangular device with a glossy screen, likely an older model smartphone or iPod. The device has a home button with a white square icon in the center at the bottom. The screen appears to be off or reflecting light, showing faint smudges or fingerprints. The background is a textured, speckled surface, possibly a table or countertop, with a bright light source in the upper right corner causing a glare on the device and the surface. The overall appearance suggests the device is placed on a flat, non-reflective surface.
GT: Shows a phone	✗ Does not show a phone	✓ Describes a phone

Figure 7. Qualitative comparison of synthetic images on VizWiz.

Question: Which fast food chain has the most popularity in the UK?
Answer: McDonald's

Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		The image is a horizontal bar chart illustrating the popularity of various fast food chains in the UK, based on the share of respondents. The chart features a white background with blue horizontal bars representing different fast food chains. Each bar is labeled with the name of the chain and its corresponding percentage of popularity. McDonald's is at the top with 57%, followed by KFC at 42%, Subway at 35%, and Burger King at 33%. Other chains, such as Pizza Hut, Domino's, Nando's, and Pizza Express, have lower percentages, ranging from 22% to 28%. The chart also includes categories like "None" (19%), "Pret A Manger" (12%), and "Other" (10%), as well as smaller percentages for chains like Krispy Kreme (9%), Wimpy (6%), and Harry Ramsden's (6%). The bottom of the chart includes a "Collapse statistic" button and a Statista 2021 watermark, indicating the source of the data. The layout is clean and organized, with percentages clearly marked next to each bar.
GT: Shows McDonald's at the top	✗ Does not show a bar chart	✓ Includes McDonald's as most popular

Question: What is the projected decrease in the average end-use price of electricity by 2050?
Answer: 9.6

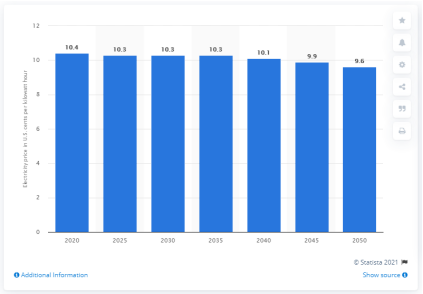
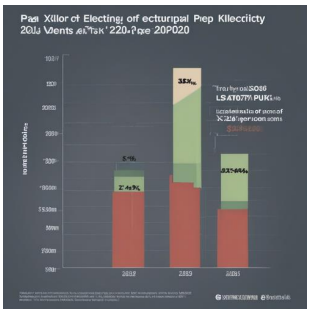
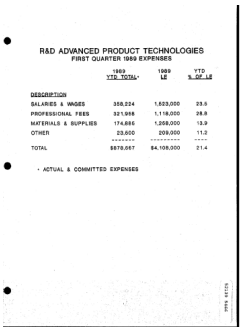
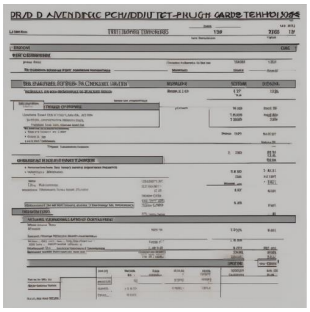
Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		The image is a bar chart depicting the projected average end-use price of electricity in U.S. cents per kilowatt hour (kWh) from 2020 to 2050. The chart has a horizontal x-axis labeled with years (2020, 2025, 2030, 2035, 2040, 2045, 2050) and a vertical y-axis labeled "Electricity price in U.S. cents per kilowatt hour," ranging from 0 to 12. Each year is represented by a blue bar, with the height of the bar corresponding to the projected price. The prices are as follows: 2020 (10.4 cents), 2025 (10.3 cents), 2030 (10.3 cents), 2035 (10.3 cents), 2040 (10.1 cents), 2045 (9.9 cents), and 2050 (9.6 cents). The chart shows a gradual decrease in electricity prices over time, with the price in 2050 projected to be 9.6 cents per kWh. The chart includes a Statista 2021 watermark in the bottom right corner, along with options for additional information and source details. The overall layout is clean and straightforward, with a white background and grid lines for reference.
GT: Shows the 2050 value "9.6"	✗ Shows an incorrect value	✓ Includes correct value

Figure 8. Qualitative comparison of synthetic images on ChartQA.

Question: Which year's expenses is mentioned in this document?
Answer: 1989

Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		The image is a document titled "R&D ADVANCED PRODUCT TECHNOLOGIES FIRST QUARTER 1989 EXPENSES." It is a financial report detailing expenses for the first quarter of 1989. The document is structured in a tabular format with three columns: "1989 YTD TOTAL," "1989 LE," and "% OF LE." The rows list various expense categories, including "SALARIES & WAGES," "PROFESSIONAL FEES," "MATERIALS & SUPPLIES," and "OTHER." Each category has corresponding numerical values for the year-to-date total, the 1989 line estimate (LE), and the percentage of the LE. At the bottom, the total expenses are summarized as $878,567 for the YTD total and $4,108,000 for the LE, with a percentage of 21.4%. The document also includes a note at the bottom indicating that the figures represent "ACTUAL & COMMITTED EXPENSES." The text is black on a white background, and the layout is clean and organized. The page number "52139 5466" is visible in the bottom right corner.
GT: Shows the year “1989”	✗ Does not show the year	✓ Includes the year “1989”

Question: What is the date mentioned in this letter?
Answer: May 22, 1978

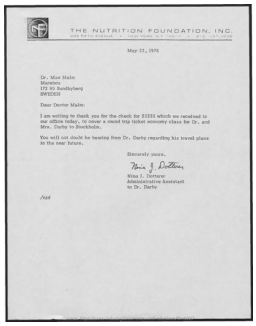
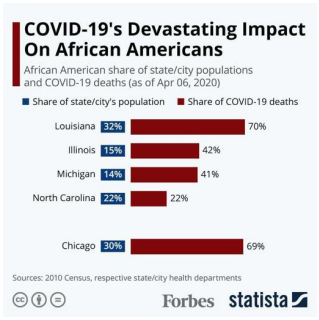
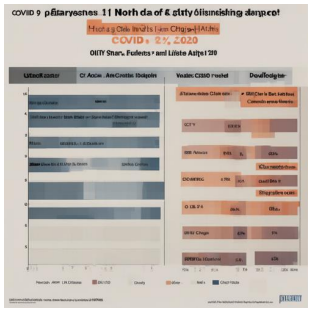
Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		The image is a formal letter on letterhead from "THE NUTRITION FOUNDATION, INC." located at 489 Fifth Avenue, New York, N.Y. 10017, with a phone number listed as 212-687-4830. The letterhead includes a logo featuring a stylized "N" in a square. The date "May 22, 1978" is prominently displayed near the top center of the page. The letter is addressed to "Dr. Max Malm" at Marabou, 172 85 Sundbyberg, Sweden. The body of the letter, written by "Nina J. Dotterer, Administrative Assistant to Dr. Darby," thanks Dr. Malm for a check of $2235 received to cover a round-trip economy class ticket for Dr. and Mrs. Darby to Stockholm. It mentions that Dr. Darby will contact Dr. Malm regarding travel plans. The letter is signed with a handwritten signature followed by Nina J. Dotterer's typed name and title. The bottom left corner has a typed initials "/njd." The overall layout is clean and professional, with a formal tone.
GT: Shows the date “May 22, 1978”	✗ Shows an incorrect date	✓ Includes the correct date

Figure 9. Qualitative comparison of synthetic images on DocVQA.

Question: What percentage of Covid-19 deaths of African American were reported in North Carolina as of Apr 06, 2020?
Answer: 22%.

Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		The image is a bar chart titled "COVID-19's Devastating Impact On African Americans," showing the African American share of state/city populations and COVID-19 deaths as of April 06, 2020. The chart compares data for Louisiana, Illinois, Michigan, North Carolina, and Chicago. Each state/city is listed on the left, with two horizontal bars representing the "Share of state/city's population" (in blue) and the "Share of COVID-19 deaths" (in red). For North Carolina, the blue bar indicates a 22% share of the population, while the red bar shows a 22% share of COVID-19 deaths. The chart uses percentages to compare these shares, and the sources are cited as the 2010 Census and respective state/city health departments. The chart is branded with the Forbes and Statista logos at the bottom.
GT: Shows the value “22%”	✗ Missing the value	✓ Includes the value

Question: How much time will the route take?
Answer: 40 HOURS.

Ground Truth Image	T2I-generated Image	Text-printed image (ours)
		The image is an infographic titled "#WildAtlanticWay," promoting Ireland's West Coast touring route. It features a bright, colorful design with illustrations and text. At the top, a green banner displays the title, followed by a brief description of the route, which stretches from Donegal to West Cork, covering 2,750 kilometers and taking 40 hours. Key details include 156 discovery points and a €10 million investment. A map highlights the route, marking cities like Donegal, Sligo, Galway, Kerry, and Cork. The infographic lists major expenses: €500 for fuel, €250 per week for car rental, and accommodation options ranging from €50-150 per night at hotels/B&Bs, €10 for camping, and €25 for a camper. It also mentions "Don't Miss" attractions like the Cliffs of Moher, The Burren, and Dingle Bay. Additional details include costs for food and drink, an old Irish blessing, and social media handles. The overall layout is organized into sections with icons, text, and visual elements to convey information clearly.
GT: Shows “40 hours”	✗ Missing the time value	✓ Includes the time value

Figure 10. Qualitative comparison of synthetic images on InfoVQA.