Visual Lexicon: Rich Image Features in Language Space

Supplementary Material

A1. Technical Details

We introduced the main technical and implementation details of our ViLex model in the main paper, here we provide a more comprehensive explanation.

Text-to-image diffusion model. Following DeDiffusion [76], we use Imagen [60] as the base text-to-image diffusion model, adapting the U-Net architecture from [49, 57] with 600M parameters, an embedding dimension of 256, and an input resolution of 64×64. The text encoder of Imagen is OpenCLIP ViT-H/14 [13, 32] with a vocabulary size of 49408. The U-Net conditions on text embeddings via a pooled embedding vector, which is added to the diffusion timestep embedding. Imagen is further conditioned on the full sequence of text embeddings by incorporating cross-attention over the text embeddings at multiple resolutions. The Imagen model uses v-prediction [61] as its objective, with a batch size of 2048, and is trained for 3 million steps. As a baseline model, Imagen achieves an FID of 6.52 on 30K 64×64 MS-COCO 2014 validation images [60]. During image generation inference, we use a super-resolution model, such as an SDXL upsampler, to upsample the image resolution from 64×64 to 512×512 for better visualizations.

Model architecture of ViLex. ViLex consists of two components: a ViT-based image encoder and a transformer-based attention pooling module. Both components are unfrozen during the training process. For the image encoder, we use a pretrained SigLIP-So400M@224 [88]. SigLIP utilizes ViT-base as the backbone and is pretrained on the WebLI dataset [11] using a sigmoid loss and trained on English image-text pairs, with input images resized to 224×224. The model architecture of the ViT-base is shape-optimized on 400M training samples for improving the model efficiency and speed. In our method, the attention pooler is implemented as a single multi-head attention layer with learnable queries, using the encoder output as both keys and values. This allows the attention pooling module to effectively aggregate embeddings of varying lengths. The attention pooling module contains n learnable queries, where $n \leq 75$, along with [SOS] and [EOS] tokens to ensure the total token count remains within the 77-context length limit defined by the CLIP text encoder [32, 53]. The attention pooling layer comprises 5 transformer blocks, which are always randomly initialized.

Model training. The training data is obtained from WebLI [11], enabling training on either images alone or with image-text pairs. We found that joint image-text training and our TFG are essential for enabling multimodal image generation. However, training without text captions does not negatively impact performance on downstream vision-language tasks. Following [60, 76], we use Adafactor optimizer [63] and a weight decay of 0.01. Training is performed with a batch size of 2048 over 300K steps, which takes approximately 2.5 days on 64 TPUv5 chips. We found that double the training steps (from 300k to 600k) can further improve the model performance on increasing the performance of a pretrained vision encoder. The ViT is initialized with a pretrained SigLIP model and the attention pooling layers are randomly initialized. We use learning rate 1×10^{-5} for the image encoder and 3×10^{-4} for the attention pooling layers, with a cosine learning rate decay and a 10K-step linear warmup, and a weight decay of 0.01. After training, our ViLex encoder maps an image to ViLex representations. We next evaluate two capabilities of these frozen ViLex representations: image generation and visual understanding.

PaliGemma experiments. To evaluate the effectiveness of the proposed ViLex approach in enhancing a pretrained vision encoder for vision-language tasks, we integrate our vision encoder into the PaliGemma [6] framework and replace the vision encoder with either the fine-tuned SigLIP-So400M [88] from ViLex or the official version without model fine-tuning, freezing the vision encoder and fine-tuning the model on downstream tasks. Following PaliGemma's official pipeline, we transfer the model to a variety of individual academic benchmarks using a unified transfer approach with minimal hyperparameter tuning. To ensure fair comparison, we applied the same hyperparameter sweeping strategy for both the baseline and our fine-tuned vision encoder, reporting the best results for each. This structured approach allows us to fairly assess the impact of the proposed ViLex method on a wide range of vision-language tasks. The sweeping parameters for these tasks are as follows: COCOCap [40] (COCO image captioning task) and COCO-35L [70] (COCO captions translated in 35 languages): learning rate (4e-6, 5e-6, 6e-6), epochs (5, 10), dropout (0, 0.02, 0.05). TextCaps [64] (image captioning with reading comprehension): learning rate (4e-6, 6e-6), and training epochs (5, 10). For SciCaps [28] (captions for scientific figures): learning rate (6e-5, 7e-5), dropout (0.1, 0.2), and label smoothing (0.1, 0.2). For VQAv2 [21] (visual question answering): label smoothing (0.0, 0.1), dropout (0.0, 0.1), and weight decay (0, 1e-6). For TextVQA [65] (visual reasoning based on text in images): learning rate (4e-6, 6e-6). For OKVQA [46] (outside knowledge VQA), ScienceQA [43] (science question answering), and VizWizVQA [22] (VQA from people who are blind): learning rate (8e-6, 1e-5), and dropout (0.0, 0.02).

		Image Captioning					Visual Question Answering						Image Segmentation						Video
Backbone	#steps	*COCOcap*	*COCO-35L*	*TextCap*	*SciCap-Val*	*SciCap-Test*	*VQAv2-Val*	*TextVQA*	*OKVQA*	*SciQA*	*VizWizQA*	*GQA*	*RC-val*	*RC-testA*	*RC-testB*	*RCp-testA*	*RCp-testB*	*RCg-test*	*MSRVTT*
Original SigLIP	-	139.7	138.6	122.1	131.7	135.5	81.4	51.9	57.1	85.9	74.3	64.8	66.2	69.0	63.6	63.3	55.3	59.6	69.4
ViLex SigLIP	150k	**140.5**	**138.8**	**122.3**	**132.6**	**135.5**	**81.4**	**52.1**	**57.3**	**86.1**	**74.5**	**65.1**	**66.5**	**69.3**	**64.2**	**64.1**	**55.6**	**60.2**	**70.6**
ViLex SigLIP	600k	**141.5**	**140.0**	**124.0**	**134.3**	**136.1**	**81.8**	**52.7**	**58.3**	**89.3**	**75.0**	**65.4**	**67.5**	**69.7**	**65.6**	**65.2**	**57.2**	**62.6**	**71.4**

Table A1. ViLex improves both image understanding and reconstruction capabilities of vision encoders by fine-tuning them using ViLex's training approach. **Extending the fine-tuning of SigLIP with the ViLex approach from 150k to 600k steps results in improved overall model performance across evaluated benchmarks.** We use PaliGemma's [6] framework for linear evaluation, replacing the vision encoder with either the fine-tuned SigLIP in ViLex or the official one, and freeze vision encoder and fine-tune the model on downstream tasks. We use the same hyper-parameters and model architecture for a fair comparison.

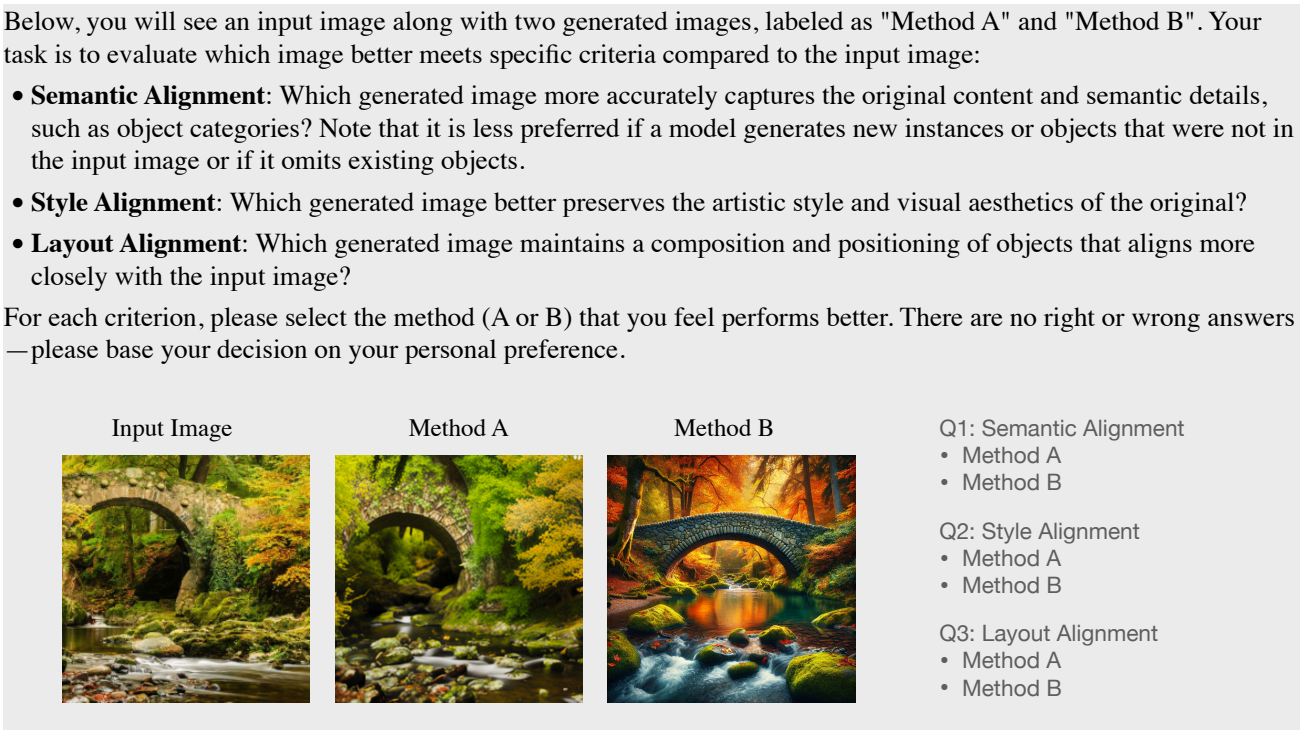

Figure A1. The instructions and question format used for human study.

For GQA [30] (VQA on image scene graphs): learning rate (5e-6, 1e-5), and dropout (0.0, 0.02, 0.05). For RefCOCO [33, 45, 83] (referring expression segmentation): label smoothing (0.1, 0.2), epochs (60, 100), and dropout (0, 0.05). For MSRVTT-Caps [78] (open-domain short video captioning): weight decay (0, 1e-6), dropout (0, 0.2), and epochs (20, 40).

A2. Human Study

We conduct human studies to evaluate the quality of generated images using an image-to-image pipeline, focusing on three criteria: Semantic Alignment, Style Alignment, and Layout Alignment. For Semantic Alignment, participants judge which generated image more accurately captures the original content and semantic details, such as object categories. Introducing new instances or omitting existing ones from the input image is considered less desirable. For Style Alignment, participants assess which generated image best retains the artistic style and visual aesthetics of the original. For Layout Alignment, participants evaluate which generated image maintains a composition and positioning of objects that closely matches the input image.

The results of this evaluation are reported in Table 2 of the main paper. Detailed instructions and the question for-

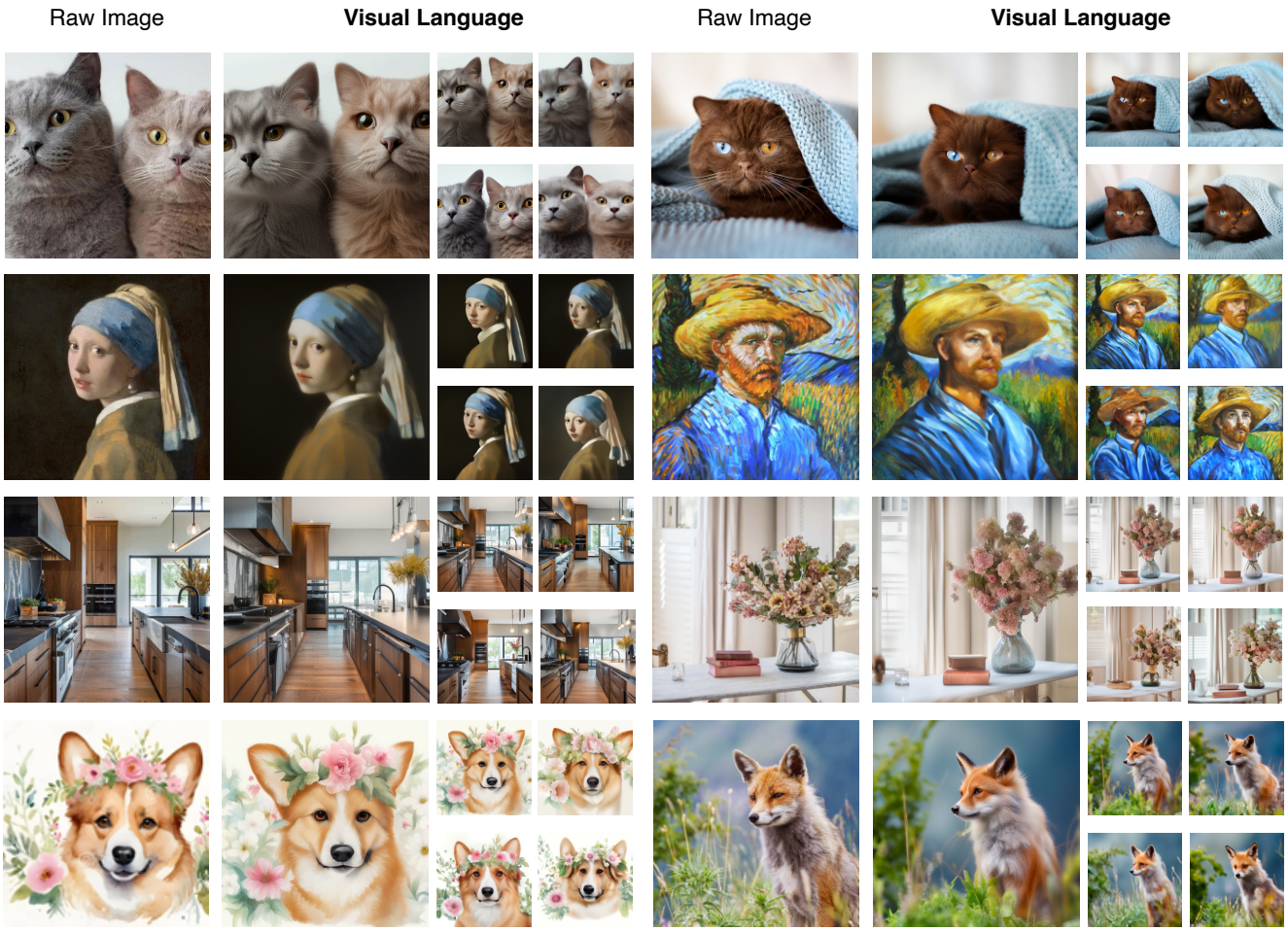

Figure A2. More demo results of generating a set of images (generated under different diffusion noises), which are highly semantically and visually similar to each other, by using ViLex tokens as "text" prompts for text-to-image diffusion models.

mat for the human study are shown in Figure A1.

A3. Ablation Study

Training Steps. We observed that extending the fine-tuning steps of the vision encoder using our ViLex pipeline leads to improved performance across nearly all evaluated benchmarks, as shown in Table A1. Specifically, increasing the training steps from 150k to 300k yields significant gains. Further extending the training to 600k steps provides marginal improvements compared to the 300k-step results. The largest improvements are observed in datasets that demand stronger spatial understanding, such as the referring expression segmentation datasets RefCOCO/+/g.

Number of attention pooling layers. Although increasing the number of attention pooling layers improves image reconstruction performance (as indicated by a lower FID score), it also introduces a trade-off with image understanding capabilities. As shown in Table A2, we found that using 5 attention pooling layers provides the optimal balance between image generation quality and developing an effective vision encoder for visual scene understanding.

#layers	FID	COCOCaps
2	2.62	140.7
5	2.58	141.5
8	2.52	141.0

Table A2. Ablation study on number of attention pooling layers.

Vision encoders. The ViLex approach effectively enhances various vision encoders for downstream visual scene understanding tasks. We initialize the vision encoder of ViLex with either the CoCa [82] pretrained ViT or the SigLIP [88] pretrained ViT-So400M. Similar to our experiments in previous sections, We observed consistent performance improvements for both image understanding tasks, such as COCOCaps [40], and video understanding tasks, such as

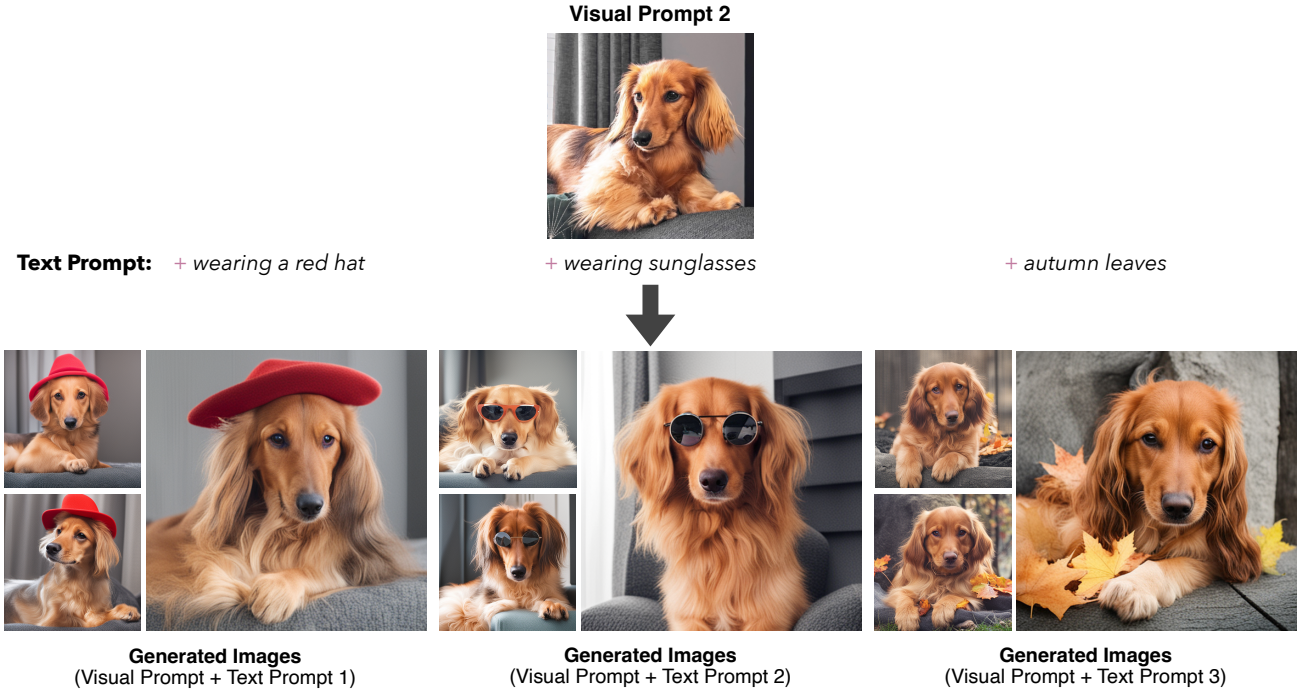

Figure A3. More demo results of zero-shot accessorization via prompting a frozen text-to-image generation model with our visual prompts (*i.e.*, ViLex tokens) and text prompts from natural language.

MSRVTT-Caps [78]. Compared to a roughly 2% improvement for SigLIP in terms of CIDEr score on COCOCaps, the gains for CoCa were even more substantial, reaching approximately 4%. The flexibility to consistently improve different pretrained models demonstrates ViLex's generalizability across various types of vision encoders.

Datasets	CoCa	F.T. w/ ViLex	SigLIP	F.T. w/ ViLex
COCOCaps	131.6	**135.8**	139.7	**141.5**
MSRVTTCaps	56.1	**60.2**	69.4	**71.4**

Table A3. Fine-tuning vision encoders with the ViLex approach enhances image understanding performance across various pretrained models, including CoCa [82] and SigLIP [88]. F.T. denotes fine-tuning the vision encoder, such as CoCa, during the ViLex model pretraining stage.

A4. Demo Results

Semantic-level image reconstruction. In this section, we present additional demo results in Figure A2, showcasing a set of images generated with varying diffusion noises and different random seeds. These images demonstrate high semantic and visual consistency, leveraging ViLex tokens as "text" prompts for text-to-image diffusion models. However, as shown in the results, our model occasionally misses small objects in the scene. This limitation primarily stems from using a low-resolution text-to-image diffusion model as the base during the ViLex model's pretraining phase. We hypothesize that this issue could potentially be mitigated by employing a higher-resolution T2I model as the base model.

Prompting a frozen T2I model with both visual and textual prompts. In the main paper, we have demonstrated that ViLex tokens can serve as a novel visual "language" for multimodal image generation. Unlike methods such as DreamBooth [58, 59] and textual inversion [18], which require: (1) learning specialized text tokens for specific instances, (2) gradient-based training for each individual image, and (3) the use of LORA adapters [29] to modify the model architecture, DreamBooth must be fine-tuned separately for each object (or each set of images corresponding to the same object). In contrast, ViLex enables several DreamBooth tasks like image re-contextualization, artistic rendition and accessorization, as illustrated in Figure A3, Figure 5 and Figure 6, by simply prompting a frozen T2I model with a combination of our visual prompts (*i.e.*, ViLex tokens) and natural language text prompts. This approach does not require changes to the architecture of a pretrained text-to-image generation model or any fine-tuning of the T2I model itself. All tasks are performed in a zero-shot and unsupervised manner.