Beyond Text: Frozen Large Language Models in Visual Signal Comprehension Supplementary Materials

Lei Zhu[1] Fangyun Wei[2]* Yanye Lu[1]
[1]Peking University [2]Microsoft Research Asia
zhulei@stu.pku.edu.cn fawe@microsoft.com yanye.lu@pku.edu.cn

A. More Implementation Details

Global Codebook Generation. To generate the global codebook, we introduce a two-phase process: (1) expanding the LLM vocabulary through the proposed vocabulary expansion technique (as shown in Figure 1); (2) applying a filtering strategy to further eliminate the entries with less semantic meaning.

We use $\mathcal{T}$ to represent the original LLM vocabulary and denote its size by N. To generate bigrams, for each $t \in \mathcal{T}$, we first input the concatenation of a text prefix (e.g., "a photo of") and t into the LLM. The LLM predicts the next word in an auto-regressive manner. We record the top-M predictions (where M is 1 by default) with the highest confidences, denoted as $\{t_1^*, \ldots, t_M^*\}$. The bigrams for each $t \in \mathcal{T}$ are represented by $\{[t, t_1^*], \ldots, [t, t_M^*]\}$. This process is repeated for all subwords in the LLM vocabulary. Ultimately, we collect a set of bigrams, denoted as $\mathcal{T}_{Bi}$, which has a size of $N \times M$. Similarly, we can build a trigram set $\mathcal{T}_{Tri}$ by feeding each bigram in $\mathcal{T}_{Bi}$ into the LLM for next-word prediction. The resulting $\mathcal{T}_{Tri}$ has a size of $N \times M \times M$. We use $\{\mathcal{T}, \mathcal{T}_{Bi}, \mathcal{T}_{Tri}\}$ to represent the expanded LLM vocabulary.

For the filtering process, we compute the CLIP similarities between each image in the training set and every entry in the expanded LLM vocabulary $\{\mathcal{T}, \mathcal{T}_{Bi}, \mathcal{T}_{Tri}\}$. We then record the top-5 entries with the highest similarity scores for each image. Finally, we aggregate these entries from all images to form the final expanded LLM vocabulary, which serves as our global codebook $\mathcal{T}_E$.

Encoder and Decoder Structures. Figure 2 details the implementation of our V2L Tokenizer's local encoder and decoder. Specifically, the local encoder shares the same basic structure as VQ-GAN [1], utilizing four residual blocks with channel dimensions [128, 256, 256, 512] to downsample the input image by a factor of 8. Similarly, our decoder mirrors the encoder's structure, employing four residual blocks with channel dimensions [512, 256, 256, 128] to upsample the image back to its original resolution. We integrate the information from global tokens into the decoding process through a cross-attention layer, which is added before the self-attention layer in the nonlocal block.

Vector Quantization Loss. The proposed V2L Tokenizer requires optimization of the encoder, the decoder and the projector. Thus, we follow VQ-VAE [4] and VQGAN [1] to implement our vector quantization loss, utilizing a straight-through gradient estimator for optimization:

$$\mathcal{L}_{vq} = ||\boldsymbol{X} - \hat{\boldsymbol{X}}||^2 + ||sg(\boldsymbol{F}) - \hat{\boldsymbol{F}}|| + \beta||sg(\hat{\boldsymbol{F}}) - \boldsymbol{F}||$$

where $sg(\cdot)$ denotes the stop-gradient operation. Note that our method involves a trainable projector to produce codebook embeddings. Thus, unlike LQAE [3] and SPAE [5], the second term in the above equation is also necessary. We set β to 0.3.

Tuning LLaMA-2 with the V2L Tokenizer. To enhance the image generation task, we propose to fine-tune an LLM model. This process begins with the V2L Tokenizer generating both global and local tokens for the training images. Subsequently, the global tokens are employed as a "text prefix". We then concatenate these global tokens with the local tokens and input them into the LLM. The auto-regression loss is applied only to the local tokens. Due to resource limitations, we fine-tune a 7B LLaMA-2 model using LoRA [2] on 12 randomly selected classes from ImageNet training dataset over 100K iterations using 32× NVIDIA V100 GPUs. LoRA weights are integrated into the query and key projection matrixes, with the hyper-parameter setting of $r = 4$, $\alpha = 32$. For optimization, we use Adam optimizer, starting with a learning rate of 3e^{-4}. This rate undergoes half-cycle cosine decay after a 5-epoch linear warm-up phase. Consequently, the tuned model is able to predict masked tokens in an auto-regressive manner. The predicted token map is input into the decoder of the V2L tokenizer to generate the reconstructed image, as demonstrated in Section 4.3 of our main paper.

B. More Ablation Studies

Vocabulary Expansion. We study the effectiveness of the proposed vocabulary expansion strategy on the 5-way-K-

*Corresponding author.

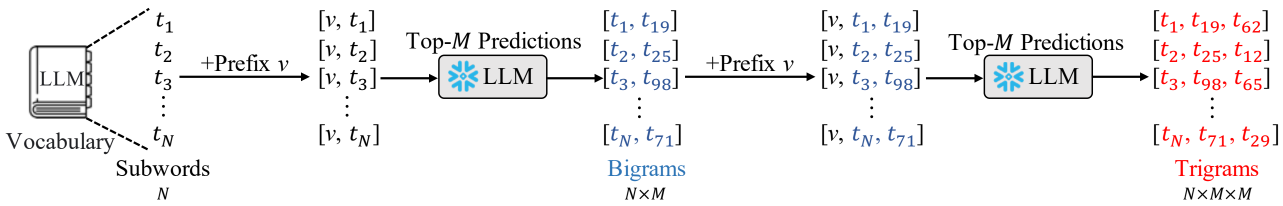


Figure 1. Illustration of the vocabulary expansion strategy. In this figure, we set $M = 1$ for illustrative purposes. The prefix v corresponds to the text phrase “a photo of”.

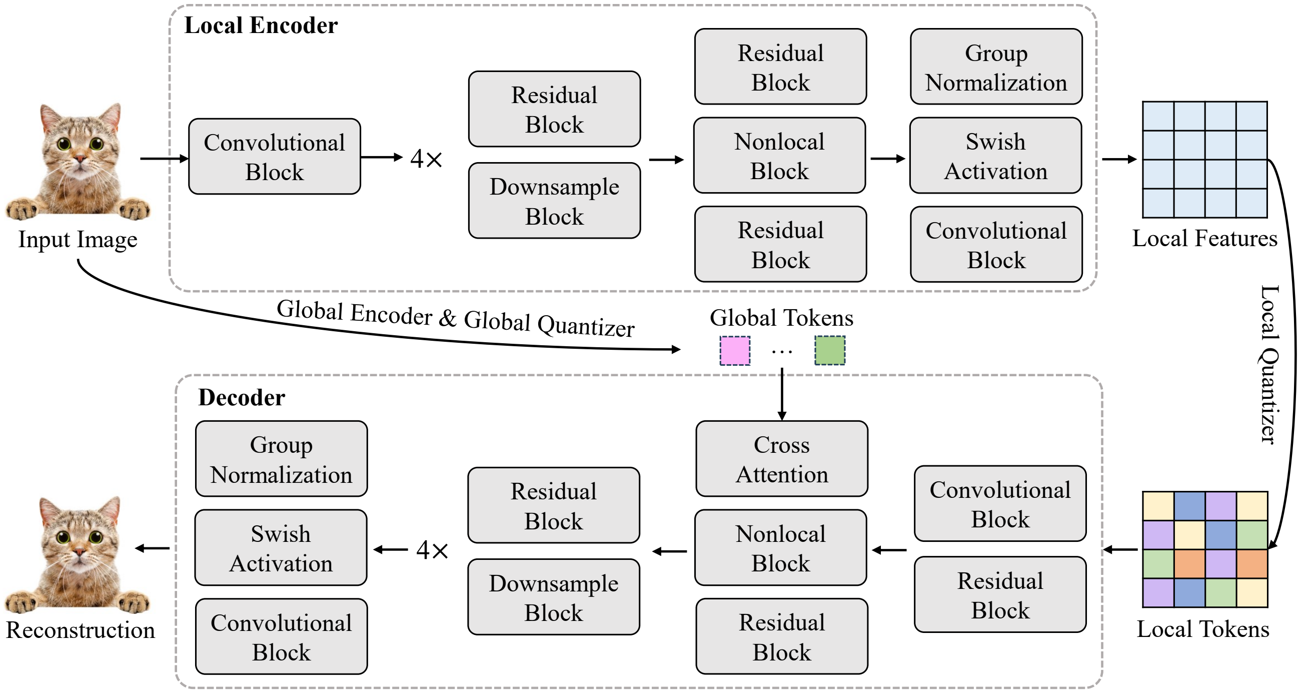


Figure 2. Illustration of the local encoder and the decoder of our V2L Tokenizer.

shot Mini-ImageNet classification benchmark. Our studies include three scenarios: utilizing the original LLM vocabulary without expansion (Subword), applying bigram expansion (Bigram), and employing trigram expansion (Trigram). The results of these scenarios are detailed in Table 1. The bigram expansion approach surpasses the non-expansion method by an average accuracy increase of +13.5 and +9.3 points with 5 and 21 global tokens, respectively. Implementing trigram expansion further elevates the average accuracy to 83.9 and 86.7. The findings demonstrate that employing vocabulary expansion significantly improves the semantic richness of the terms in the expanded LLM vocabulary, leading to enhanced classification accuracy.

Embeddings of Local Codebook. As shown in Figure 2 of the main paper, we introduce a trainable projector to project the LLM embeddings into a visual space, which enhances reconstruction quality. Table 2 presents our investigation of various LLM embeddings, including the default projected LLM embeddings (P-LLaMA-2), the original LLM embeddings (LLaMa-2), and those produced by the CLIP-text-encoder (CLIP). We observe that utilizing the CLIP text encoder for extracting language embeddings significantly boosts the quality of reconstruction. This improvement likely stems from the CLIP model’s inherent alignment between linguistic and visual spaces. By introducing a trainable projector, this alignment is further refined, leading to superior reconstruction performance.

Denoising Step and Condition Length. As shown in Figure 4 of the main paper, we denoise m masked tokens at a time using n tokens preceding them for the inpainting task, where m and n denote denoising step and condition length, respectively. We vary the values of m and n and report the FID scores for inpainting task in Figure 3. As the denoising step increases, the performance decreases. Additionally, an excessively long condition length leads to suboptimal performance since the LLM struggles to handle the complex context of a new “foreign language” in the visual modality.

		Task Induction:		✓	✓	✓	✓	✓	✓	
Method	#Tokens	Inner-shot:	1	1	3	5	1	1	1	Avg
		Repeats:	0	0	0	0	1	3	5	
Subword	5	LLaMA-2 (70B)	31.8	65.6	82.8	85.6	68.8	69.9	69.3	67.7
Bigram			40.6	83.1	91.7	92.6	86.5	87.0	86.9	81.2
Trigram			41.7	87.1	94.8	96.1	88.9	89.2	89.1	83.9
Subword	21	LLaMA-2 (70B)	34.3	74.1	90.1	91.8	79.6	80.2	80.7	75.8
Bigram			44.8	84.1	95.0	95.5	91.6	92.3	92.5	85.1
Trigram			46.5	89.1	96.9	97.8	91.4	92.7	92.9	86.7

Table 1. Ablation study for the proposed vocabulary expansion strategy on the 5-way-K-shot Mini-ImageNet classification benchmark.

Vocabulary	Embedding	FID↓	LPIPS↓	PSNR↑
LLaMA-2	LLaMA-2	9.51	0.17	21.48
LLaMA-2	CLIP	4.58	0.11	23.58
LLaMA-2	P-LLaMA-2	3.41	0.08	23.56

Table 2. Ablation study on various LLM embeddings. We report results on ImageNet-1K val set.

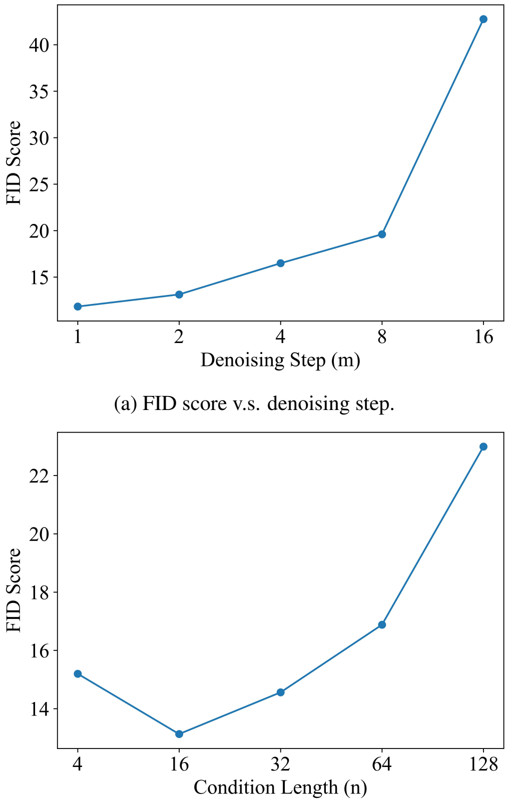
C. More Qualitative Results

Semantic Interpretation. We provide qualitative results for semantic interpretation in Figure 6 of the main paper. Here, we show additional visualizations in Figure 4.

Image Captioning and Visual Question Answering. Figure 5 of the main paper visualizes the results of image captioning and VQA. In Figures 5 and 6, we compare our approach with SAPE [5] using additional samples. Our model consistently generates more reasonable image captions and provides more accurate answers.

Image Reconstruction. In Table 3 of the main paper, we report the quantitative results for reconstruction evaluation. In this study, we show several qualitative visualizations. In Figure 7, we compare our approach with VQ-GAN [1], LQAE [3] and SPAE [5]. Our approach is notable for its ability to reconstruct images with a high level of detail.

Image Denoising. We show visualizations for image denoising in Figure 7 of the main paper. Here, we provide extra visualizations for inpainting (Figure 8), outpainting (Figure 9), deblurring (Figure 10), rotation restoration (Figure 11) and shift restoration (Figure 12).



(a) FID score v.s. denoising step.

(b) FID score v.s. condition length.

Figure 3. Ablation study on the denoising step (m) and the condition length (n) for the image inpainting task, using a 7B LLaMA-2.

References

[1] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1, 3

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[3] Hao Liu, Wilson Yan, and Pieter Abbeel. Language quantized autoencoders: Towards unsupervised text-image alignment. *arXiv preprint arXiv:2302.00902*, 2023. 1, 3

[4] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[5] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, et al. Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. *arXiv preprint arXiv:2306.17842*, 2023. 1, 3, 4

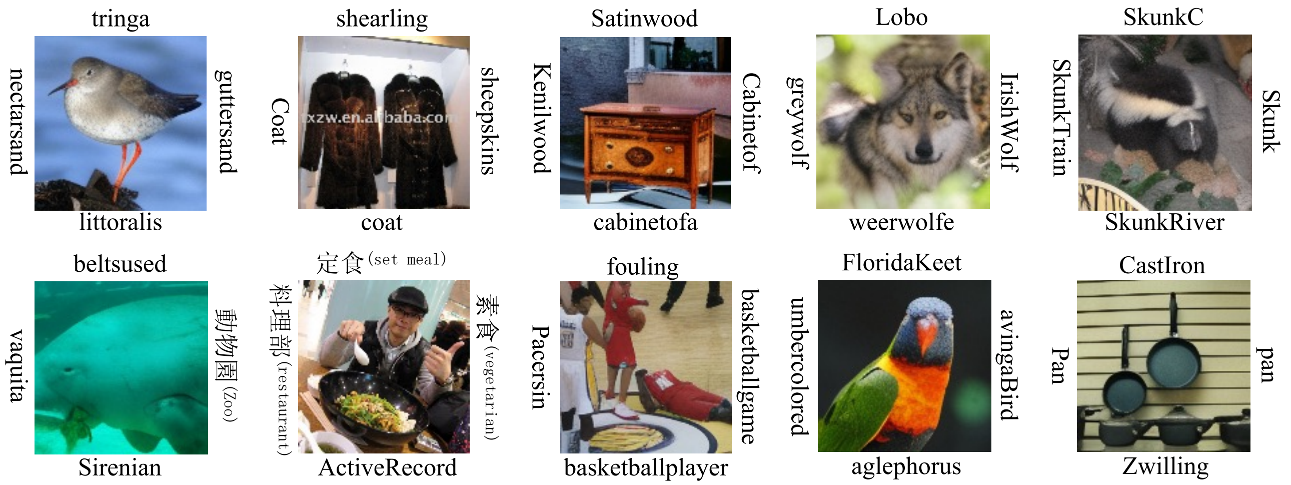
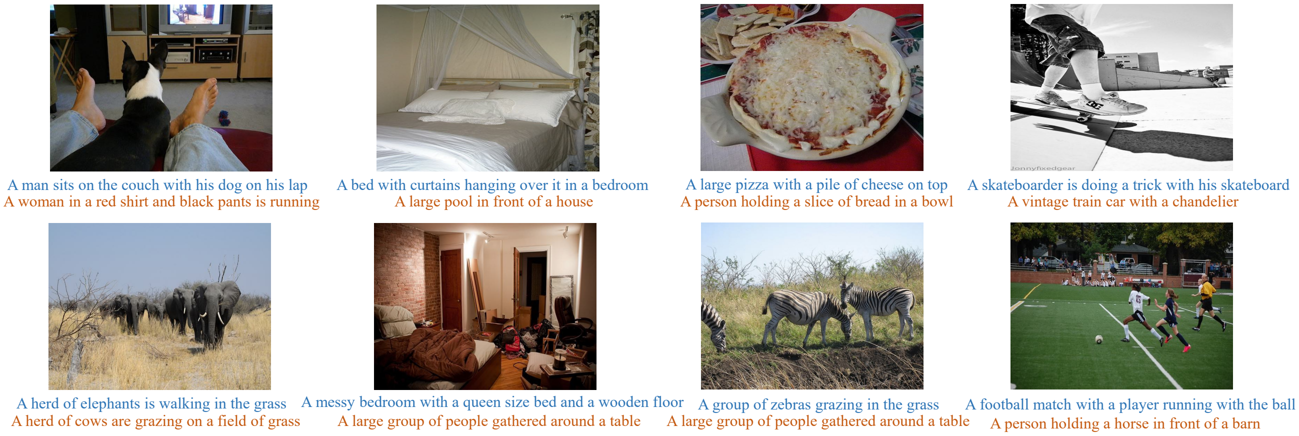


Figure 4. More visualizations for semantic interpretation.



A man sits on the couch with his dog on his lap
A woman in a red shirt and black pants is running

A bed with curtains hanging over it in a bedroom
A large pool in front of a house

A large pizza with a pile of cheese on top
A person holding a slice of bread in a bowl

A skateboarder is doing a trick with his skateboard
A vintage train car with a chandelier

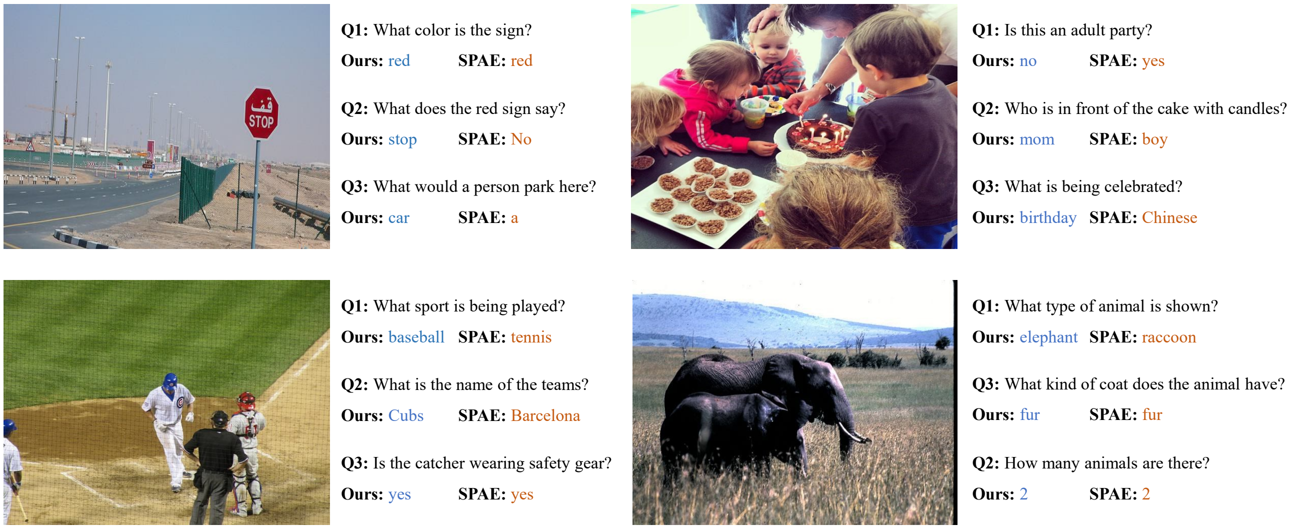
A herd of elephants is walking in the grass
A herd of cows are grazing on a field of grass

A messy bedroom with a queen size bed and a wooden floor
A large group of people gathered around a table

A group of zebras grazing in the grass
A large group of people gathered around a table

A football match with a player running with the ball
A person holding a horse in front of a barn

Figure 5. Visualizations for image caption. Blue: ours. Orange: SPAE [5] (re-implementation).



Q1: What color is the sign?
Ours: red **SPAE:** red

Q2: What does the red sign say?
Ours: stop **SPAE:** No

Q3: What would a person park here?
Ours: car **SPAE:** a

Q1: Is this an adult party?
Ours: no **SPAE:** yes

Q2: Who is in front of the cake with candles?
Ours: mom **SPAE:** boy

Q3: What is being celebrated?
Ours: birthday **SPAE:** Chinese

Q1: What sport is being played?
Ours: baseball **SPAE:** tennis

Q2: What is the name of the teams?
Ours: Cubs **SPAE:** Barcelona

Q3: Is the catcher wearing safety gear?
Ours: yes **SPAE:** yes

Q1: What type of animal is shown?
Ours: elephant **SPAE:** raccoon

Q3: What kind of coat does the animal have?
Ours: fur **SPAE:** fur

Q2: How many animals are there?
Ours: 2 **SPAE:** 2

Figure 6. Visualizations for visual question answering. Blue: ours. Orange: SPAE [5] (re-implementation).

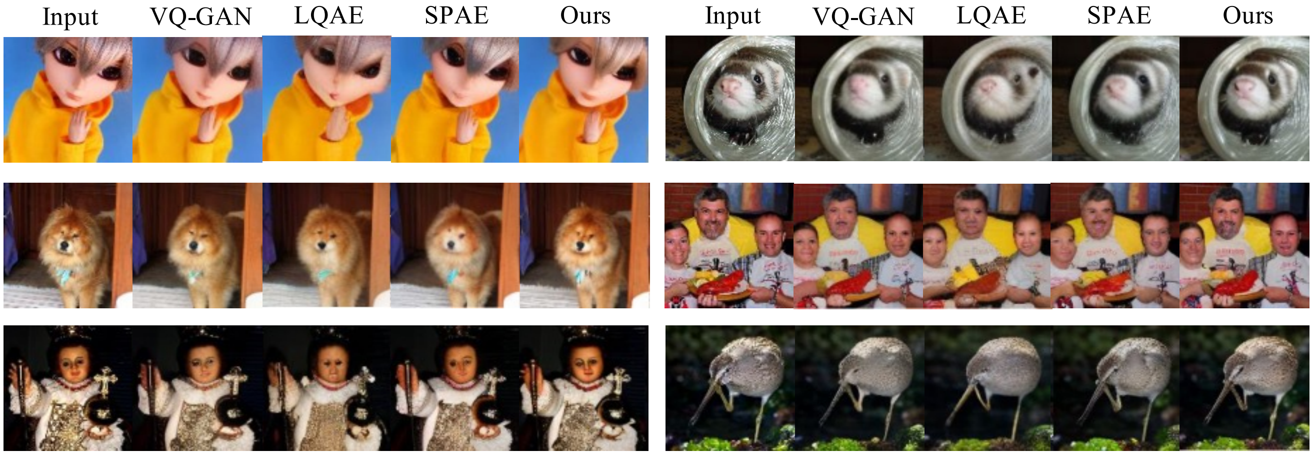


Figure 7. Visualizations for image reconstruction.

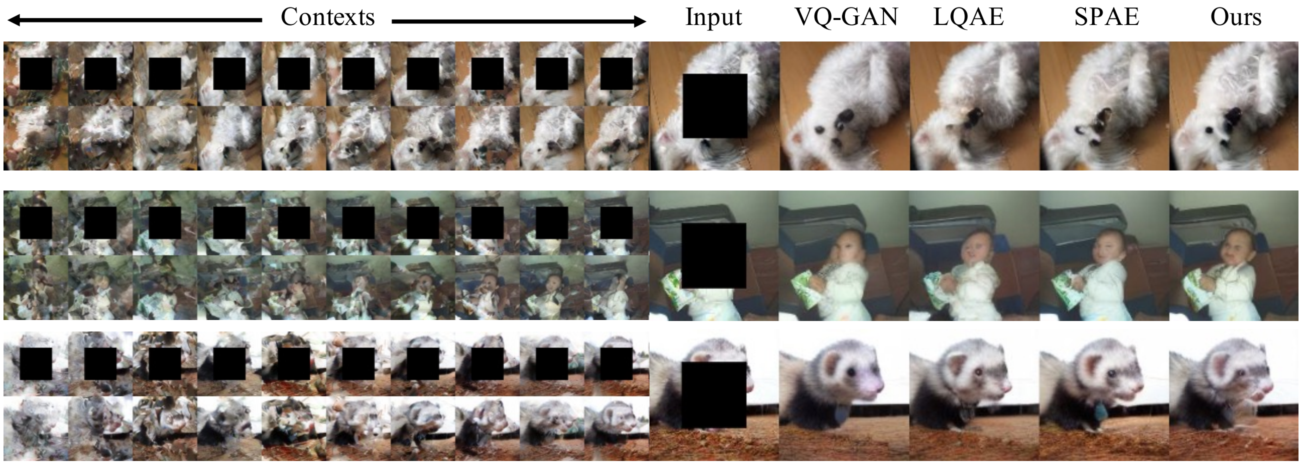


Figure 8. Visualizations for image inpainting.

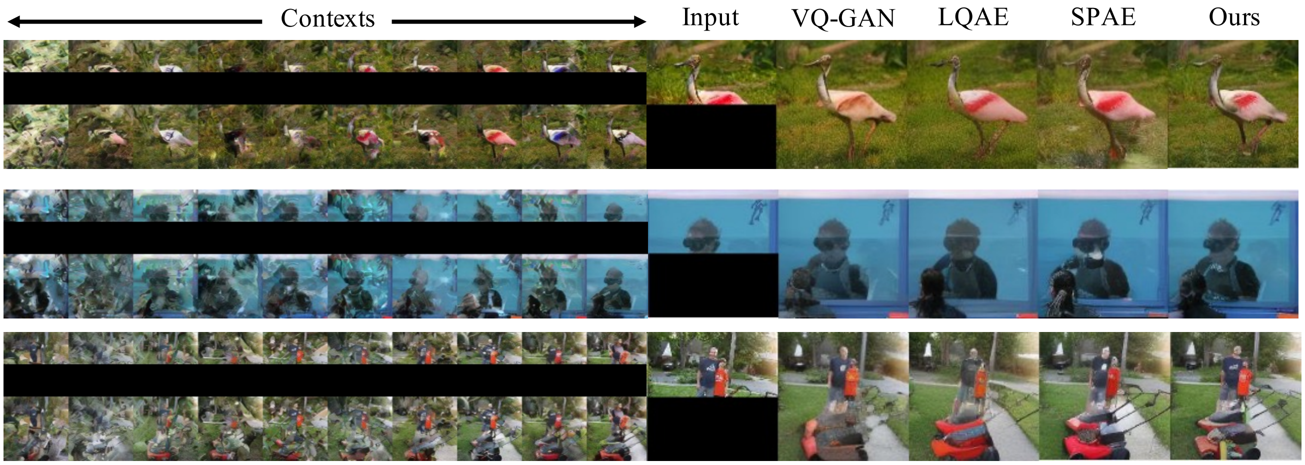


Figure 9. Visualizations for image outpainting.

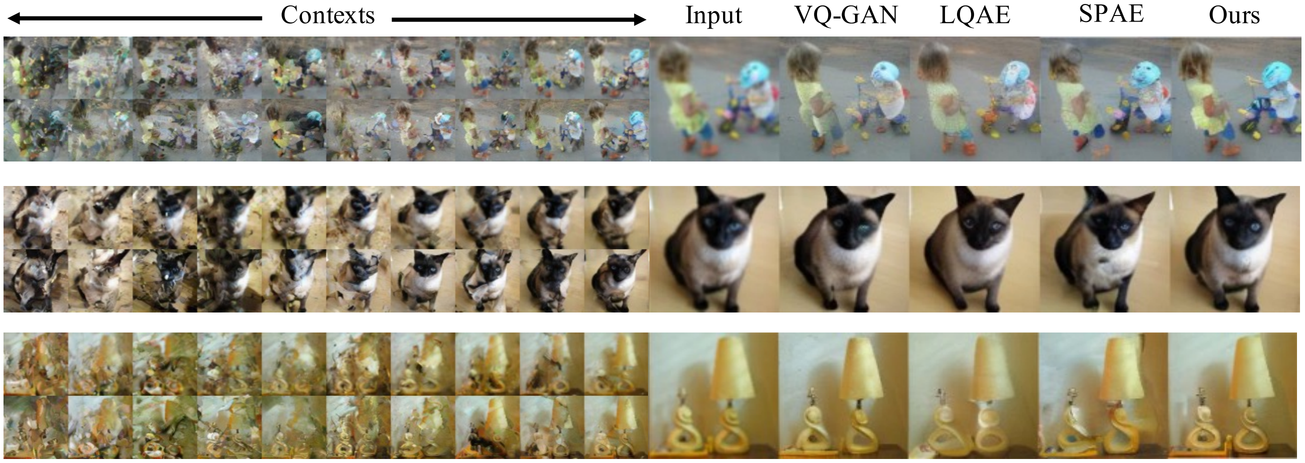


Figure 10. Visualizations for image deblurring.

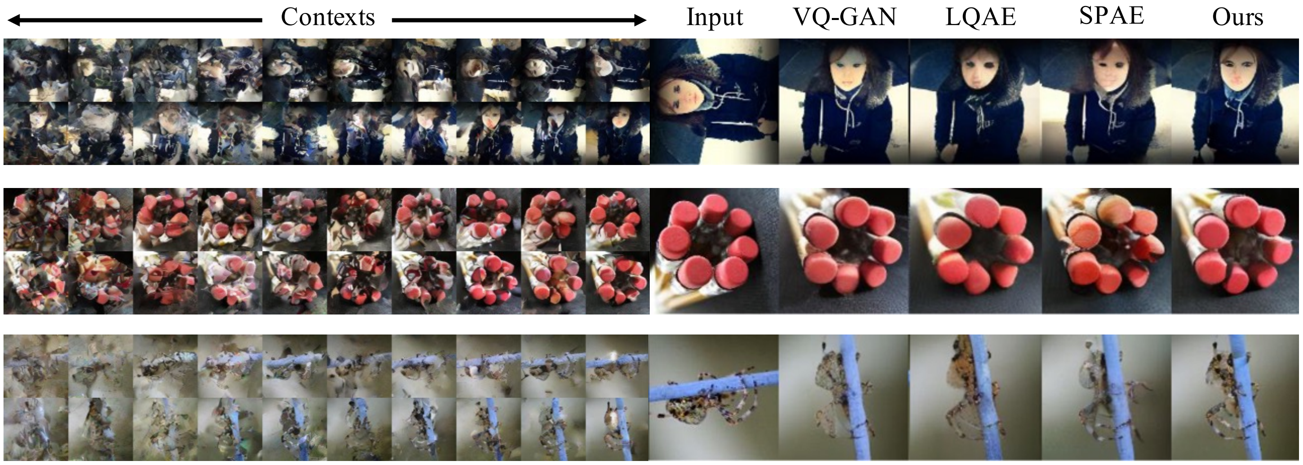


Figure 11. Visualizations for rotation restoration.

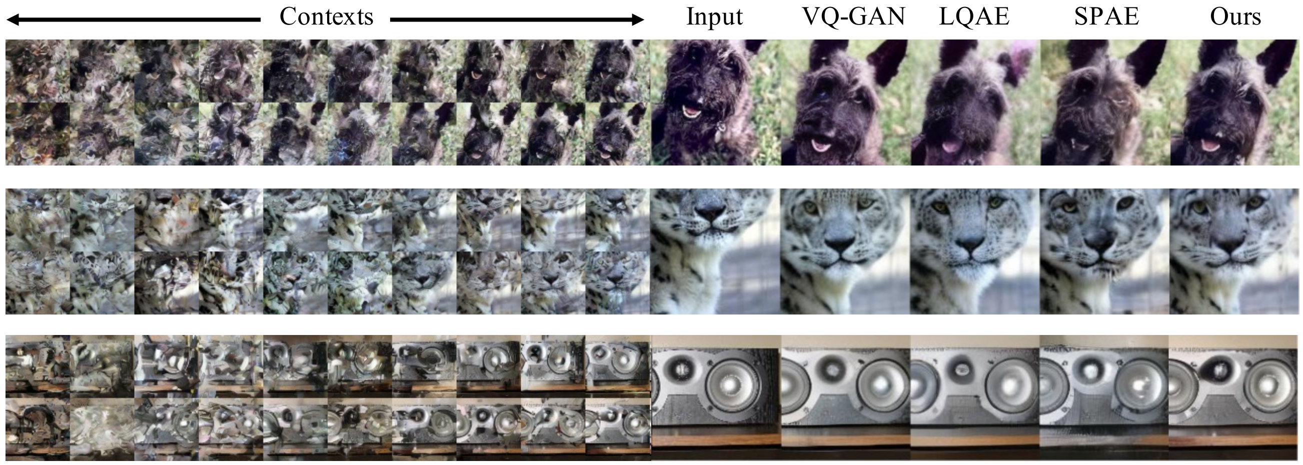


Figure 12. Visualizations for shift restoration.