A. Training Details

We apply BPE dropout with a rate of 0.1. We also apply residual and attention dropouts with a rate of 0.1, and label smoothing for both image and text loss computation with a rate of 0.1. We train both ARGVLT and MAGVLT models using AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.96$, $\epsilon = 10^{-8}$, weight decay coefficient of 4.5×10^{-2}, and the learning rate of 4.5×10^{-4} with a cosine annealing. The gradients are clipped by a norm using a threshold of 4, prior to applying the Adam update. When training ARGVLT, we observe that calculating the predictive losses on the context tokens along with the generation tokens improves the overall performance. Hence, we compute the losses on the whole concatenated token sequence with the loss coefficients to 0.9 and 0.1 for generation modality and conditional modality, respectively. The data augmentations used in [46] are applied to the images before encoding them using VQ-GAN. For positional embedding, we adopt a learnable absolute position encoding, for both image and text modalities. The encoded image tokens are flattened by the raster scan order before being fed into the transformer. MAGVLT was trained on 128 V100 GPUs for 40K updates with a batch size of 4,096, which takes about 3 days.

Table 8 describes the detailed architecture hyperparameters for the transformers we used including the large models.

Parameter	**Model**	
	ARG/MAGVLT	ARG/MAGVLT$_{\text{Large}}$
Params	371M	840M
Layers	24	36
Embed Dim	1024	1280
Heads	8	10

Table 8. Detailed architecture hyperparmeters. The left model column represents the default model described in the main paper, while the right column indicates the large model that will be presented in the next section.

B. Model Scaling

It is well known that scaling up the pretrained generative model generally improves the generalization ability, and recently VL models often have more than 1B parameters. Therefore, we also scale up our VLTs and evaluate those for the tasks of zero-shot I2T and T2I on MS-COCO. As shown in Table 8, the large model (MAGVLT$_{\text{Large}}$) contains 840M parameters for the transformer and 916M parameters including VQ-GAN in total. MAGVLT$_{\text{Large}}$ was trained on 128 V100 GPUs for 80K updates with a batch size of 4096, which takes about 12 days.

The zero-shot T2I results on MS-COCO are shown in Table 9. Notably, the large-scale models of both VLTs significantly improve FID and IS scores with large margin, compared to their respective default models. In addition, the degree of sampling speed reduction by model scaling is relatively smaller in MAGVLT than that in ARGVLT. Note that MAGVLT$_{\text{Large}}$ is slightly slower than the default MAGVLT (6.97× vs 8.12×), however it is still much faster than the default ARGVLT which has much fewer parameters.

Model	FID (↓)	IS (↑)	Speed
ARGVLT	16.93	22.50	1.00×
ARGVLT$_{\text{Large}}$	13.01	23.75	0.51×
MAGVLT	12.08	22.75	**8.12×**
MAGVLT$_{\text{Large}}$	**10.14**	**25.15**	6.97×

Table 9. *Zero-shot* T2I results on MS-COCO validation.

Model	CIDEr	SPICE
MS-COCO		
ARGVLT	45.5	11.2
ARGVLT$_{\text{Large}}$	43.6	11.2
MAGVLT	60.4	14.3
MAGVLT$_{\text{Large}}$	**68.1**	**15.5**
NoCaps		
ARGVLT	33.4	6.4
ARGVLT$_{\text{Large}}$	34.1	6.1
MAGVLT	46.3	8.7
MAGVLT$_{\text{Large}}$	**55.8**	**9.8**

Table 10. *Zero-shot* I2T results on MS-COCO Karpathy test (**Top**) and NoCaps validation (**Bottom**).

The zero-shot I2T results on MS-COCO and NoCaps datasets are presented in Table 10. Similar to the T2I results, the large-scale models of both VLTs show better I2T scores compared to their respective default models. Note that in case of ARGVLT, the performance gap between the default and large models is marginal on MS-COCO dataset, while MAGVLT improves the performance significantly on both datasets, as the model size is increased. These results imply that our MAGVLT is more effective in model scaling.

C. Finetuning on Downstream Tasks

In order to verify the transferability of MAGVLT by task-specific finetuning, we perform finetuning on two downstream tasks, one for generation and the other for understanding. In this finetuning setting, ARGVLT and MAGVLT are initialized from their 40K pretrained checkpoint, and MAGVLT$_{\text{Large}}$ is initialized from 60K pretrained checkpoint.

Image Captioning. We finetune ARGVLT and MAGVLT on the image caption generation task of MS-COCO 2014 dataset. In specific, we finetune the VLTs with the cross entropy loss for 100 epochs with a batch size of 512. The learning rate is set to 10^{-5} for ARGVLT and MAGVLT, and

2×10^{-5} for MAGVLT$_{\text{Large}}$. Note that we do not use the additional tasks, UnrollMask and MixSel, in finetuning. The captioning performances are presented in Table 11. Similar to zero-shot I2T results, MAGVLT shows better results compared to ARGVLT. Moreover, the large-scale model of MAGVLT improves the performances compared to its respective default model.

Model	B-4	M	C	S
ARGVLT	28.6	25.2	94.7	18.1
MAGVLT	29.3	27.1	103.3	20.5
MAGVLT$_{\text{Large}}$	**32.3**	**27.9**	**110.7**	**21.0**

Table 11. Comparisons of finetuned models on MS-COCO Karpathy splits.

Visual Question Answering. Masked pretraining is well known as a good representation learning approach for VL *understanding* tasks. Therefore, even though we use a variable mask ratio rather than a low fixed ratio during training for obtaining generation capability of MAGVLT, we can also evaluate the transferability of MAGVLT on a discriminative task. For this, we perform experiments on visual question answering (VQA) task, which is a VL understanding task that requires a model to answer a question given an image, on the commonly used VQAv2 dataset [23]. Following [64], we treat this task as a classification task where an auxiliary classifier predicts an answer from 3,129 candidates. The tokens of the question mark '?' and <MASK> token are sequentially added to the tail of the input sequence $[X; Y]$ where $[\cdot]$ is the concatenation operator. The top layer output of <MASK> is used as an input for the classifier. We finetune the classifier and the corresponding model with the cross entropy loss for 20 epochs with a batch size of 2,048 and a learning rate of 5×10^{-5}, and the dropout rate of the top layer output is set to 0.6.

The results are shown in Table 12. Compared to the latest algorithms [14, 25], MAGVLT performs slightly worse, however it can be confirmed that the discriminative representation for understanding has been learned by MAGVLT to some extent. While VLKD [14] and MetaLM [25] use large-scale language-only data and leverage a language model, we pretrain our model from scratch using only paired image-text datasets. And, our model is basically trained for generation, and moreover, it can even generate images by a single model.

Model	test-dev	test-std
VLKD$_{\text{ViT-B/16}}$ [14]	69.8	-
MetaLM [25]	**74.4**	**74.5**
MAGVLT	63.0	63.4
MAGVLT$_{\text{Large}}$	65.7	66.2

Table 12. Experimental results on VQAv2.

D. Unconditional Image+Text Generation Result

Since we train MAGVLT with the three multi-modal tasks including IT2IT, the model is able to produce both image and text at a time. Namely, all of the tokens of X and Y are masked at first, and then refined through the iterative decoding. For the target length prediction, the target length is randomly initialized in a range from 8 to 16 and then iteratively predicted as the refinement step proceeds. Here, we provide unconditional image+text generation results which are presented in Figure 8. Note that the generated images are very diverse and generally have high quality, and the generated texts also describe the images properly.

E. MixSel Analysis

Here, we demonstrate the effectiveness of the proposed *MixSel* task. As described in subsection 3.4, MixSel mixes two different contexts and selects one of them to be used for generation. We hypothesize that our MixSel training task allows the model to attend more carefully to the proper cross-modal context and accordingly to reduce the overlooking of the cross-modal context. In order to verify this, we first consider *MixRandom* setting which is the same as MixSel, but different in that the target is randomly selected without the additional special token to inform which one is selected, *i.e.*<LEFT> and <RIGHT> or <TOP> and <BOTTOM>. This MixRandom can be seen as the perturbation of the input context alone for regularization like data augmentations. In Table 13, MAGVLT$_{\text{MixRandom}}$, which indicates the trained MAGVLT along with UnrollMask and MixRandom, deteriorates the performances of both the zero-shot I2T and the zero-shot T2I, in comparison to MAGVLT with the use of MixSel training.

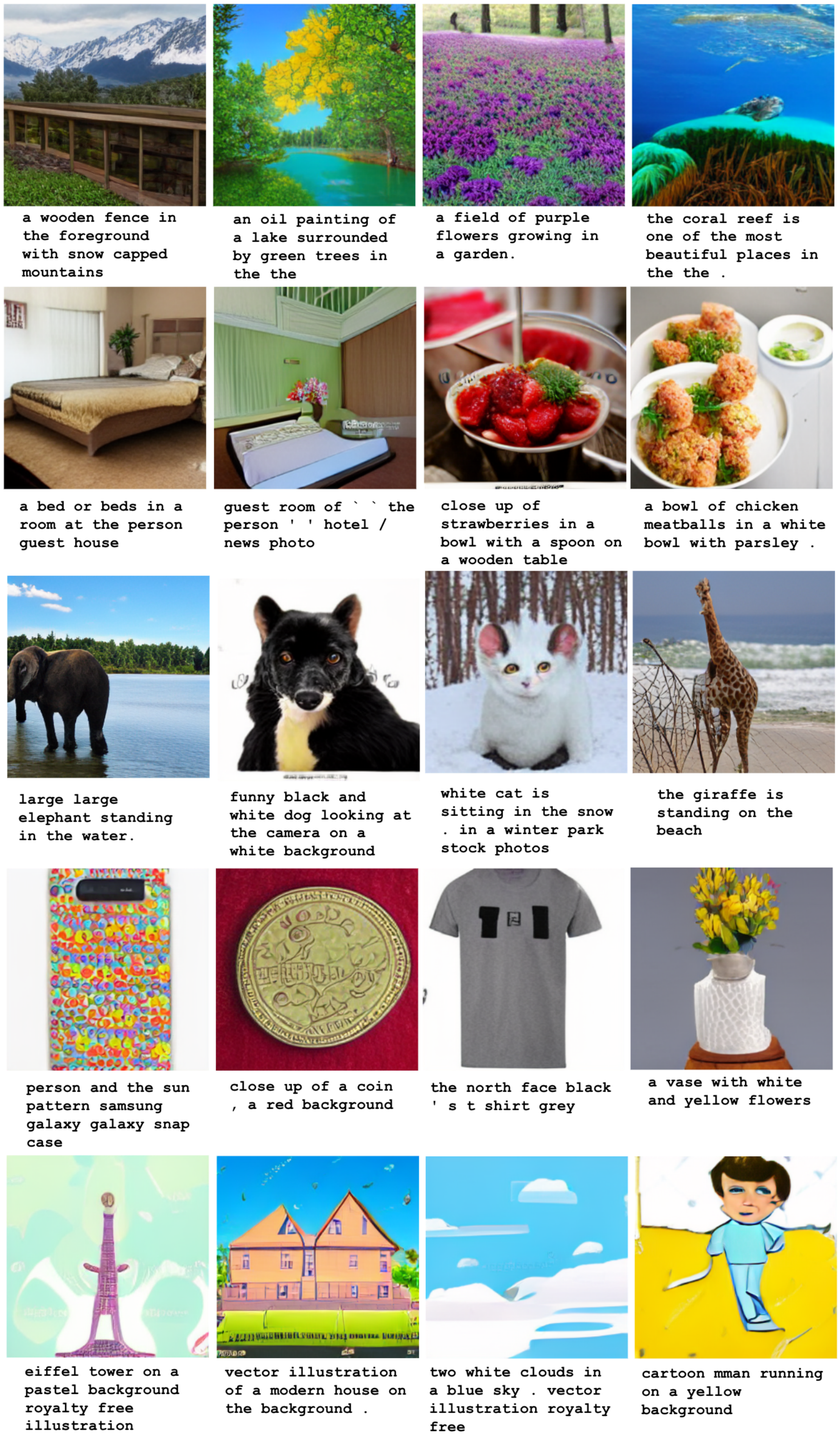
Model	CIDEr (↑)	FID (↓)
MAGVLT$_{\text{MixSel}}$	**60.4**	**12.08**
MAGVLT$_{\text{MixRandom}}$	57.9	13.43

Table 13. Comparison of MixSel and MixRandom on *Zero-shot* I2T and T2I.

Furthermore, in Figure 9, we qualitatively show by visualization of cross-modal attention maps that MixSel pretraining task makes the model to attend more to the cross-modal context appropriately compared to the model trained without MixSel training.

F. Additional Samples

Here, we present more qualitative results of image and text generation tasks described in subsection 4.2 and subsection 4.3. The image generation and inpainting results are presented in Figure 10, Figure 11, respectively. The image captioning and text infilling results are shown in Figure 12 and Figure 13, respectively. For text generation tasks, we resize and center-crop the validation images. Overall, our proposed MAGVLT shows better results than ARGVLT.



a wooden fence in the foreground with snow capped mountains

an oil painting of a lake surrounded by green trees in the the

a field of purple flowers growing in a garden.

the coral reef is one of the most beautiful places in the the .

a bed or beds in a room at the person guest house

guest room of ` ` the person ' ' hotel / news photo

close up of strawberries in a bowl with a spoon on a wooden table

a bowl of chicken meatballs in a white bowl with parsley .

large large elephant standing in the water.

funny black and white dog looking at the camera on a white background

white cat is sitting in the snow . in a winter park stock photos

the giraffe is standing on the beach

person and the sun pattern samsung galaxy galaxy snap case

close up of a coin , a red background

the north face black ' s t shirt grey

a vase with white and yellow flowers

eiffel tower on a pastel background royalty free illustration

vector illustration of a modern house on the background .

two white clouds in a blue sky . vector illustration royalty free

cartoon mman running on a yellow background

Figure 8. Unconditional image+text generation results obtained by MAGVLT. Note that the generated images cover diverse categories, such as natural scenery (1st row), indoor scenes & foods (2nd row), animals (3rd row), objects (4th row), and illustrations (5th row). Also, the generated texts are well aligned with generated images.

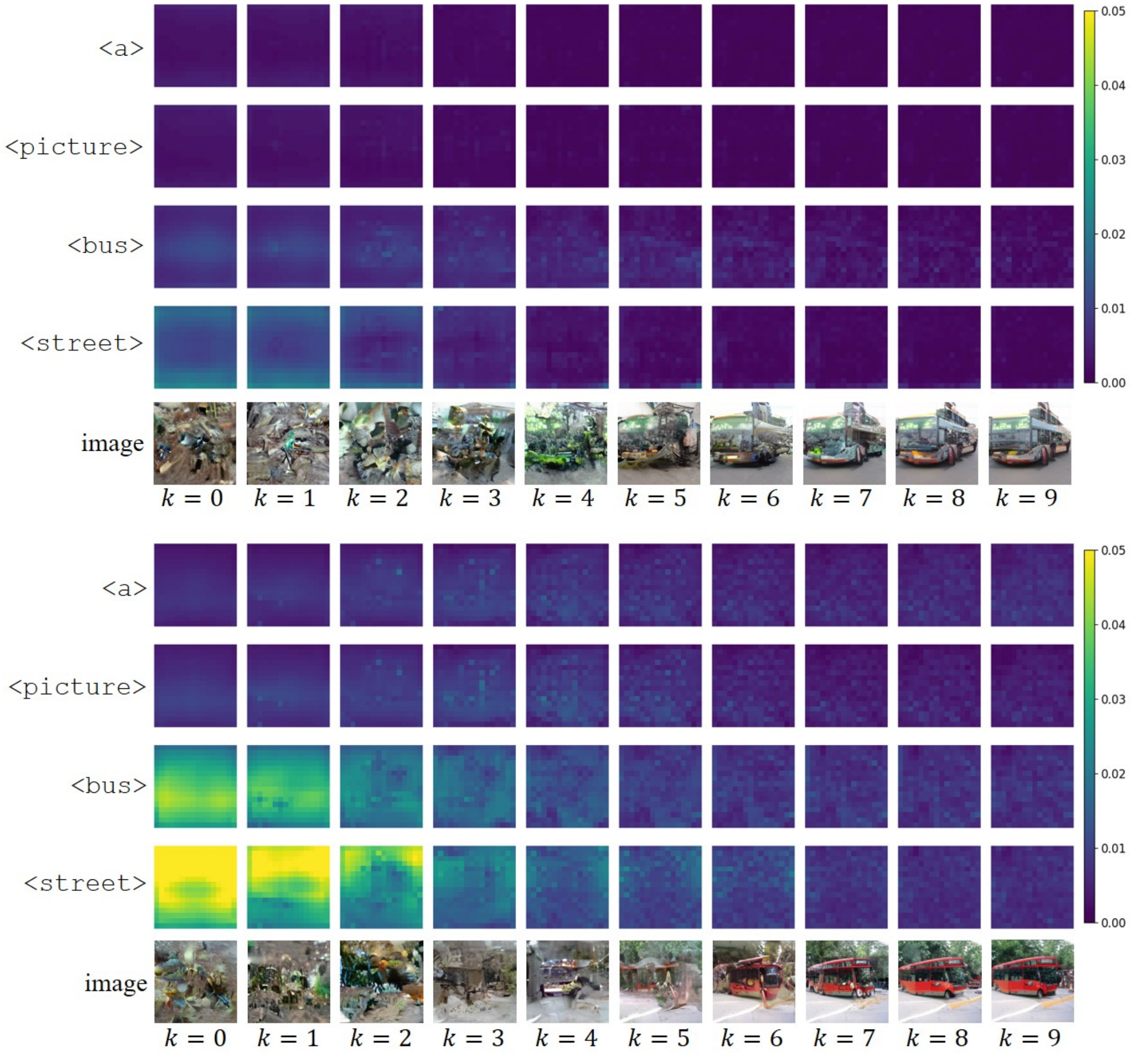


Figure 9. Visualization of cross-modal attention maps and generated images at different refinement steps. Given the text ”a picture of bus in the street.”, images are generated using MAGVLTs trained without the use of UnrollMask and MixSel (**Top**) and with the use of UnrollMask and MixSel (**Bottom**). To visualize each attention map, cross-attention scores between all 256 image tokens (queries) and a specific text token (a key, corresponds to each row) are computed and then reshaped to 16x16. Image tokens more attend to object text tokens (<bus> and <street>) when the model trained with the use of UnrollMask and MixSel.

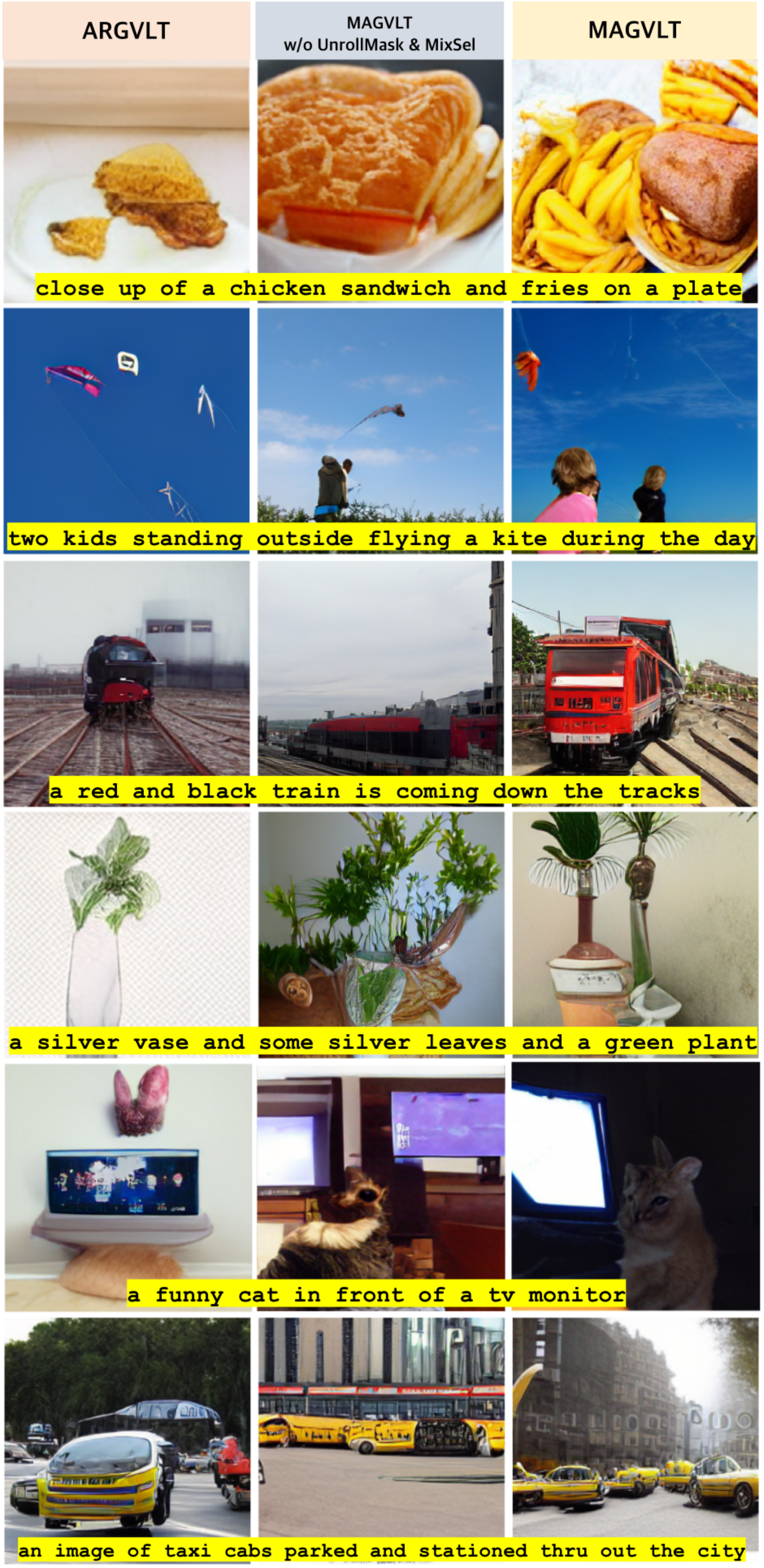

ARGVLT
MAGVLT w/o UnrollMask & MixSel
MAGVLT
close up of a chicken sandwich and fries on a plate
two kids standing outside flying a kite during the day
a red and black train is coming down the tracks
a silver vase and some silver leaves and a green plant
a funny cat in front of a tv monitor
an image of taxi cabs parked and stationed thru out the city

Figure 10. More samples of text to image on MS-COCO dataset.

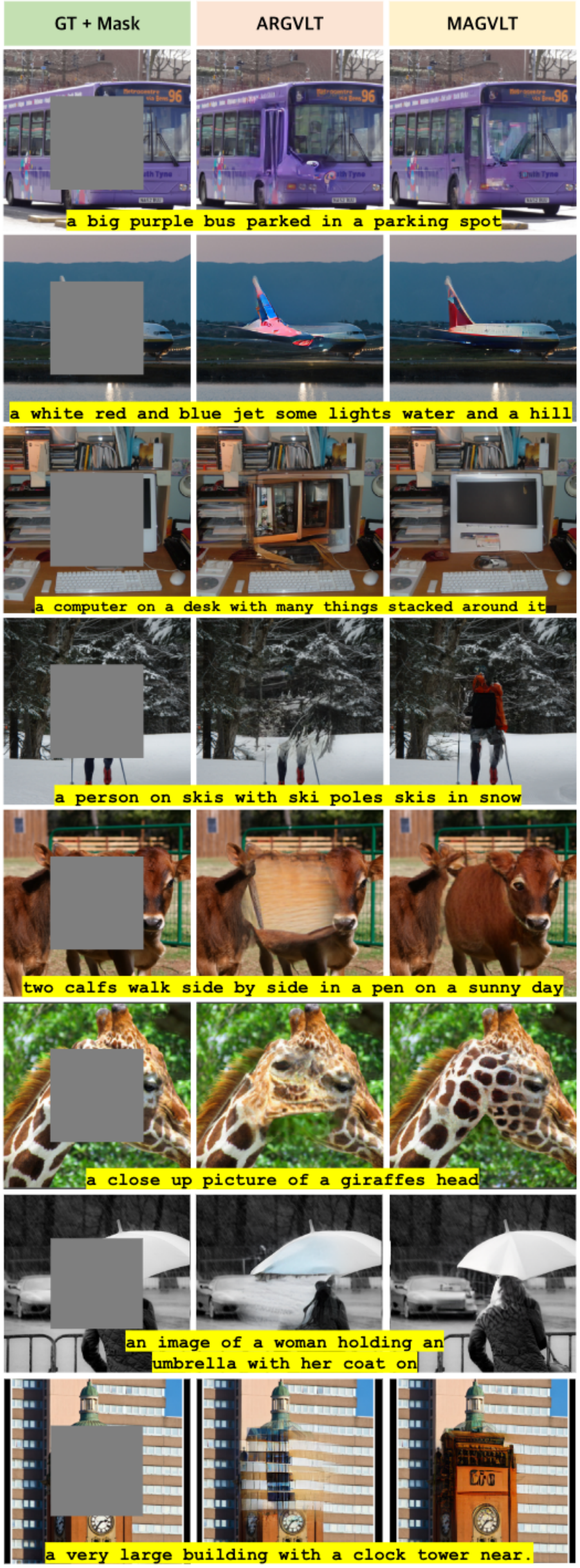

GT + Mask
ARGVLT
MAGVLT
a big purple bus parked in a parking spot
a white red and blue jet some lights water and a hill
a computer on a desk with many things stacked around it
a person on skis with ski poles skis in snow
two calfs walk side by side in a pen on a sunny day
a close up picture of a giraffes head
an image of a woman holding an umbrella with her coat on
a very large building with a clock tower near.

Figure 11. Image inpainting samples on MS-COCO dataset. MAGVLT generated the masked parts to be more blended with the surrounding context, and more proper to the captions.

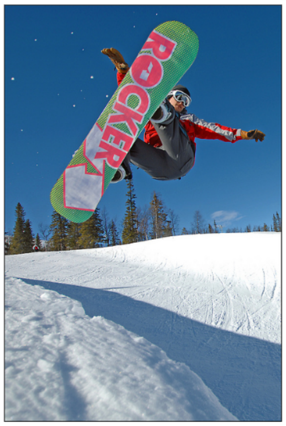
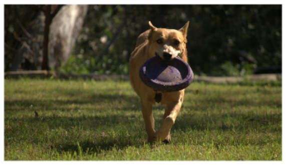
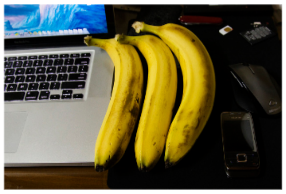
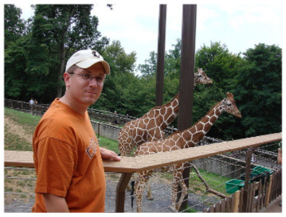
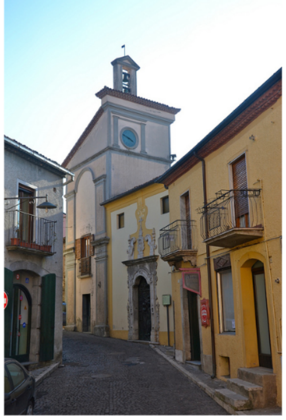
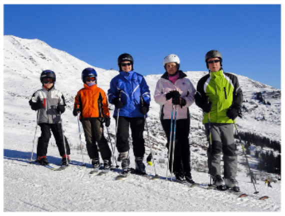
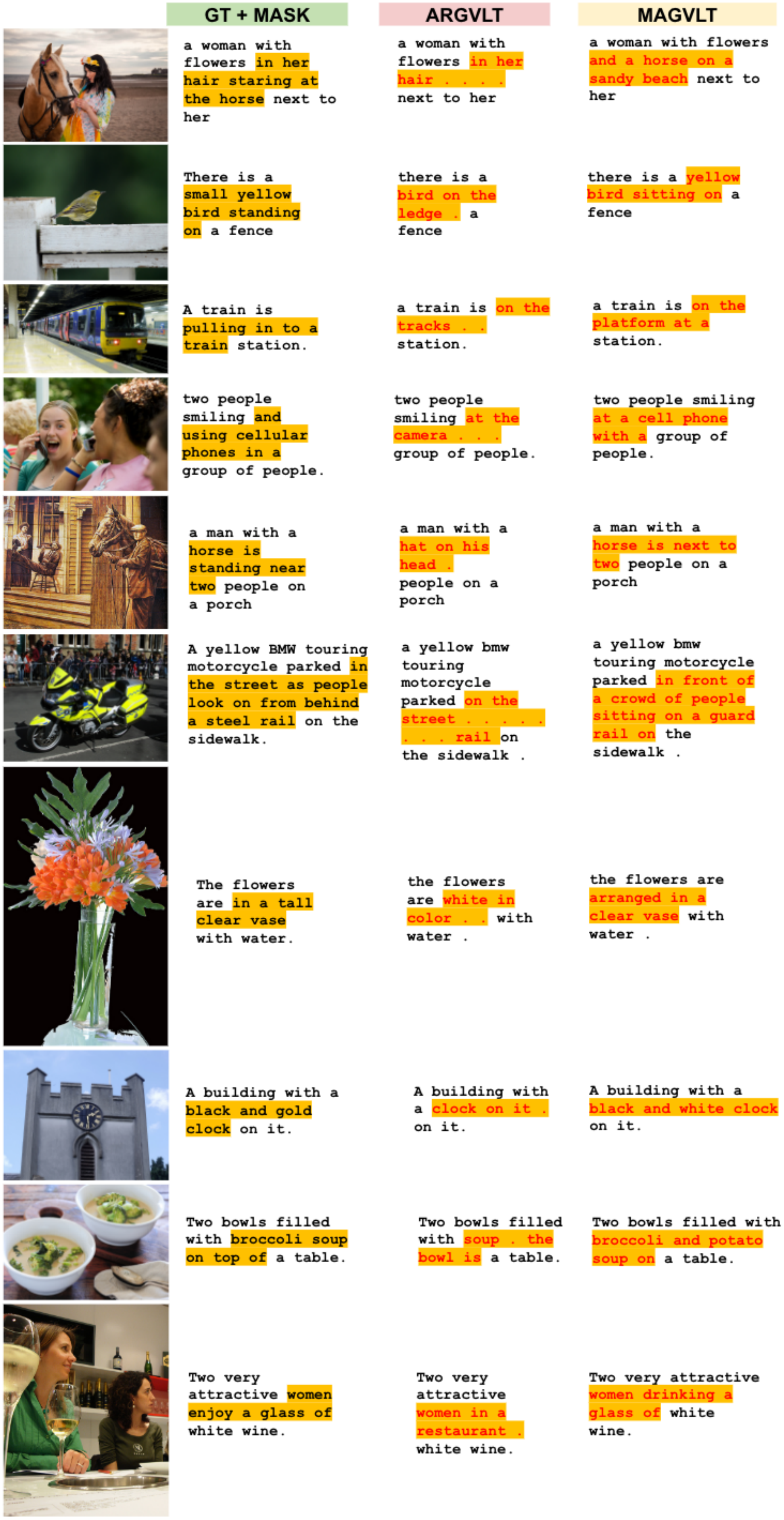
	GT	ARGVLT	MAGVLT
	A person launching into the air on a snowboard.	A person wearing a helmet.	A snowboarder jumping in the air.
	A very cute brown dog with a disc in its mouth.	The dog is brown.	The dog has a purple frisbee in its mouth.
	Three bananas that are sitting next to a laptop and cellphones.	A white cord	A bunch of bananas on the side of the laptop.
	Man posing in front of a pair of giraffes in background.	A man wearing a orange shirt.	A man in an orange shirt watching a a giraffe.
	A street with various buildings on each side and a clock tower.	A window on a building.	A typical street scene in a town in central italy , italy.
	A group of young and old are skiing on the snow.	A person wearing a helmet.	A group of skiers standing on a snowhill .

Figure 12. More samples of image captioning on MS-COCO dataset.



GT + MASK	ARGVLT	MAGVLT
a woman with flowers in her hair staring at the horse next to her	a woman with flowers in her hair . . . . next to her	a woman with flowers and a horse on a sandy beach next to her
There is a small yellow bird standing on a fence	there is a bird on the ledge . a fence	there is a yellow bird sitting on a fence
A train is pulling in to a train station.	a train is on the tracks . . station.	a train is on the platform at a station.
two people smiling and using cellular phones in a group of people.	two people smiling at the camera . . . group of people.	two people smiling at a cell phone with a group of people.
a man with a horse is standing near two people on a porch	a man with a hat on his head . people on a porch	a man with a horse is next to two people on a porch
A yellow BMW touring motorcycle parked in the street as people look on from behind a steel rail on the sidewalk.	a yellow bmw touring motorcycle parked on the street . . . . rail on the sidewalk .	a yellow bmw touring motorcycle parked in front of a crowd of people sitting on a guard rail on the sidewalk .
The flowers are in a tall clear vase with water.	the flowers are white in color . . with water .	the flowers are arranged in a clear vase with water .
A building with a black and gold clock on it.	A building with a clock on it . on it.	A building with a black and white clock on it.
Two bowls filled with broccoli soup on top of a table.	Two bowls filled with soup . the bowl is a table.	Two bowls filled with broccoli and potato soup on a table.
Two very attractive women enjoy a glass of white wine.	Two very attractive women in a restaurant . white wine.	Two very attractive women drinking a glass of white wine.

Figure 13. Text infilling samples on MS-COCO dataset. The locations to be infilled are shaded with orange color. The words infilled by MAGVLT are better aligned with the surrounding context words, and more appropriate on the corresponding images.