MAGVLT: Masked Generative Vision-and-Language Transformer

Sungwoong Kim⁠¹,†, Daejin Jo⁠²,*, Donghoon Lee⁠²,*, Jongmin Kim⁠²,*
¹Department of Artificial Intelligence, Korea University, Seoul, South Korea
²Kakao Brain, Seongnam, South Korea

swkim01@korea.ac.kr, {daejin.jo, dhlee, jmkim}@kakaobrain.com

Abstract

While generative modeling on multimodal image-text data has been actively developed with large-scale paired datasets, there have been limited attempts to generate both image and text data by a single model rather than a generation of one fixed modality conditioned on the other modality. In this paper, we explore a unified generative vision-and-language (VL) model that can produce both images and text sequences. Especially, we propose a generative VL transformer based on the non-autoregressive mask prediction, named MAGVLT, and compare it with an autoregressive generative VL transformer (ARGVLT). In comparison to ARGVLT, the proposed MAGVLT enables bidirectional context encoding, fast decoding by parallel token predictions in an iterative refinement, and extended editing capabilities such as image and text infilling. For rigorous training of our MAGVLT with image-text pairs from scratch, we combine the image-to-text, text-to-image, and joint image-and-text mask prediction tasks. Moreover, we devise two additional tasks based on the step-unrolled mask prediction and the selective prediction on the mixture of two image-text pairs. Experimental results on various downstream generation tasks of VL benchmarks show that our MAGVLT outperforms ARGVLT by a large margin even with significant inference speedup. Particularly, MAGVLT achieves competitive results on both zero-shot image-to-text and text-to-image generation tasks from MS-COCO by one moderate-sized model (fewer than 500M parameters) even without the use of monomodal data and networks.

1. Introduction

Generalizable multimodal modeling has recently made a lot of progress, especially in the field of vision-and-language (VL) modeling [3, 17, 30, 43, 45, 46, 48, 62, 63, 66, 67]. In particular, a massive amount of image-text data [9, 11, 51, 52, 54, 55] allows robust pretraining of large-scale multimodal VL models that can be easily transferred to various downstream tasks including image captioning [2, 38],

---

*Contributed equally.
†Work done at Kakao Brain. Corresponding author.
on the mask prediction has been proposed for language [22],
image [10, 36], and video processing [24, 59]. Masked
modeling is usually employed for representation learning to
solve understanding tasks in language, vision, and VL do-
 mains. However, with an iterative refinement-based gener-
ation and a variable mask ratio during training, it has been
shown to be used as a promising generative modeling. In
this regard, for our generative VL modeling, we propose
Masked Generative VL Transformer (MAGVLT). In con-
trast to AR-based generative VL transformer (ARGVLT),
the proposed MAGVLT is able to exploit bidirectional con-
ditioning and fast generation through a small number of re-
finement steps and parallel token predictions.

In specific, MAGVLT can generate any or both of an im-
age and a text sequence conditioned also on any or both
of them. Namely, it can perform any kind of task in a
form of image-and-text-to-image-and-text (IT2IT), includ-
ing image-to-text (I2T) and text-to-image (T2I) tasks. Fol-
lowing the previous masked generative modeling [10, 22],
we conduct sampling by iterative denoising based on the
masked token prediction and train MAGVLT by the masked
token prediction objective with a randomly sampled mask
ratio to take into account various denoising steps. Here, to
perform robust training of MAGVLT especially with only
image-text pairs from scratch, MAGVLT is learned basi-
cally by the composition of image-to-text, text-to-image,
and joint image-and-text mask prediction objectives. We
observe that our cross-modal masking (joint image-and-text
mask prediction) during training helps in improving both
performances of I2T and T2I tasks over single-modal mask-
ing (image-to-text + text-to-image mask predictions). Note
that only masked generative modeling used in MAGVLT
enables this cross-modal mask prediction during training.

In addition, we propose to use two additional tasks based
on the step-unrolled mask prediction and the selective pre-
diction on the mixture of two image-text pairs. The former
one is motivated by SUNDAE [50] and is modified to per-
form the mask prediction on the unrolled prediction, which
simulates the masked input samples encountered at the in-
termediate refinement steps. On the other hand, the latter
one learns to reconstruct the masked tokens in accordance
with a selected context between two VL contexts that are
mixed as a noisy input. This selective prediction improves
cross-modal attention for an accurate generation.

Through experiments on various downstream VL gener-
ation tasks, we empirically demonstrate that our MAGVLT
significantly outperforms ARGVLT even with greatly re-
duced inference time. Especially, to the best of our knowl-
dge, MAGVLT is the first model that obtains strong per-
formances on both zero-shot I2T and zero-shot T2I generation
tasks of MS-COCO benchmark [38] by a single moderate-
sized model (fewer than 500M parameters) without relying
on monomodal data and networks. Previously, as unified
generative VL models, L-Verse [32] and UPGen [4] have
not showed zero-shot I2T results while OFA [62] has used
monomodal data and also has not showed zero-shot I2T and
T2I results. Extensive ablations also validate the contribu-
tion of each component for MAGVLT.

To summarize, our main contributions are: (1) a masked
generative VL transformer as a unified generative VL model
that can produce both images and texts; (2) a robust train-
ing on image-text pairs by multiple training tasks that in-
clude the cross-modal mask prediction in tandem with the
step-unrolled mask prediction and the selective prediction
on the mixed context; and (3) an empirical validation of
MAGVLT that outperforms the autoregressive model and
moreover shows competitive performances on both of zero-
shot I2T and T2I generation tasks for the first time without
employing extra monomodal data and networks.

2. Related Work

2.1. Multimodal Vision-and-Language Modeling

There has been a lot of research on multimodal VL
training. Especially, in recent years, large-scale image-
text datasets have made great progress in VL pretraining
with various training objectives. For example, image-text
matching [39, 57, 65] and contrastive learning on image-text
data [30, 43, 60, 66, 68] has been widely used for discrimi-
native representation learning.

Since BERT [15] has shown impressive performances on
many natural language processing tasks, masked language
modeling has been widely adopted for VL pretraining. In
particular, [13] formulated multiple VL tasks as a text gen-
eration task and applied masked token prediction objective.
[6, 8, 21, 35, 63] have proposed to use a unified masked
data modeling with a shared multimodal transformer while
several algorithms have combined a number of objectives
including image-text matching, contrastive VL loss, and
masked language modeling [7, 14, 33, 71]. However, most
of these masked VL pretraining algorithms have been de-
veloped for VL understanding tasks.

Meanwhile, AR generative modeling has also recently
received lots of interest for VL pretraining due to its power-
ful generalization ability. While many algorithms have pro-
posed to utilize the generative language modeling paradigm
for VL understanding tasks [3, 25, 31, 61, 64], recent large-
scale AR transformers trained on a large amount of image-
text pairs have shown powerful performances for text-
guided image generation [16, 17, 46, 67]. A few generative
models based on AR decoding have been shown to produce
both images and text sequences [32, 62], however, they have
disproportionate competitive performances on both modalities.

2.2. Non-Autoregressive Generative Modeling

Non-AR generative models have been increasingly used
to lift certain limitations of AR models such as unidi-
rectional attention and slow decoding. Among them,
diffusion-based models have recently shown remarkable
performances on the task of text-guided image generation
3. Masked Generative Vision-and-Language Transformer

3.1. Masked Image-Text Modeling

MAGVLT is based on the previous masked generative modeling for image and language processing [10, 22]. Given an image-text pair \((I, T)\), the image input \(I\) is mapped to latent tokens \(X = [x_{ij}]_{i=1}^{N_I}\) by VQ-GAN [18], where \(N_I\) is the number of image tokens (e.g., 16 × 16), and the text sequence \(T\) is also converted to the tokenized sequence \(Y = [y_{ij}]_{j=1}^{N_T}\) by byte pair encoding (BPE) [53].

However, for text generation, diffusion models [5, 29] are still limited in achieving competitive performances compared to AR models.

Similar to diffusion-based generative models, masked generative models also perform iterative refinements for data generation and simulate various denoising steps during training. On top of that, masked generative models often conduct fewer refinement steps leading to faster generations. Therefore, masked generative models have recently been employed for language [22], image [10, 36], and video processing [24, 59]. However, there have been almost no masked generative models that can generate both text and image data. Very recently, a concurrent work [4] has tried to combine representation and generative learning for VL tasks into a single model that is based on masked token prediction. However, their generation performances are very poor on both modalities in contrast to the strong performances of MAGVLT on both modalities, and furthermore, our MAGVLT differs in that we use multiple cross-modal tasks for robust generative training on image-text pairs.

For training MAGVLT on image-text pairs, we first sample a mask ratio \(\gamma(r) \in (0, 1)\) where \(r \in [0, 1]\) indicates the simulated refinement step ratio and is also uniformly sampled considering various steps during generation. Then, we uniformly sample \([\gamma \cdot N]\) tokens where \(N = N_I + N_T\) and replace them with a special token \(<\text{MASK}>\). Here, we separately apply this masking for each modality such that \([\gamma_I \cdot N_I]\) image tokens and \([\gamma_T \cdot N_T]\) text tokens are masked. Let \(M_I = [m_{ij}]_{i=1}^{N_I}\) and \(M_T = [m_{ij}]_{j=1}^{N_T}\) be the resulting binary image mask and binary text mask, respectively, such that \(x_{ij}\) is replaced with \(<\text{MASK}>\) if \(m_{ij} = 1\) while \(y_{ij}\) is replaced with \(<\text{MASK}>\) if \(m_{ij} = 0\). Note that we set \(\gamma_I(\cdot)\) and \(\gamma_T(\cdot)\) as the cosine function and the linear function, respectively, following [10] and [22].

As shown in Figure 1, MAGVLT is trained basically by the composition of three mask prediction losses: \(\mathcal{L}_{IT2T}\) for the I2T task, \(\mathcal{L}_{T2I}\) for the T2I task, and \(\mathcal{L}_{IT2IT}\) for the IT2IT task. And these losses are defined by the negative log-likelihood of the masked tokens:

\[
\mathcal{L}_{IT2T} = -\mathbb{E}_{(X, Y) \in \mathcal{D}} \left[ \sum_{\forall j \in [1, N_T], m_{ij}^T = 1} \log p(y_j | Y_{\bar{M}_T}, X) \right],
\]

\[
\mathcal{L}_{T2I} = -\mathbb{E}_{(X, Y) \in \mathcal{D}} \left[ \sum_{\forall i \in [1, N_I], m_{ij}^I = 1} \log p(x_i | X_{\bar{M}_I}, Y) \right],
\]

\[
\mathcal{L}_{IT2IT} = -\mathbb{E}_{(X, Y) \in \mathcal{D}} \left[ \sum_{\forall j \in [1, N_T], m_{ij}^T = 1} \log p(y_j | Y_{\bar{M}_T}, X_{\bar{M}_I}) + \sum_{\forall i \in [1, N_I], m_{ij}^I = 1} \log p(x_i | X_{\bar{M}_I}, Y_{\bar{M}_T}) \right].
\]
the ground-truth length \( N \) predicted. The loss of the auxiliary target length predictor, on test time, the text sequence is generated after the length is

number of refinement steps is generally small (\( \lfloor \frac{K}{D} \rfloor \)). Thus, unmasked tokens again in the following iterations. Since the

image tokens in previous steps are excluded in computing

noted that following the masking strategy in [10], unmasked

again. This process is described in Figure 2. Here, it is

all the tokens in parallel. For the next \( k \)th iteration, the most

[\( \lfloor \frac{K}{N} \rfloor \)] unconfident tokens are masked out and predicted

This process is described in Figure 2. Here, it is

During inference, the target sequence is predicted by it-

erative decoding. The mask ratio is defined as a function

of the decoding steps as \( \gamma(\frac{k}{K}) \) \( \in \{0, 1, \ldots, K - 1\} \)

and \( K \) is the total number of iterations. For the first iteration

(\( k = 0 \)), all the tokens are masked, and the model predicts

all the tokens in parallel. For the next \( k \)th iteration, the most

[\( \frac{K}{N} \)] unconfident tokens are masked out and predicted

again. This process is described in Figure 2. Here, it is

noted that following the masking strategy in [10], unmasked

image tokens in previous steps are excluded in computing

confidences and accordingly will never be masked again.

On the other hand, unmasked text tokens can be selected as

masked tokens again in the following iterations. Since the

number of refinement steps is generally small (e.g., 10), this

iterative decoding with parallelizable predictions is signif-

icantly faster than the autoregressive decoding, especially

when the number of tokens is very large.

**Target Length Prediction.** Since the length of text se-
quence is varied in contrast to the fixed number of image to-

kens, non-AR models like MAGVLT need to perform target

length prediction for text generation. We follow the length

decor proposed in [22] where the output of `<BOT>` that

is located between \( X \) and \( Y \) is the predicted text length. At

test time, the text sequence is generated after the length is

predicted. The loss of the auxiliary target length predictor,

\( \mathcal{L}_{TL}(N_T, \hat{N}_T) \), is defined by the cross entropy loss between

the ground-truth length \( N_T \) and the predicted length \( \hat{N}_T \)
given the maximum possible number of text tokens.

**3.3. Step-Unrolled Mask Prediction**

Although a variable mask ratio during training reflects

various intermediate refinement steps, there still exists a gap

between a corruption on the target tokens at training time

and a corruption on the partially predicted tokens at test

time. SUNDAE [50] tries to resolve this issue by optimiz-

ing the model conditioned on a corrupted target sequence

which is sampled through one step generative unrolling dur-

ing training and achieves significant performance improve-

ments of the non-AR autoencoder for text generation.

Here, we adopt and modify this step-unrolled denois-

ing as an additional training task for MAGVLT. Since

MAGVLT is based on the masked token prediction, we re-

mask the one-step predicted sequence where the mask ratio

is reduced from the previous mask ratio and then predict the

re-masked tokens by MAGVLT. We call this task as step-

unrolled mask prediction, dubbed UnrollMask. We apply

UnrollMask only to the I2T and T2I training tasks to main-

tain the uncorrupted cross-modal context. Figure 3 visually

depicts this UnrollMask especially for the I2T task. We de-

ote the UnrollMask loss as \( \mathcal{L}_{UM} \), and for example \( \mathcal{L}_{UM,I2T} \)

on the I2T task can be defined as

\[
\mathcal{L}_{UM,I2T} = - \mathbb{E}_{(X,Y) \in \mathcal{D}} \left[ \sum_{y \in \{1, N_T\}, M_T^{(1)} = 1} \log p(y|\hat{Y}_M^{(1)}(X), X) \right], \tag{4}
\]

where \( \hat{Y}_M^{(1)}(X) \) indicates the re-masked one-step unrolled

prediction of \( Y_M^{(1)} \).

**3.4. Selective Prediction on Mixed Context**

In this multimodal generative modeling, the model often

ignores the cross-modal context and produces an output that

is biased to the within-modal statistics. For example, in the

I2T task of Figure 4, the model should predict the masked

word token as ‘cat’ by the given image, however, the model

often rather outputs ‘dog’ since ‘dog jumps’ is more likely

occurred than ‘cat jumps’ before the text of ‘on grass field’

in the set of training text sequences.

Thus, in order to reduce such bias, we propose a simple

yet effective additional learning task, named selective pre-

diction on the mixed context (MixSel), which is described
in Figure 5. As shown in the figure, two different input contexts are mixed in a half-and-half concatenated manner, and one of them is randomly selected to be the target context in generation. Here, a special token is appended to inform the selected context, for instance <LEFT> or <RIGHT> is used for the horizontally combined image or the concatenated text sequence while <TOP> or <BOTTOM> is used for the vertically combined image. Also, when two different text sequences are concatenated, another special token, <SEP>, is inserted between them. The MixSel objective is denoted as \( \mathcal{L}_{MS} \), and for instance \( \mathcal{L}_{MS} \) on the I2T task can then be defined as

\[
\mathcal{L}_{MS, I2T} = \mathbb{E}_{(X, Y) \in D} \left[ \log p(y_{\ell}^{f} | \hat{Y}_{\tau}^{L}, \phi(X^1, X^2)) \right],
\]

where \( \phi \) is the mixture function on the two images \( X^1 \) and \( X^2 \), and \( \ell \in \{1, 2\} \) represents the selected context.

From this MixSel training task, the model is able to attend more carefully to the appropriate span of the cross-modal context and improve the accuracy of the cross-modal attention by mixing the original cross-modal content with randomly unrelated one. This could make the model to tend more carefully to the appropriate span of the cross-modal content and circumventing the within-modal bias problem in test-time generation. Note that it is different from the previous mix-based data augmentation techniques [26, 69, 70] in that we retain the information of the original contexts entirely and randomly select the target one for generation. Moreover, although MixSel training is relevant to classifier-free guidance (CFG) [28] in that both try to strengthen the effect of the condition, MixSel does not have to perform the forward processing twice at test time. Also, we can adapt CFG along with MixSel training.

### 3.5. Multitask Pretraining

As we mentioned above, MAGVLT is basically trained via three types of multimodal tasks: I2T, T2I, and IT2IT. During training, a task \( \tau \in \{I2T, T2I, IT2IT\} \) is sampled from the categorical distribution with the predefined sampling probability \( p_{\tau} \) for each iteration (batch-wise), and then apply the associated mask prediction loss \( \mathcal{L}_{mask, \tau} \in \{\mathcal{L}_{I2T}, \mathcal{L}_{T2I}, \mathcal{L}_{IT2IT}\} \). Along with this mask prediction loss, we also add the target length prediction loss \( \mathcal{L}_{TL, \tau} \), the UnrollMask loss \( \mathcal{L}_{UM, \tau} \), and the MixSel loss \( \mathcal{L}_{MS, \tau} \) according to the selected task \( \tau \). Overall, the final objective of MAGVLT with respect to \( \tau \) is

\[
\mathcal{L}_{\tau} = \mathcal{L}_{mask, \tau} + \lambda_{TL} \mathcal{L}_{TL, \tau} \cdot \mathbb{I}[^{\tau} \neq T2I] + \lambda_{UM} \mathcal{L}_{UM, \tau} \cdot \mathbb{I}[^{\tau} \neq IT2IT] + \lambda_{MS} \mathcal{L}_{MS, \tau},
\]

where \( \mathbb{I}[^{\tau}] \) is the indicator function, and \( \lambda_{TL} \), \( \lambda_{UM} \), and \( \lambda_{MS} \) are relative loss weights. Here, we fix \( \lambda_{TL} = 0.01 \), \( \lambda_{UM} = 1.0 \), and \( \lambda_{MS} = 0.5 \) for all our experiments.

### 4. Experiment

In this section, we elaborate the experiments on the VL generation tasks with extensive ablation studies to identify the contribution of each factor in the proposed algorithm. Official codes will be available\(^{†}\).

#### 4.1. Experimental Setup

**Model.** VLTs (i.e. ARGVLT and MAGVLT) have 447M parameters (24 layers, 1024 hidden dimension, and 8 attention heads) including VQ-GAN in total. We also perform experiments about scaling of VLTs, and the results are presented in Appendix. As an image encoder, VQ-GAN [18] converts a 256 \( \times \) 256 image into 16 \( \times \) 16 tokens with 16,384 codebook size. For text sequence, we adopt the BPE tokenizer [53] used in CLIP [43] with 49,408 vocabulary size. We fix the text sequence length to 64.

**Dataset.** We pretrain ARGVLT and MAGVLT from scratch using paired image-text datasets. Our pretraining data consists of Conceptual Captions 3M (CC3M) [54], Conceptual Captions 12M (CC12M) [11], SBU Caption [41], and Visual Genome [34] datasets. Together, there are about 17M image-text pairs.

**Pretraining.** There are many options to train VLTs. Note that T2I and I2T are available for both ARGVLT and

\[\text{https://github.com/kakaobrain/magvlt}^{†}\]
Figure 6. Text-to-Image samples on MS-COCO captions. The images in the second column are sampled from MAGVLT trained without UnrollMask and MixSel. MAGVLT generated more appropriate images on the corresponding caption. More samples will be found in Appendix.

MAGVLT while IT2IT is only available for MAGVLT. We experiment various subsets of the multimodal tasks in order to investigate the effectiveness of each task. In specific, we pretrain ARGVLT on three different subsets which consist of T2I only, I2T only, and T2I & I2T. Likewise, we pretrain MAGVLT on three different subsets which consist of T2I only, I2T only, and T2I & I2T & IT2IT. More details of pre-training will be found in Appendix.

Evaluation. We compare generative VL models based on cross-modality generation tasks especially in zero-shot settings in order to evaluate the generalization ability of the proposed method. We also provide more results including finetuning VLTs on downstream tasks in Appendix.

Sampling. For text-to-image generation, following [32], we obtain 32 samples from each trained VLT (i.e. ARGVLT and MAGVLT) and calculate similarity scores between the sampled images and the conditioning text by CLIP [43] to select a top ranked image (clip reranking). Likewise, for image-to-text generation (image captioning), we produce 64 samples from each trained VLT and select a top ranked text by CLIP scores. The number of refinement steps is set to \( K = 10 \) and \( K = 12 \) for image generation tasks and text generation tasks, respectively.

4.2. Image Generation

Text-to-Image. We evaluate the zero-shot generalization capability of MAGVLT under text-to-image generation. We measure quantitative metrics of quality of generated images by Fréchet Inception Distance (FID) [27] and Inception Score (IS) [49] on MS-COCO [38] validation. In addition, we compare the relative decoding speed of MAGVLT against ARGVLT. The results are shown in Table 1, where the models are grouped according to: (1) whether they are AR-based or not, and (2) the modality they can generate; the models in the first two groups are able to generate image only while the models in the last group can generate both image and text. MAGVLTs significantly outperform AR-based methods including ARGVLTs as well as obtain comparable scores to the state-of-the-art diffusion-based methods. Note that all the other models in the non-AR group have more than 1B parameters while MAGVLTs have less than 500M parameters. And the performance gap between the task-specific MAGVLT (T2I only) and the universal MAGVLT is small. Moreover, MAGVLTs generate an image more than eight times faster than ARGVLTs. We provide qualitative samples in Figure 6.

Image Inpainting. One of the key advantages of MAGVLT over ARGVLT is that it enables bidirectional encoding of conditional information. MaskGIT [10] already demonstrated this advantage. To reconfirm it on MAGVLT, we conduct similar image inpainting experiments. In detail, the central 8×8 image tokens corresponding to the central 50% of the whole 16×16 image tokens are masked out, and then replaced with newly-generated tokens conditioning on the unmasked image tokens and the ground-truth text tokens. The output images are blended with the input images along the mask boundary following [10]. Quantitatively, as shown in Table 2, MAGVLT outperforms ARGVLT, which is also observed in qualitative samples of Figure 11 in Appendix.

### Table 1. Zero-shot T2I results on MS-COCO validation set. Here, we compute FID and IS on a subset of 30,000 captions sampled from COCO validation.

<table>
<thead>
<tr>
<th>Model</th>
<th>FID (↓)</th>
<th>IS (↑)</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CM3-Medium (2.7B) [1]</td>
<td>36.78</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DALL-E (12B) [46]</td>
<td>27.5</td>
<td>17.9</td>
<td>-</td>
</tr>
<tr>
<td>CogView (4B) [16]</td>
<td>27.1</td>
<td>18.2</td>
<td>-</td>
</tr>
<tr>
<td>CogView2 (6B) [17]</td>
<td>24.0</td>
<td>22.4</td>
<td>-</td>
</tr>
<tr>
<td>Parti-350M (350M) [67]</td>
<td>14.10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Make-A-Scene (4B) [20]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARGVLT (T2I only) (447M)</td>
<td>21.80</td>
<td>19.27</td>
<td>1.00×</td>
</tr>
<tr>
<td>MAGVLT (T2I only) (447M)</td>
<td>10.74</td>
<td>23.94</td>
<td>8.12×</td>
</tr>
</tbody>
</table>

| Non-AR based           |         |        |       |
| GLIDE (3.5B) [40]      | 12.24   | -      | -     |
| DALL-E-2 (6.5B) [45]   | 10.39   | -      | -     |
| Imagen (4.9B) [48]     | 7.27    | -      | -     |
| ERNIE-ViLG 2.0 (24B) [19] | 6.75 | -      | -     |
| MAGVLT (T2I only) (447M) | 10.74 | 23.94  | 8.12× |

| Available for both T2I & I2T |         |        |       |
| UPGen (307M) [4]          | 65.25   | -      | -     |
| L-Verse (500M) [32]       | 37.92   | -      | -     |
| ARGVLT (447M)             | 16.93   | 22.50  | 1.00× |
| MAGVLT (447M)             | 12.08   | 22.75  | 8.12× |

### Table 2. Zero-shot image inpainting results on MS-COCO validation.

<table>
<thead>
<tr>
<th>Model</th>
<th>FID (↓)</th>
<th>IS (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARGVLT</td>
<td>5.62</td>
<td>29.21</td>
</tr>
<tr>
<td>MAGVLT</td>
<td>3.17</td>
<td>30.79</td>
</tr>
</tbody>
</table>
### 4.3. Text Generation

**Image Captioning.** We evaluate the zero-shot image caption generation on MS-COCO Caption [38], and NoCaps [2]. We measure quantitative metrics of the quality of the generated caption compared to the ground truth by BLEU-4 (B-4), METEOR (M), CIDEr (C), and SPICE (S). The evaluation results on MS-COCO are shown in Table 3. Likewise as in text-to-image generation, we can verify that MAGVLTs improve performances over ARGVLTs significantly. Moreover, MAGVLT outperforms some baselines that leverage external language models in CIDEr and SPICE which are specifically designed for the captioning task. Note that MAGVLT has about six times smaller parameters than Flamingo-3B [3], and even can generate images by a single model. In addition, the universal MAGVLT performs better than the task-specific MAGVLT (I2T only) maybe due to the synergetic improvement between the modalities. Regarding the marginal speedup (1.56×) by MAGVLT over ARGVLT for I2T, compared to the fixed number of image tokens (256), the numbers of text tokens are quite small and generally less than 32, thus, the relative speedup by parallel predictions for 12 steps is reduced in text generation. Here, MAGVLT generates a text sequence given the pre-predicted target length. We provide qualitative samples in Figure 7.

Here, since L-Verse [32], which can generate both modalities, provides I2T results only by scratch training on MS-COCO, for more comparison to L-Verse, we also perform the training from scratch on MS-COCO alone and obtain a much better FID score for T2I but a slightly lower CIDEr score for I2T compared to L-Verse (18.49 vs. 45.8 in FID, 85.3 vs. 102.2 in CIDEr).

The evaluation results on NoCaps are shown in Table 4. Basically, including MS-COCO in training is beneficial in performing on NoCaps since the interface of the caption collection for NoCaps closely resembles that used for the collection of the MS-COCO captions. Yet, MAGVLT shows comparable performances compared to FewVLM [31]. Also in this task, MAGVLT significantly outperforms ARGVLT. MAGVLTs underperform in comparisons to VLKD [14] and SimVLM [64]. This might be due to that VLKD (RN50x16) has larger parameters (≃700M) than MAGVLT and also leverages an external language model (BART [37]) unlike MAGVLT. SimVLM uses ALIGN dataset [30] which contains 1.8B image-text pairs as well as C4 [44] dataset which contains 360M text-only instances while MAGVLT uses only 17M image-text pairs for pretraining.

**Text Infilling.** Similar to image inpainting, we conduct the text infilling experiment where the central 50% part of the text is erased by a mask and then replaced with generated tokens from the trained model. The qualitative samples are shown in Figure 13 in Appendix, where we can see that the infilled words generated by MAGVLT are better aligned with surrounding context words, compared to ARGVLT. Also, as shown in Table 5, MAGVLT quantitatively outperforms ARGVLT.
4.4. Ablation Studies

Variants of MAGVLT. Here, we investigate the effects of sampling weights for three multimodal tasks corresponding to $p_r$, in learning MAGVLT, and the results are shown in Table 6. We denote the three weights in a form of $T2I:I2T:IT2IT$ in the table. The trained model by $T2I$ only produces the best performance for text-to-image generation. However, in contrast to this, the trained model by $I2T$ only shows the worst performance for captioning. We observe that the training loss of the mask prediction of this $I2T$ only model is the lowest compared to the other models. This may suggest that the bias issue we mentioned in subsection 3.4 can be more serious, especially in non-AR methods. As shown in the result, learning $T2I$ along with $I2T$ resolves the issue to some extent. The ratio of $8:2:0$ shows better $I2T$ performance in CIDEr but inferior $T2I$ performance in FID than that of $0:0:1$. We observe that the inclusion of the uncorrupted cross-modal context is necessary. The model trained with $2:1:1$ weights shows the best captioning performance but the worst $T2I$ performance at a time. As the portion of $T2I$ training is getting larger, the model performs better in $T2I$ but worsens in $I2T$. It means that there is a trade-off between $T2I$ and $I2T$ according to the sampling ratio. We use the most balanced one ($8:1:1$) super-scripted by * as the default setting for MAGVLT.

It is noted that regarding the performance drop by $IT2IT$ modeling rather than $T2I$-only modeling for the $T2I$ task, compared to $I2T$, in $T2I$ the benefit of focusing on the cross-modal context is less significant, especially in the later refinement steps. Therefore, improved cross-modal attention by $IT2IT$ training could be less effective for the $T2I$ task. Having said that, the performance drop by MAGVLT for $T2I$ is small even though it can also perform $I2T$. And we observe that in terms of CLIP scores, our $IT2IT$ training is slightly better than $T2I$-only training ($0.3176$ vs. $0.3145$) due to the enhanced cross-modal alignment. Moreover, the capacity of our moderate-sized model would be still limited in generating both modalities. Under this limited capacity, there exist some performance trade-offs, and here it would be more leaned to $I2T$. Table 9 in Appendix shows that when we increase the model capacity about two times, the large model with $IT2IT$ performs better than the medium model with $T2I$-only training on the $T2I$ task.

Effectiveness of Additional Tasks. Here, we investigate the effects of the additional tasks (i.e. $UnrollMask$ and $MixSel$), and the results are shown in Table 7. MAGVLTs without the additional tasks are clearly underperformed in total. Applying $UnrollMask$ in pretraining significantly improves the performances of the base models on both $T2I$ and $I2T$ tasks. Including $MixSel$ along with $UnrollMask$ also improves the performances of the $T2I$-only model and especially the models on $I2T$. Similar to $IT2IT$ training, this improved cross-modal attention by $MixSel$ could be less effective for $T2I$ task compared to $I2T$ task. We also experimentally found that the performance gain by $MixSel$ is somewhat marginal for ARGVLT. We hypothesize that the causal attention of AR models would be hard to encode the randomly changing span of context.

## 5. Conclusion

In this work, we propose MAGVLT as a unified generative VL model that can produce both image and text data. MAGVLT is robustly trained on image-text pairs by multiple cross-modal tasks and significantly outperforms ARGVLT achieving strong performances on both of zero-shot $I2T$ and $T2I$ tasks. In future work, we first need to scale-up MAGVLT in terms of both model and data for better generalizability. Also, we aim to leverage a pretrained language model to enhance natural language processing and eventually to enable in-context VL learning. We will also try to exploit an encoder-decoder architecture to robustly perform on both understanding and generation tasks.

## Acknowledgements

We would like to thank Brain Cloud Team at Kakao Brain for their support. This work was also supported by Korea University Grant (K2304351) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI).
References


[29] Emil Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forrê, and Max Welling. Armgax flows and multinomial dif-

30. Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021. 1, 2, 7


[65] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Yuan Cao. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. In ACL, 2021. 2


[67] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789, 2022. 1, 2, 6


[69] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junseuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In ICCV, 2019. 5
