MAGVLT:

Masked Generative Vision-and-Language Transformer

Sungwoong Kim[1]*, Daejin Jo[2]*, Donghoon Lee[2]*, Jongmin Kim[2]*

THU-PM-261

1 KOREA UNIVERSITY 2 kakaobrain

Multi-Modal Generative Modeling

Can one model generate multi-modal data?

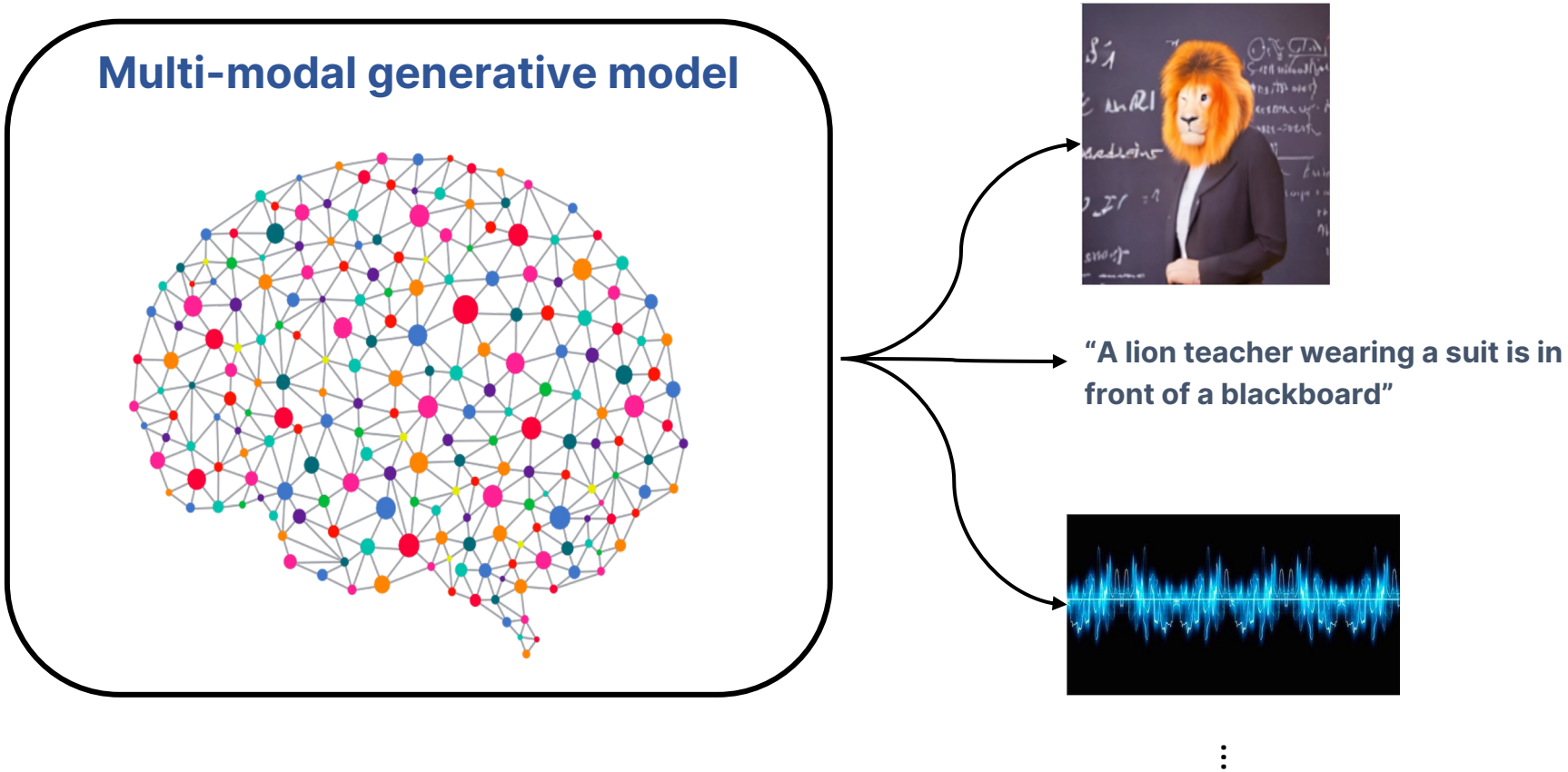
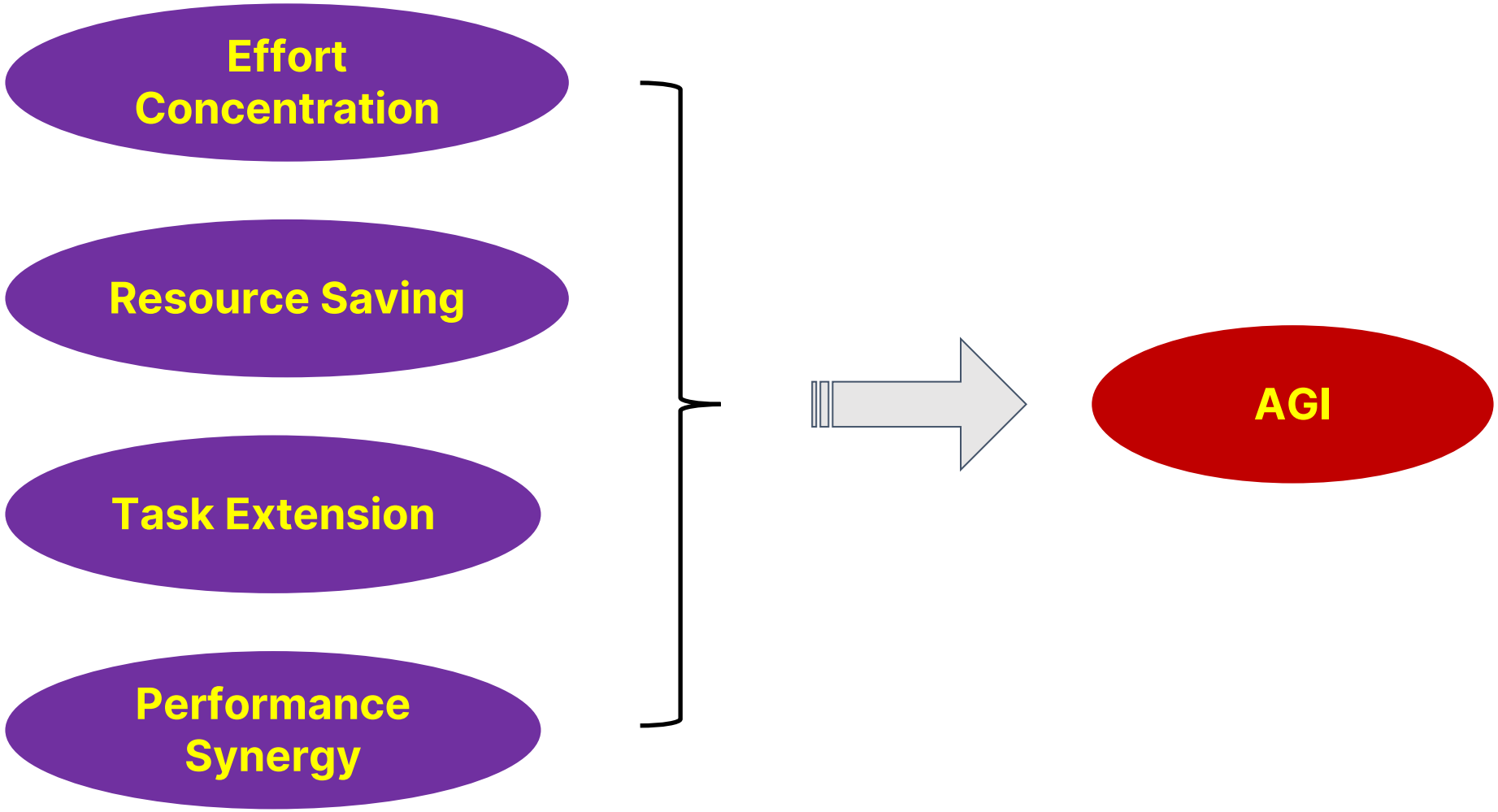


Image from Ding et al. (2022)

Benefits of One Universal Generative Model



MAGVLT: Masked Generative Vision-and-Language Transformer

- **Contributions**

  - **MAGVLT: Unified Generative Vision-and-Language (VL) Model based on Masked Generative Transformer.**

  - **Robust training on image-text pairs: cross-modal mask prediction + step-unrolled mask prediction + selective prediction on the mixed context.**

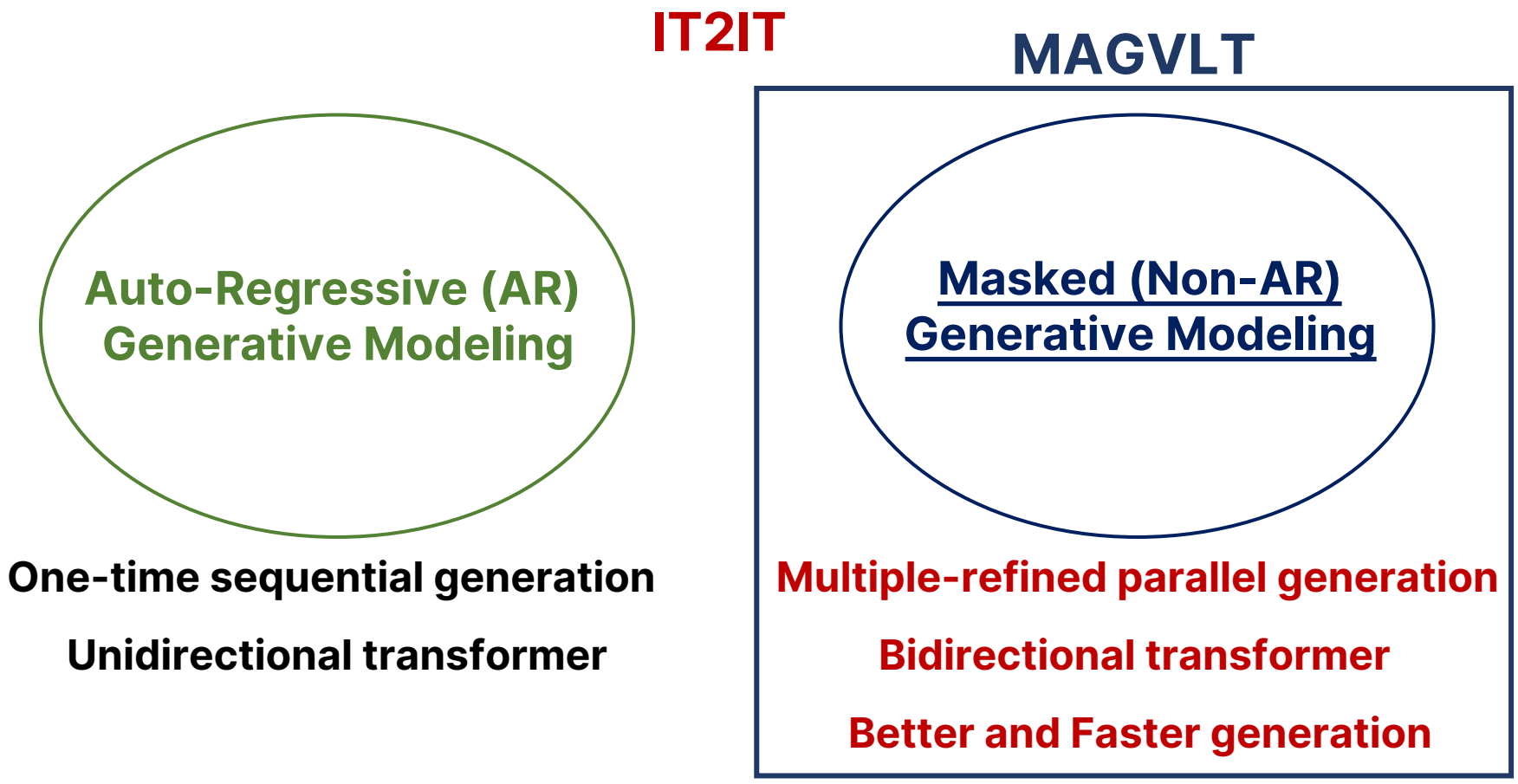
  - **Competitive performances of MAGVLT on both of zero-shot image-to-text (I2T) and text-to-image (T2I) generation tasks for the first time.**

MAGVLT: Generative Vision-and-Language Model

One universal IT2IT (Image+Text to Image+Text) model can do …

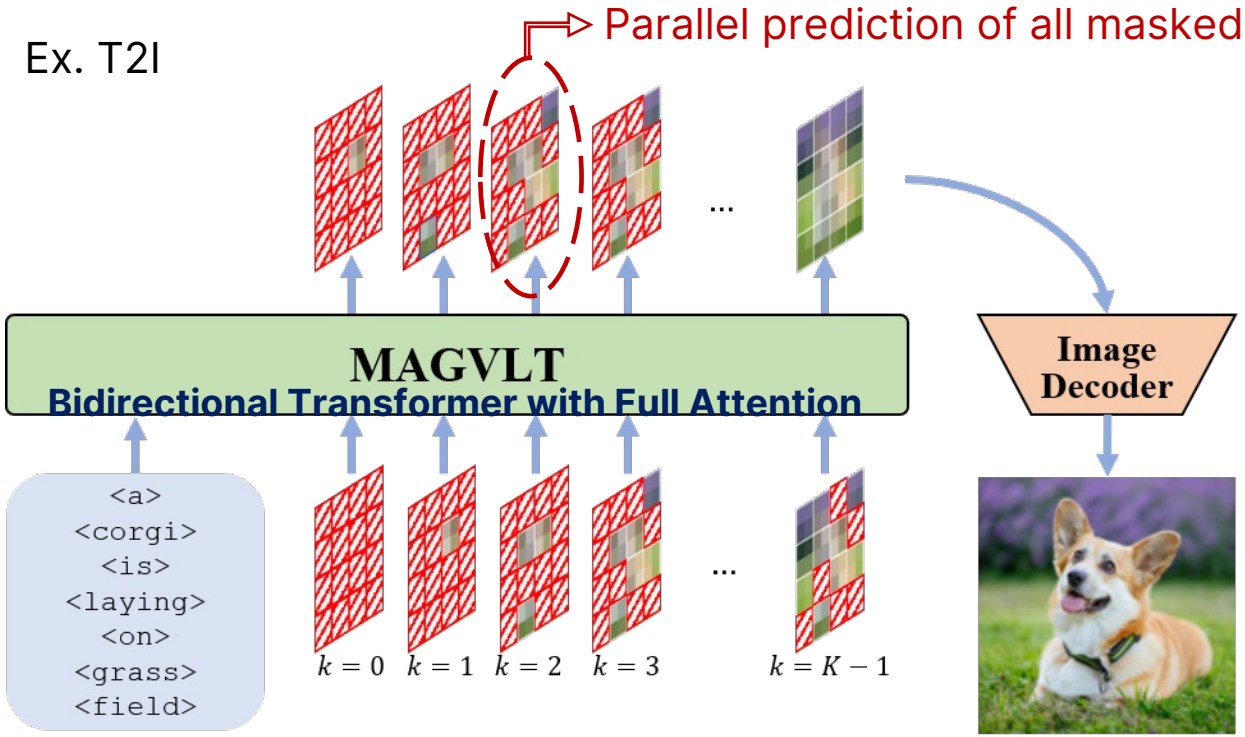
T2I	-> Text-guided image generation
I2T	-> Image captioning
I2I	-> Image transfer, inpainting
T2T	-> Translation, QA, text infilling
IT2I	-> Text-guided image editing
IT2T	-> Visual QA
T2IT	-> Visual story generation
ITITIT…2IT	-> In-context vision+language generation

MAGVLT: Generative Vision-and-Language Model



MAGVLT: Masked Generative Vision-and-Language Transformer

Inference (Iterative Refinement)



Generation: Iterative refinement (denoising)

Train: Variable mask ratio, simulating one of refining (denoising) steps

} **Similar to Discrete Diffusion**

** I2T Generation: Target Length Prediction on <BOT> trained by CE loss $\mathcal{L}_{\mathrm{TL}}(N_T, \hat{N}_T)$

MAGVLT: Masked Generative Vision-and-Language Transformer

Train

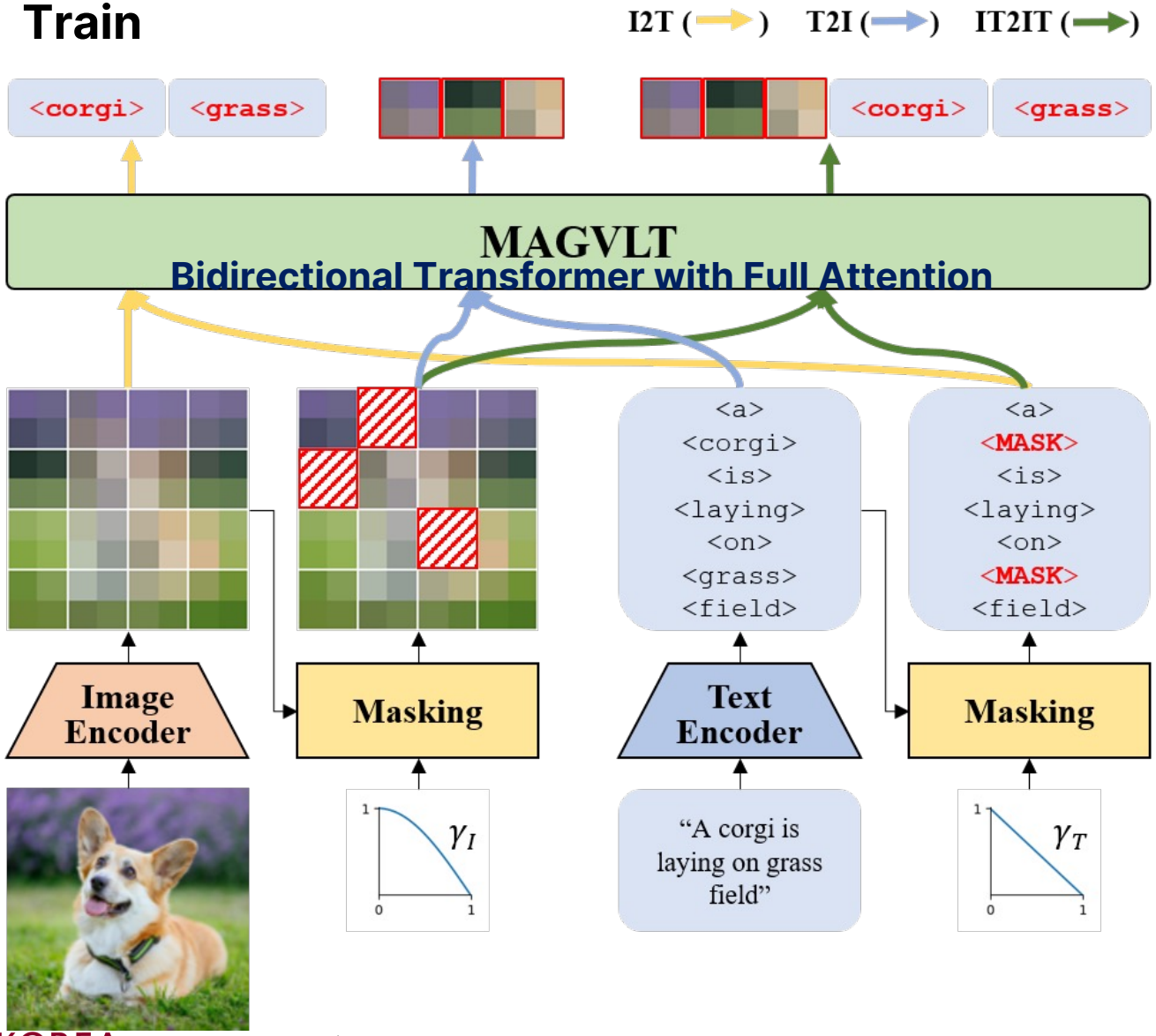


Image tokens: $X = [x_i]_{i=1}^{N_I}$

Text tokens: $Y = [y_j]_{j=1}^{N_T}$

Masked image tokens: $X_{\bar{M}_I}$

Masked text tokens: $Y_{\bar{M}_T}$

$$\mathcal{L}_{\text{I2T}} = -\mathop{\mathbb{E}}_{(X,Y)\in\mathcal{D}}\left[\sum_{\forall j\in[1,N_T],m_j^T=1}\log p(y_j|Y_{\bar{M}_T},X)\right], (1)$$

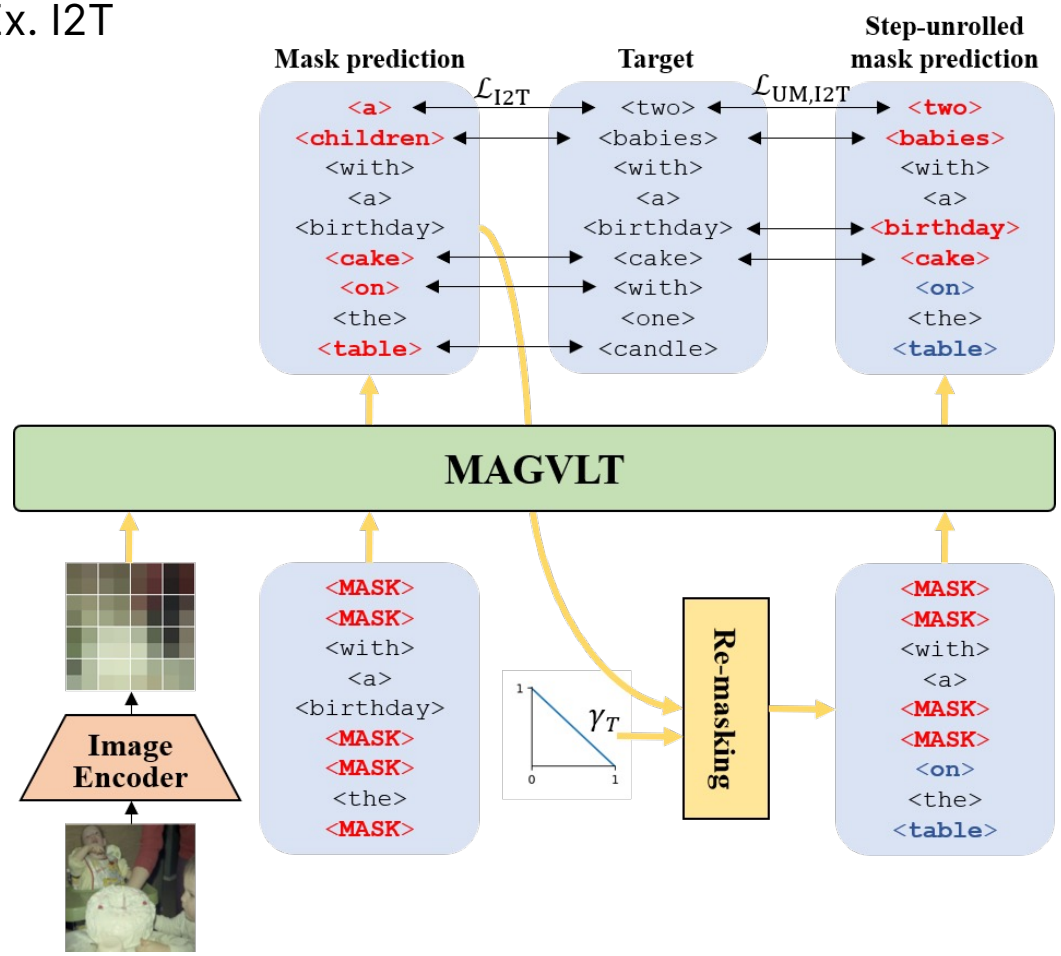
$$\mathcal{L}_{\text{T2I}} = -\mathop{\mathbb{E}}_{(X,Y)\in\mathcal{D}}\left[\sum_{\forall i\in[1,N_I],m_i^I=1}\log p(x_i|X_{\bar{M}_I},Y)\right], (2)$$

$$\mathcal{L}_{\text{IT2IT}} = -\mathop{\mathbb{E}}_{(X,Y)\in\mathcal{D}}\left[\sum_{\forall j\in[1,N_T],m_j^T=1}\log p(y_j|Y_{\bar{M}_T},X_{\bar{M}_I}) + \sum_{\forall i\in[1,N_I],m_i^I=1}\log p(x_i|X_{\bar{M}_I},Y_{\bar{M}_T})\right], (3)$$

MAGVLT: Masked Generative Vision-and-Language Transformer

Step-Unrolled Mask Prediction (UnrollMask)

Ex. I2T



SUNDAE* tries to reduce **the gap** between
a corruption on the target tokens at training and
a corruption on the partially predicted tokens at testing.

Motivated by this, MAGVLT **remasks the one-step predicted sequence** then predicts the re-masked tokens.

$$\mathcal{L}_{\text{UM,I2T}} = -\mathop{\mathbb{E}}_{(X,Y)\in\mathcal{D}}\left[\sum_{\forall j\in[1,N_T],m_j^{T(+1)}=1}\log p(y_j|\hat{Y}^{(+1)}_{\bar{M}_T^{(+1)}},X)\right], \quad (4)$$

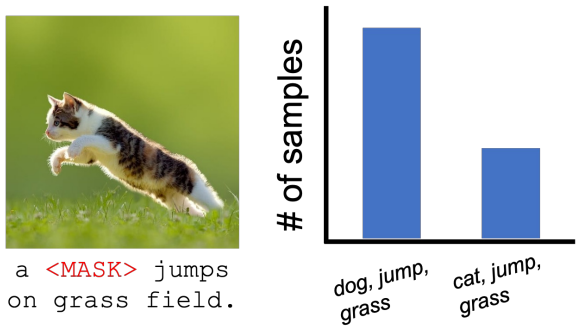
$\hat{Y}^{(+1)}_{\bar{M}_T^{(+1)}}$: re-masked one-step unrolled prediction of $Y_{\bar{M}_T}$

N. Savinov, et al., Step-unrolled denoising autoencoders for text generation. In ICLR, 2022

MAGVLT: Masked Generative Vision-and-Language Transformer

Selective Prediction on Mixed Context (MixSel)

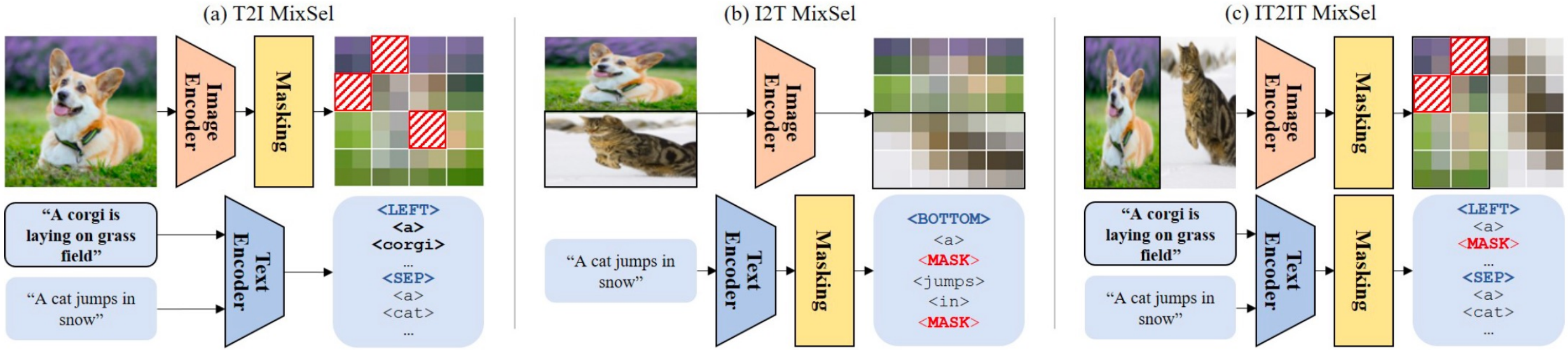
Ex. within-modal bias



To reduce this bias and overlooking the cross-modal context, **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 with a prepended special token to inform the selected context.

$$\mathcal{L}_{\text{MS, I2T}} = -\mathop{\mathbb{E}}_{(X,Y)\in\mathcal{D}}\left[\sum_{\forall j\in[1,N_T], m_j^T=1} \log p(y_j^\ell | \hat{Y}^\ell_{\bar{M}_T}, \phi(X^1, X^2))\right]$$

ϕ: mixture function
$\ell \in \{1, 2\}$: the selected context



MAGVLT: Masked Generative Vision-and-Language Transformer

Multitask Pretraining

$$\mathcal{L}_\tau = \mathcal{L}_{\text{mask},\tau} + \lambda_{\text{TL}}\mathcal{L}_{\text{TL},\tau} + \lambda_{\text{UM}}\mathcal{L}_{\text{UM},\tau} + \lambda_{\text{MS}}\mathcal{L}_{\text{MS},\tau}$$

$$\tau \in \{\text{I2T, T2I, IT2IT}\},\ \lambda_{\text{TL}} = 0.01,\ \lambda_{\text{UM}} = 1.0,\ \lambda_{\text{MS}} = 0.5$$

A task $\tau \in \{\text{I2T, T2I, IT2IT}\}$ is sampled from the predefined p_τ for each iteration (batch-wise).

Experiments

- **Model**
  - 447M parameters in total including VQ-GAN

	ARG/MAGVLT
Params	371M
Layers	24
Embed Dim	1024
Heads	8

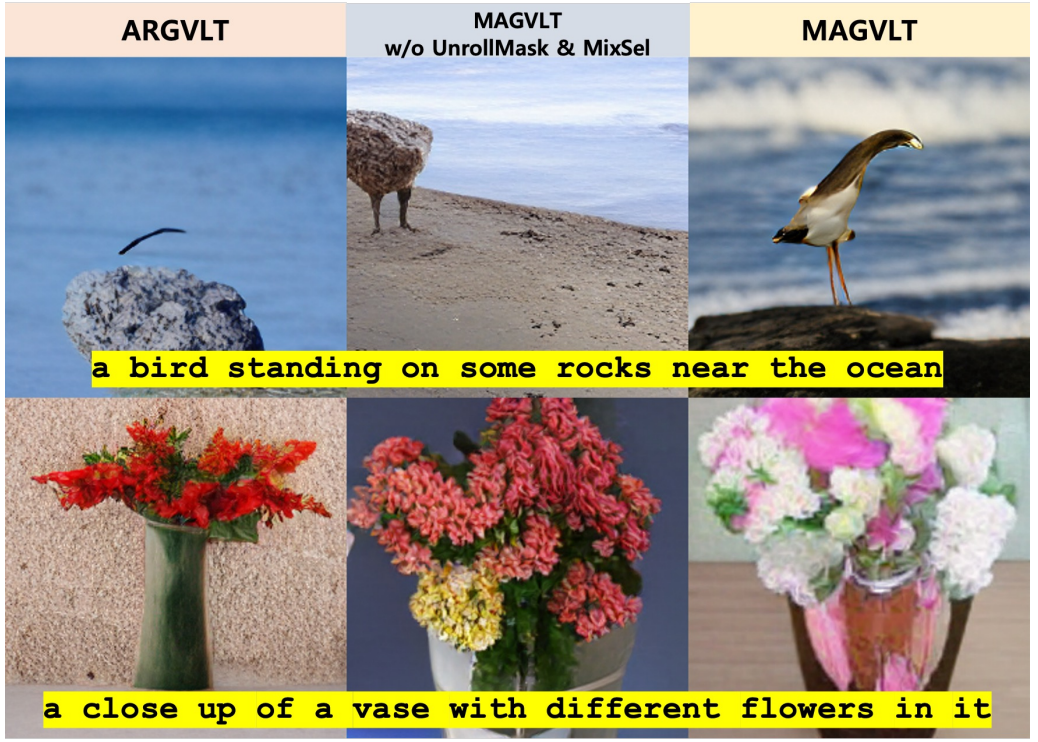
  - VQ-GAN converts a 256x256 image into 16x16 tokens with 16,384 codebook size.
  - BPE tokenizer with 49,408 vocab size converts a sentence to a text token sequence (max length: 64).
- **Pretraining**
  - CC3M + CC12M + SBU + VG: total 17M image-text pairs
  - 40K updates with a batchsize of 4,096 from scratch w/ only image+text pairs
- **Evaluation**
  - Zero-shot T2I on MS-COCO
  - Zero-shot I2T on MS-COCO and NoCaps
- **Sampling**
  - # of refinement steps: 10 for image generation, 12 for text generation
  - Clip reranking

Experiments

- Image Generation

Model	FID (↓)	IS (↑)	Speed
AR based			
CM3-Medium (2.7B) [1]	36.78	-	-
DALL-E (12B) [46]	27.5	17.9	-
CogView (4B) [16]	27.1	18.2	-
CogView2 (6B) [17]	24.0	22.4	-
Parti-350M (350M) [67]	14.10	-	-
Make-A-Scene (4B) [20]	**11.84**	-	-
ARGVLT (*T2I only*) (447M)	21.80	19.27	1.00×
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]	37.2	-	-
ARGVLT (447M)	16.93	22.50	1.00×
MAGVLT (447M)	**12.08**	**22.75**	8.12×

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.



Experiments

- Text Generation

Model	B-4	M	C	S	Speed
with external language model					
ZeroCap (345M) [58]	2.6	11.5	14.6	5.5	-
MAGIC (1.5B) [56]	12.9	17.4	49.3	11.3	-
VLKD$_{\text{ViT-B/16}}$ (406M) [14]	**16.7**	**19.7**	58.3	**13.4**	-
Flamingo-3B (3B) [3]	-	-	**73.0**	-	-
without external language model					
SimVLM$_{\text{huge}}$ (632M) [64]	11.2	14.7	32.2	8.5	-
ARGVLT (*I2T only*) (447M)	11.4	15.1	47.4	11.4	1.00×
ARGVLT (447M)	10.9	14.9	45.5	11.2	1.00×
MAGVLT (*I2T only*) (447M)	12.9	17.1	53.5	12.9	1.56×
MAGVLT (447M)	**14.6**	**19.0**	**60.4**	**14.3**	1.56×

Table 3. *Zero-shot* I2T results on MS-COCO Karpathy test.

GT	ARGVLT	MAGVLT
A hairy brown cow laying on top of a field.	A brown and white cow.	A brown and white cow laying in the grass.
A cat behind flowers in a vase, small pumpkins, a wine bottle and a glass of wine.	A white and black cat.	A cat with a bottle of wine glasses, and a glass vase with some flowers in the background.
A sepia toned photo of a baby snuggling with a giant teddy bear.	A woman with a teddy bear.	The baby is sitting on on the floor and holding the arm of the bear.

Experiments

- **Ablations**

Task sample weights T2I:I2T:IT2IT	CIDEr (↑)	FID (↓)
1:0:0 (*T2I only*)	-	**10.74**
0:1:0 (*I2T only*)	53.5	-
0:0:1 (*IT2IT only*)	55.3	12.06
8:2:0 (*T2I & I2T*)	59.7	13.09
2:1:1	**61.7**	15.17
6:1:1	60.7	12.65
8:1:1*	60.4	12.08
10:1:1	59.2	12.07

Table 6. Variants of MAGVLT.

Model	CIDEr (↑)	FID (↓)
MAGVLT (*T2I only*)	-	**10.74**
w/o MixSel	-	10.97
w/o UnrollMask and MixSel	-	11.72
MAGVLT (*I2T only*)	**53.5**	-
w/o MixSel	51.3	-
w/o UnrollMask and MixSel	48.0	-
MAGVLT	**60.4**	**12.08**
w/o MixSel	58.9	**12.07**
w/o UnrollMask	56.5	13.26
w/o UnrollMask and MixSel	53.8	14.12

Table 7. Effectiveness of additional training tasks.

Experiments

- **Model Scaling**

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

T2I

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.

I2T

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**).

Experiments

- **Finetuning on Downstream Tasks**

  - Generation & Understanding

Finetuned I2T

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

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

Finetuned VQAv2

Model	test-dev	test-std
$\text{VLKD}_{\text{ViT-B/16}}$ [14]	69.8	-
MetaLM [24]	**74.4**	**74.5**
MAGVLT	63.0	63.4
$\text{MAGVLT}_{\text{Large}}$	65.7	66.2

Table 12. Experimental results on VQAv2.

Experiments

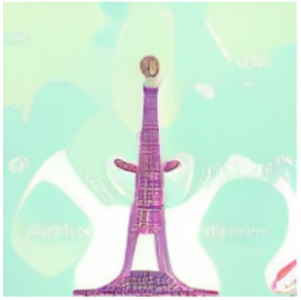
- **Unconditional Image+Text Generation**

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

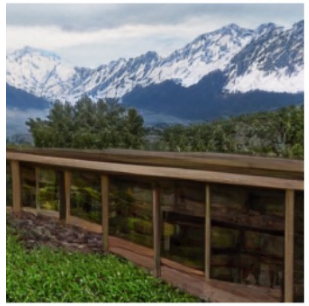


eiffel tower on a pastel background royalty free illustration

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



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

Conclusion

- MAGVLT as a unified generative VL model that can produce both image and text data.
- Masked generative transformer with a robust training on image-text pairs by multiple training tasks.
- MAGVLT outperforms ARGVLT and shows competitive performances on both of zero-shot image-to-text and text-to-image generation tasks for the first time.

Ongoing works

- Scale-up both model and data
- Leverage language model and data
- Universal model for both understanding and generation
- Increase modalities (e.g. audio)

Poster: THU-PM-261

https://github.com/kakaobrain/magvlt