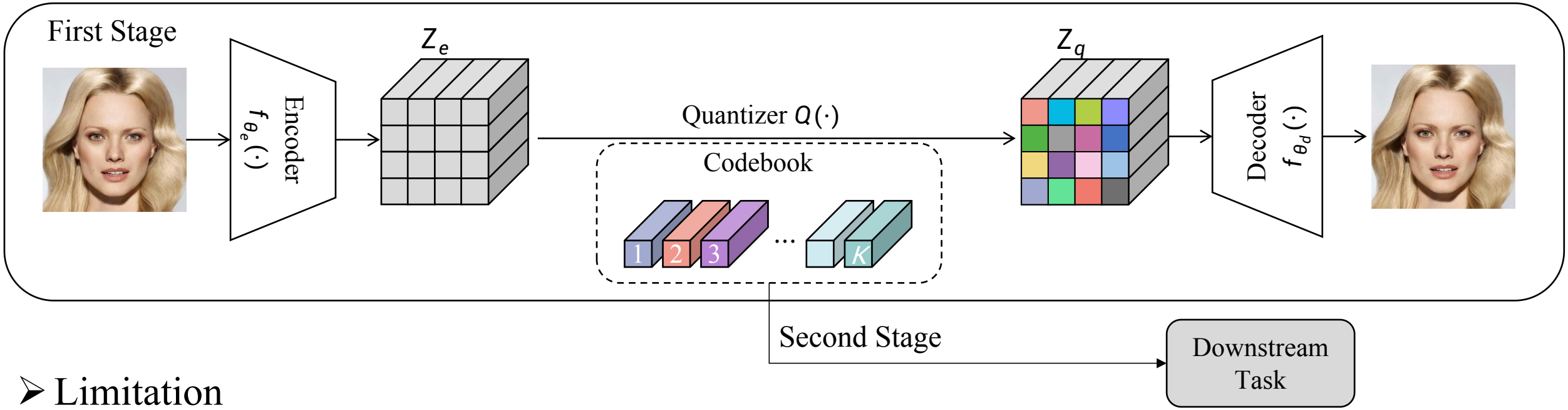
LG-VQ: Language-Guided Codebook Learning

Guotao Liang[1,2], Baoquan Zhang[1], Yaowei Wang[2], Xutao Li[1], Yunming Ye[1]
Huaibin Wang[1], Chuyao Luo[1], Kola Ye[3], linfeng Luo[3]
[1]Harbin Institute of Technology, Shenzhen
[2]Peng Cheng Laboratory, [3]SiFar Company
{lianggt, wangyw}@pcl.ac.cn
{zhangbaoquan, 22S051022}@stu.hit.edu.cn
{lixutao, yeyunming}@hit.edu.cn
{luochuyao.dalian, kolaygm, llf10811020205}@gmail.com

Motivation

➢ Vector Quantization



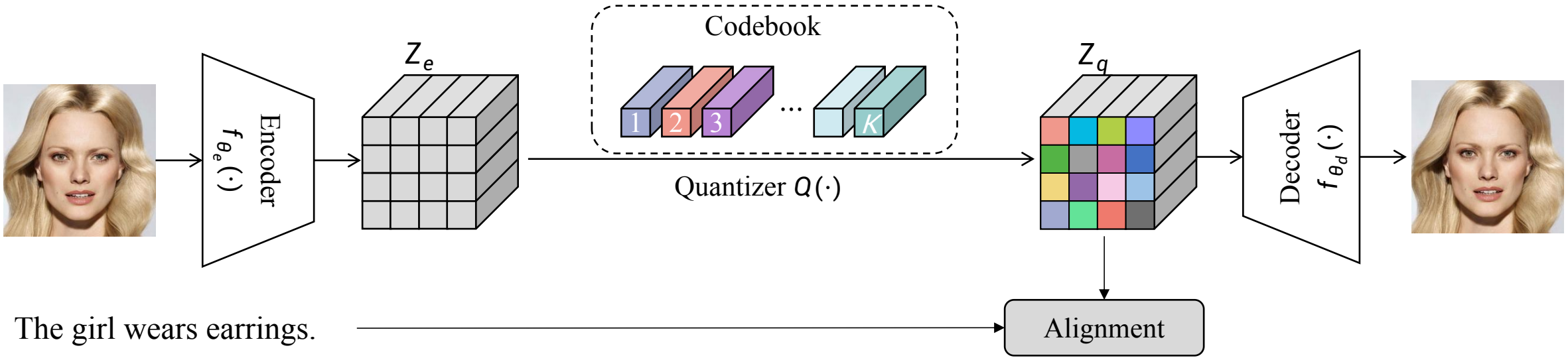
➢ Limitation

- Most methods only focus on learning ***a single-modal codebook***, resulting in suboptimal performance when the codebook is applied to multi-modal downstream tasks

Idea

➢ Limitation

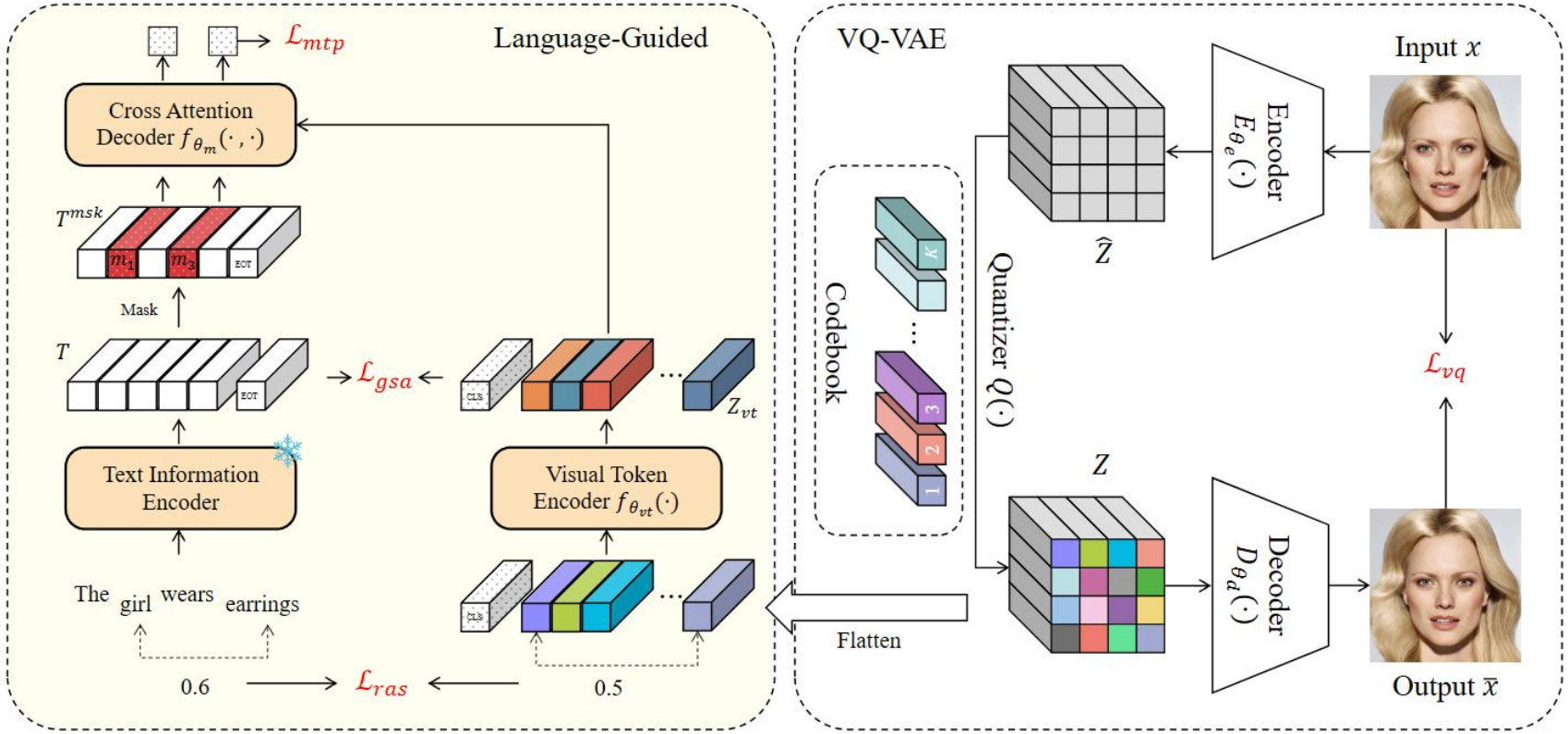
- Most methods only focus on learning ***a single-modal codebook***, resulting in suboptimal performance when the codebook is applied to multi-modal downstream tasks



➢ The novelty lies in utilizing **pre-trained text semantics** to guide the model to learn **text-aligned codebook**

Proposed Method

➢ The novelty lies in utilizing **pre-trained text semantics** to **guide** the model to learn **text-aligned codebook**



- Semantic Alignment Module

  $\mathcal{L}_{gsa}$ $\mathcal{L}_{mtp}$

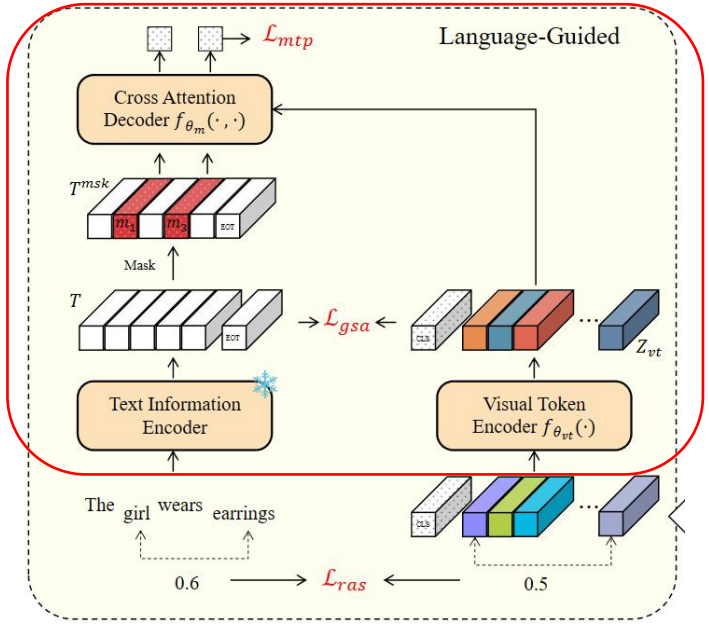
- Relationship Alignment Module

  $\mathcal{L}_{ras}$

Proposed Method

➢ Semantic Alignment Module

- Insight: Considering that **paired image and text data** have consistent semantic information and the missing information of masked data can be completed from the other modality. We propose **global semantic alignment** and **masked text prediction**.

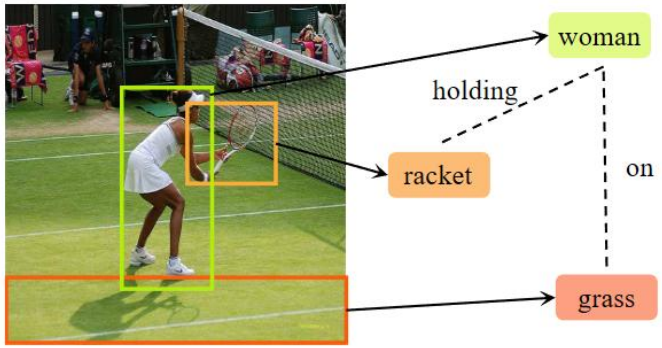


$$\mathcal{L}_{gsa} = -\sum_{i\in\mathcal{B}} \log \frac{\exp(s(e^i_{CLS}, e^i_{EOT}))}{\sum_{j\in\mathcal{B}} \exp(s(e^j_{CLS}, e^j_{EOT}))}.$$

$$\mathcal{L}_{mtp} = -\mathbb{E}_{(Z_{vt},T^{msk})\sim\mathcal{B}} H(y_{msk}, f_{\theta_m}(Z_{vt}, T^{msk})).$$

Proposed Method

- Semantic Alignment Module **Limiting**
  - Cannot satisfy more complex reasoning tasks like image captioning and VQA.



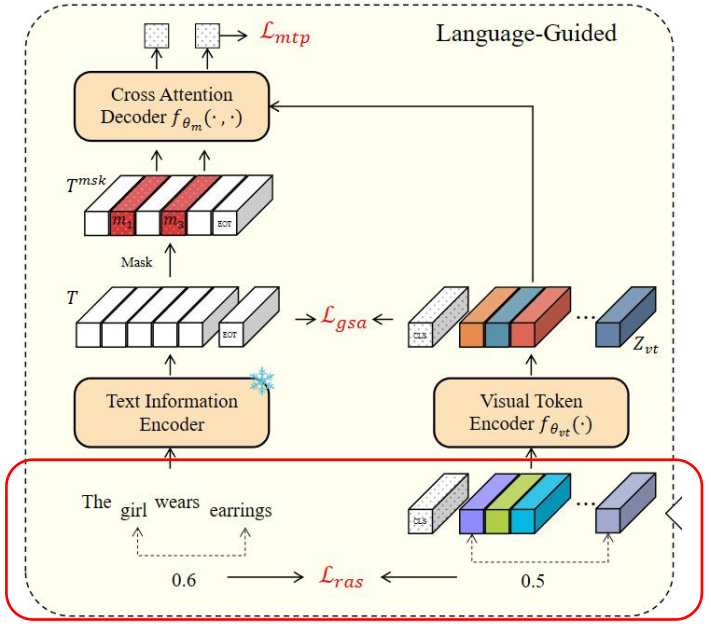
Q: What is this woman holding?

Figure 1: To answer the question, one not only needs to identify "women" and "racket" but also understand the semantic relationship between them ("holding").

- **Relationship Alignment Module**

$$e_{z_i} = Z[\underset{e_z \in Z_{vt[1:]}}{argmax}\, s(e_{w_i}, e_z), :], \qquad e_{z_j} = Z[\underset{e_z \in Z_{vt[1:]}}{argmax}\, s(e_{w_j}, e_z), :].$$

$$\mathcal{L}_{ras} = \sum_{(w_i, w_j) \in t} (s(e_{w_i}, e_{w_j}) - s(e_{z_i}, e_{z_j}))^2.$$



Experiments

➢ Reconstruction Performance

Table 1: Results of image reconstruction on TextCaps, CelebA-HQ, CUB-200, and MS-COCO. "VQ-VAE+LG" denotes considering our method LG-VQ based on VQ-VAE.

Models	TextCaps		CelebA-HQ		CUB-200		MS-COCO	
	FID↓	PSNR↑	FID↓	PSNR↑	FID↓	PSNR↑	FID↓	PSNR↑
VQ-VAE	82.31	**21.96**	41.45	**25.57**	54.92	24.38	86.21	**23.55**
VQ-VAE+LG	**81.93**	21.95	**40.53**	25.04	**36.55**	**25.60**	**79.54**	23.40
VQ-GAN	24.08	19.64	5.66	**24.10**	3.63	22.19	14.45	20.21
VQ-GAN+LG	**20.35**	**19.92**	**5.34**	23.75	**3.08**	**22.47**	**10.72**	**20.50**
CVQ	16.35	**20.24**	5.19	23.15	3.61	22.29	9.94	20.48
CVQ+LG	**15.51**	20.21	**4.90**	**24.48**	**3.33**	**22.47**	**9.69**	**20.71**

Experiments

➢ Ablation Study

Table 2: Ablation study of our three loss functions on TextCaps and CUB-200.

	Setting	TextCaps FID↓	CUB-200 FID↓
(i)	Baseline(VQ-GAN)	24.08	3.63
(ii)	+ $\mathcal{L}_{gsa}$	23.01	3.39
(iii)	+ $\mathcal{L}_{mtp}$	21.54	3.49
(iv)	+ $\mathcal{L}_{mtp}$ + $\mathcal{L}_{ras}$	20.77	3.32
(v)	+ $\mathcal{L}_{mtp}$ + $\mathcal{L}_{gsa}$	20.46	3.34
(vi)	+ $\mathcal{L}_{mtp}$ + $\mathcal{L}_{gsa}$ + $\mathcal{L}_{ras}$	20.35	3.08

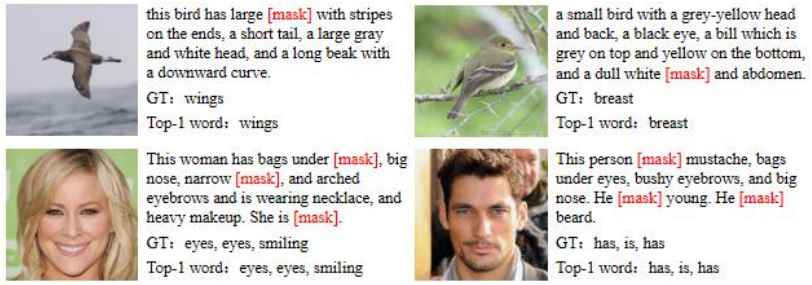


Figure 5: Examples of the top-1 word predicted on masked word prediction task.

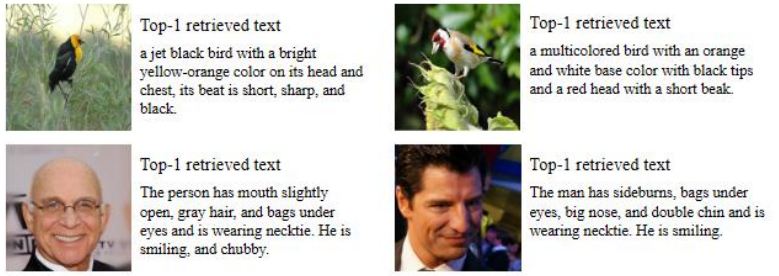


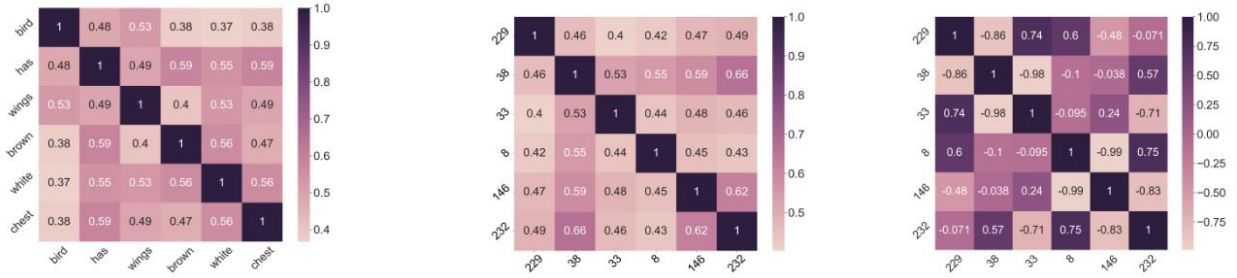
Figure 4: Examples of the top-1 most similar text selected on image-to-text retrieval task.

Table 3: Results (Recall@1) of masked word prediction on CelebA-HQ and CUB-200. "Mask-1" denotes that text is randomly masked one word.

Dataset		Recall@1
CelebA-HQ	Mask-1	99.55
	Mask-3	99.24
CUB-200	Mask-1	83.65
	Mask-3	80.17

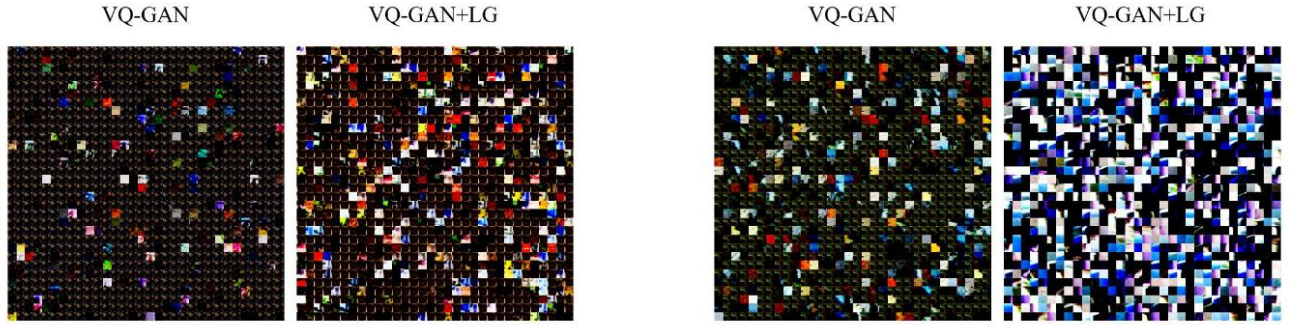
Experiments

➢ Ablation Study



(a) word similarity (b) LG-VQ code similarity (c) VQ-GAN code similarity

Figure 6: Visualization of words similarity and image codes similarity aligned with the word. We extract some representative words from the text as a demonstration.



(a) The usage of codebook on VQ-GAN is 18.62% and VQ-GAN+LG is 43.58% on TextCaps

(b) The usage of codebook on VQ-GAN is 40.09% and VQ-GAN+LG is 97.89% on MS-COCO

Figure 7: Visualization of the codebook of VQ-GAN and LG-VQ on TextCaps and MS-COCO.

Table 4: Results of similarity evaluation between codes and words on CUB-200 all test data.

Method	VQ-GAN	VQ-GAN+LG
MSE↓	0.6374	0.0351

Experiments

➢ Application

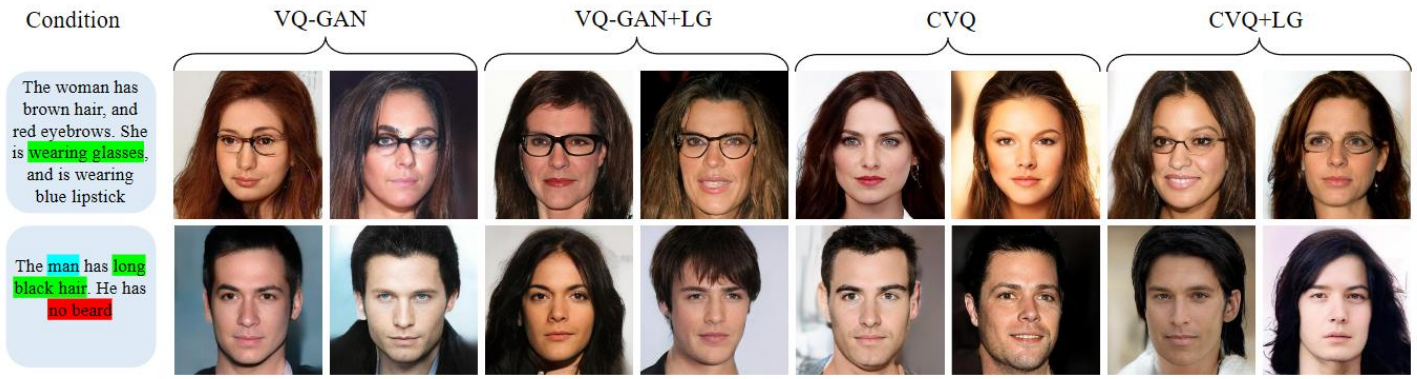
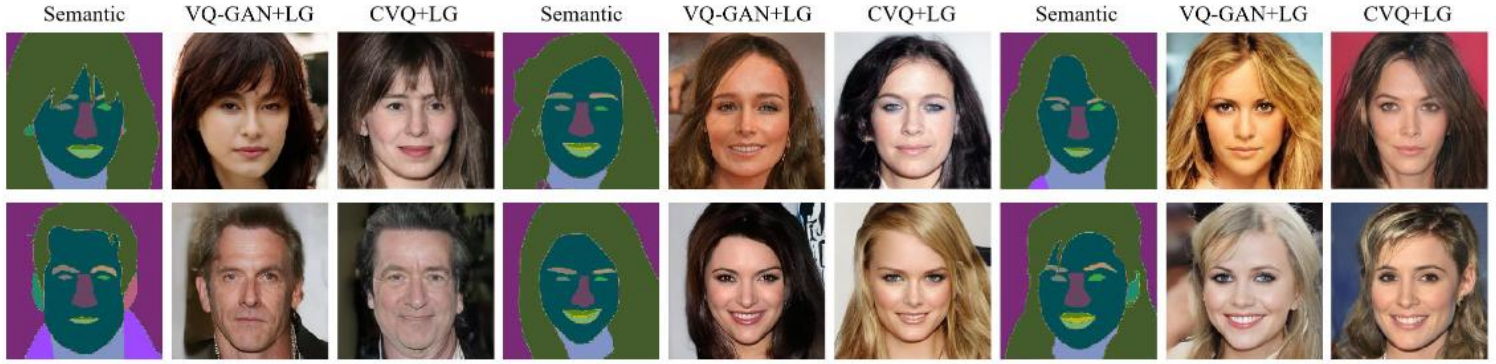


Figure 8: Text-to-image synthesis and semantic image synthesis on CelebA-HQ. Text with background color emphasizes generated details

Table 9: Results of text-to-image on CelebA-HQ dataset.

Model	Text-to-image FID↓
AttnGAN (2018)	125.98
ControlGAN (2019)	116.32
TediGAN (2021)	106.37
Unite and Conqu (2023)	26.09
Corgi (2023)	19.74
LAFITE (2022)	12.54
VQ-GAN+LG	12.61
CVQ+LG	12.33



Setting	VQ-VAE		VQ-GAN		CVQ	
	w/o LG	w LG	w/o LG	w LG	w/o LG	w LG
text-to-image	49.51	**49.36**	15.29	**12.61**	13.23	**12.33**
image synthesis	48.72	**48.23**	11.53	**11.46**	11.04	**11.03**

Experiments

➢ Application

Figure A: Examples of **unconditional image generation** on CelebA-HQ based on VQ-GAN+LG.

Table 3: Result (FID↓) of unconditional image synthesis on CelebA-HQ.

Model	CelebA-HQ
Style ALAE*	19.2
DC-VAE*	15.8
VQ-GAN*	10.2
VQ-GAN+LG	9.1
Improvement	**10.78%**

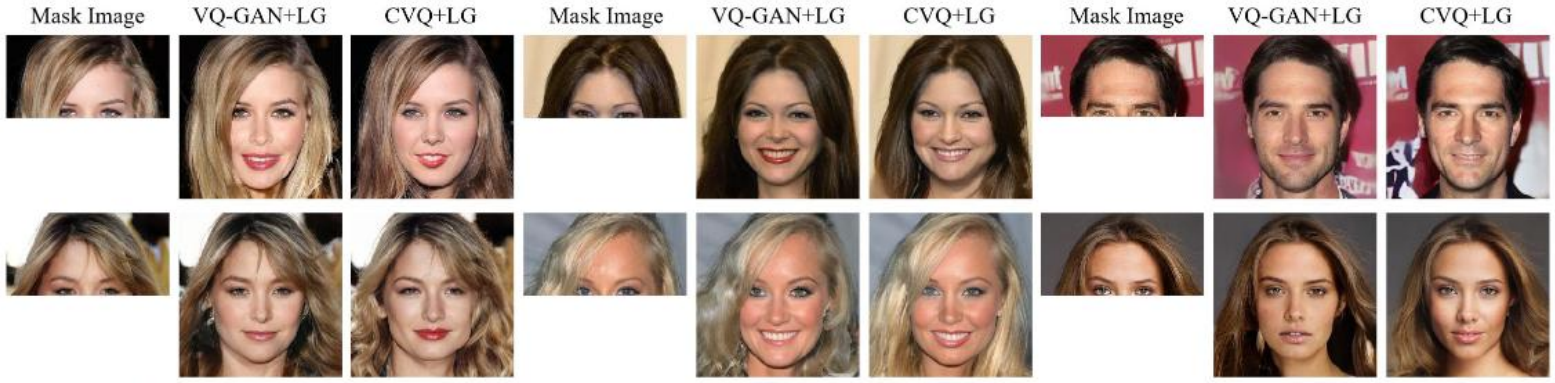


Table 4: Result (FID↓) of image completion on CelebA-HQ.

Model	CelebA-HQ
VQ-GAN	9.02
VQ-GAN+LG	8.14
Improve	**9.76%**

Experiments

➢ Application

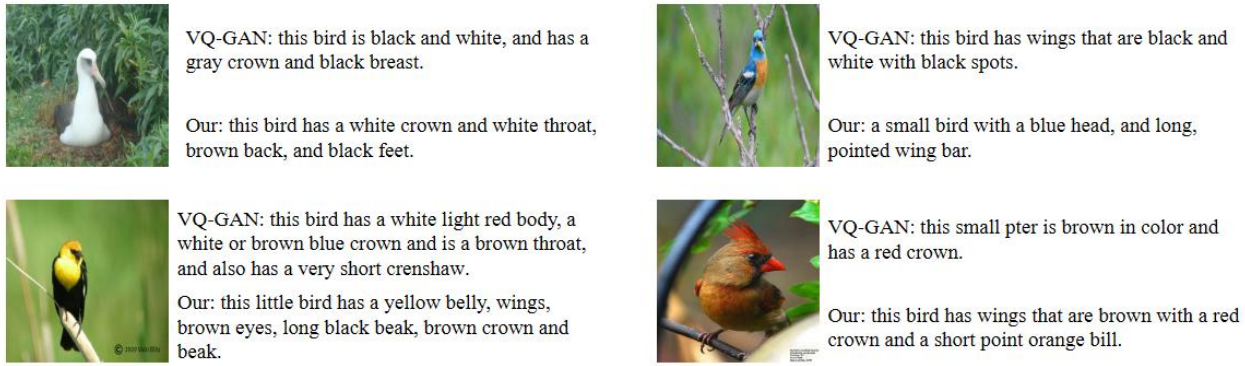


Figure 15: Image Captioning on CUB-200 based on VQ-GAN and VQ-GAN+LG.

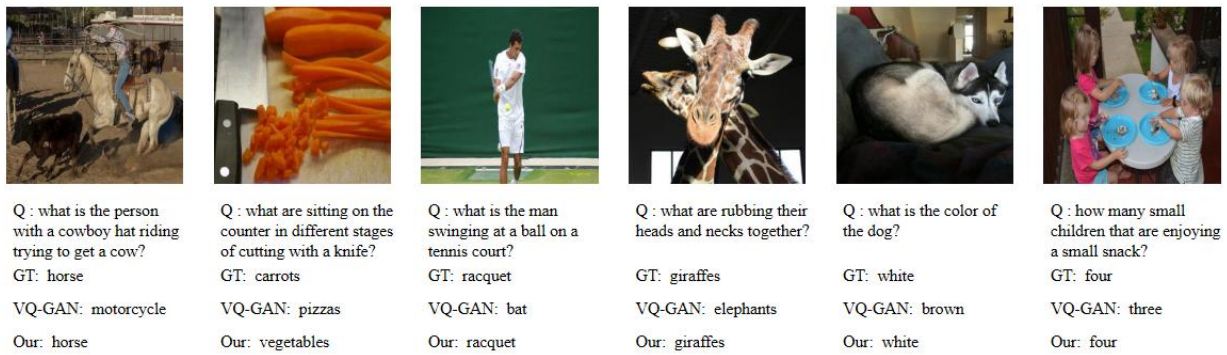


Figure 16: VQA on COCO-QA based on VQ-GAN and VQ-GAN+LG.

Table 2: Results of image captioning on CUB-200 datasets.

Model	Image Captioning			
	BLEU4↑	ROUGE-L↑	METEOR↑	CIDEr-D↑
VQ-GAN	1.29	33.40	24.47	93.62
V2L Tokenizer	1.59	30.65	25.76	**104.14**
VQCT	1.38	26.50	24.63	98.22
VQ-GAN+LG	**1.69**	**34.73**	**25.78**	102.77

Table 6: Results of (Accuracy and WUPS [39]) VQA on COCO-QA [28] dataset using MS-COCO's codebook.

Setting	VQA	
	Accuracy↑	WUPS↑
VQ-GAN	37.82 ± 0.97	82.06 ± 0.54
VQ-GAN+LG	40.97 ± 0.13	83.56 ± 0.11

Experiments

➢ Application

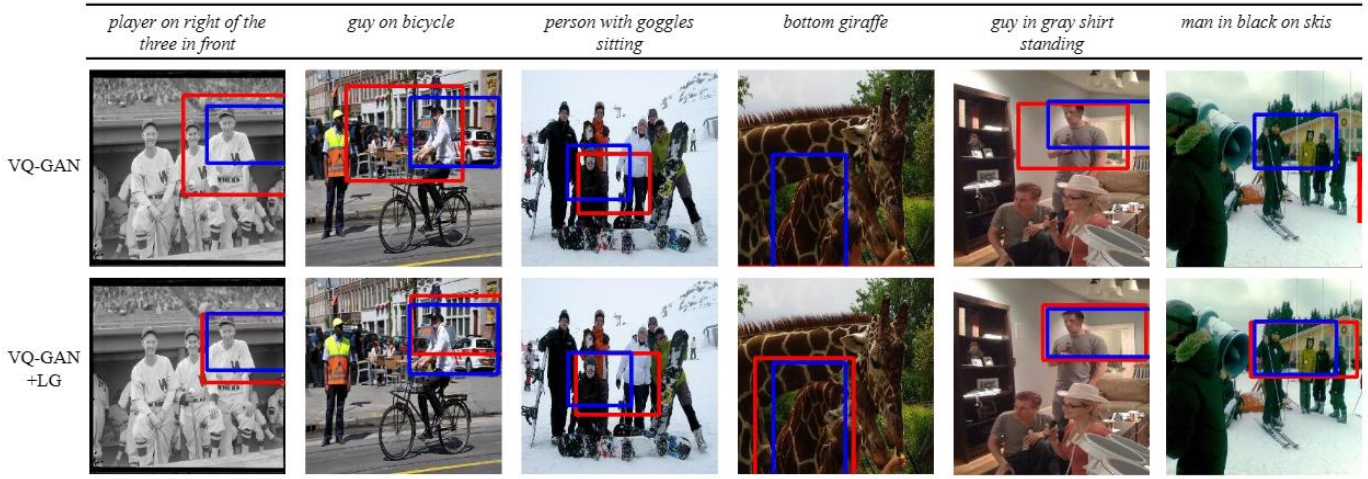


Figure C: Examples of **visual grounding** on refcoco. Blue boxes are the ground-truth, red boxes are the predictions of the model.

TABLE VIII: Results (Accuracy (0.5) ↑) of visual grounding on refcoco datasets [57] using MS-COCO's codebook.

Model	Visual Grounding Accuracy (0.5)↑
VQ-GAN [7]	9.14
VQCT [27]	9.46
LG-VQ [38]	9.62
VQCT-VLT	10.44

Thanks for Listening !