LG-VQ: Language-Guided Codebook Learning

Guotao Liang^{1,2}, Baoquan Zhang¹, Yaowei Wang², Xutao Li¹, Yunming Ye¹ Huaibin Wang¹, Chuyao Luo¹, Kola Ye³, linfeng Luo³

¹Harbin Institute of Technology, Shenzhen

²Peng Cheng Laboratory, ³SiFar Company
{lianggt, wangyw}@pcl.ac.cn
{zhangbaoquan, 22S051022}@stu.hit.edu.cn
{lixutao, yeyunming}@hit.edu.cn
{luochuyao.dalian, kolaygm, llf10811020205}@gmail.com







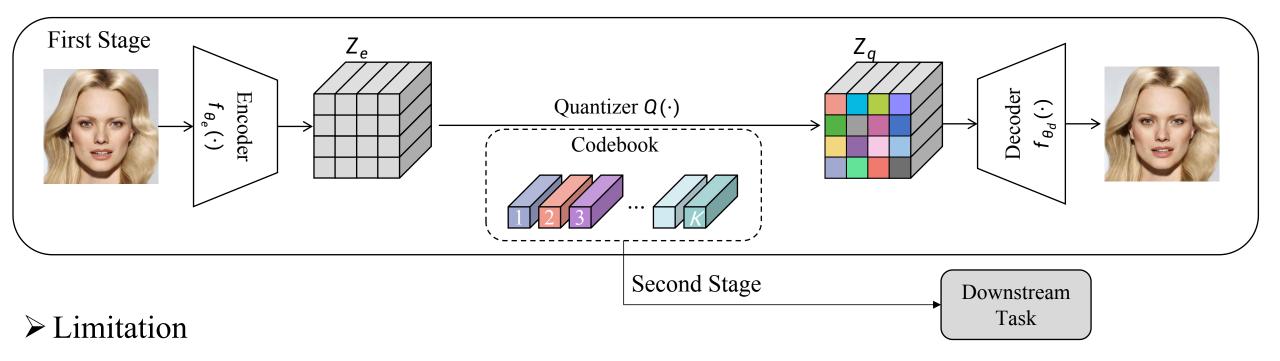
Motivation







➤ Vector Quantization



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

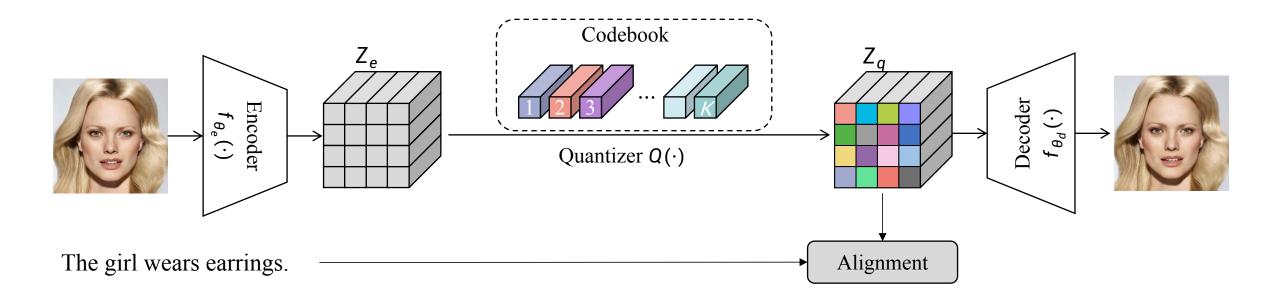






➤ Limitation

• Most methods only focus on learning <u>a single-modal codebook</u>, 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

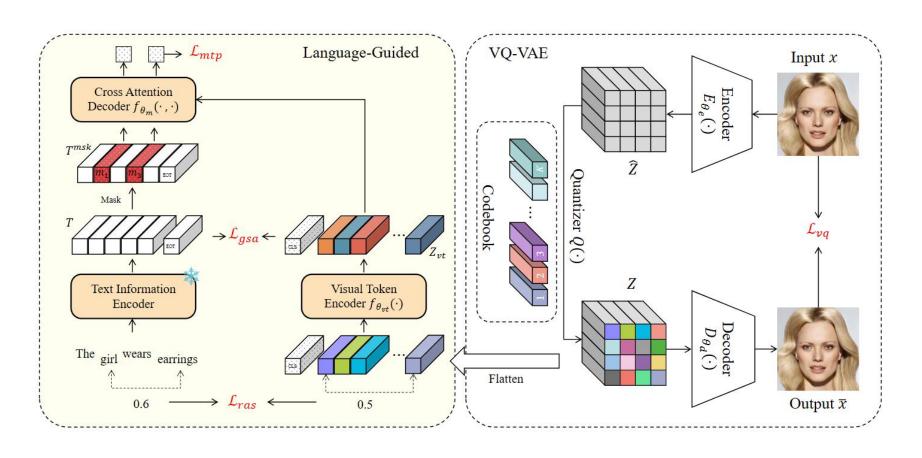
Proposed Method







> 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



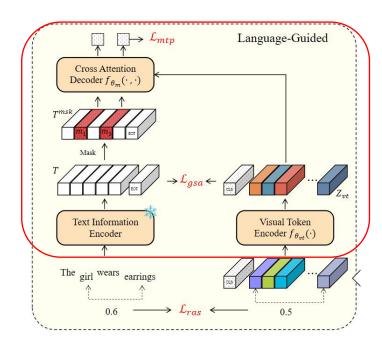




- > Semantic Alignment Module
 - Insight: Considering that **paired image and text data** have consistent semantic information and the missing information of masked data can be completed from the other modality. We propose **global semantic alignment** and **masked text prediction**.

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

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



Proposed Method





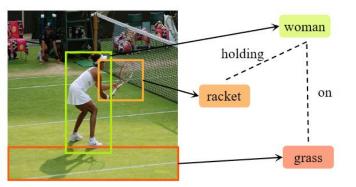


- Semantic Alignment Module Limiting
 - Cannot satisfy more complex reasoning tasks like image captioning and VQA.

Relationship Alignment Module

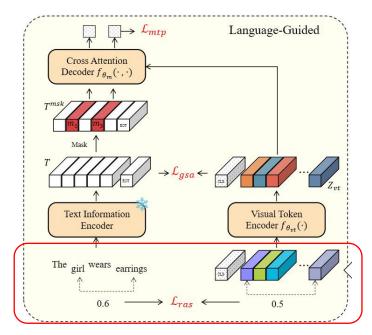
$$e_{z_i} = Z[\underset{e_z \in Z_{vt[1:]}}{argmax} \, s(e_{w_i}, e_z), :], \qquad e_{z_j} = Z[\underset{e_z \in Z_{vt[1:]}}{argmax} \, s(e_{w_j}, e_z), :].$$

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



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").









> 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	Tex	tCaps	Celel	oA-HQ	CU	B-200	MS-	COCO
Models	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







➤ 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)	$+ \mathcal{L}_{qsa}$	23.01	3.39
(iii)	$+\mathcal{L}_{mtp}$	21.54	3.49
(iv)	$+ \mathcal{L}_{mtp} + \mathcal{L}_{ras}$	20.77	3.32
(v)	$+ \mathcal{L}_{mtp} + \mathcal{L}_{qsa}$	20.46	3.34
(vi)	$+\mathcal{L}_{mtp}+\mathcal{L}_{gsa}^{\sigma}+\mathcal{L}_{ras}$	20.35	3.08



this bird has large [mask] with stripes on the ends, a short tail, a large gray and white head, and a long beak with a downward curve.

GT: wings Top-1 word: wings



This woman has bags under [mask], big nose, narrow [mask], and arched eyebrows and is wearing necklace, and heavy makeup. She is [mask].

GT: eyes, eyes, smiling Top-1 word: eyes, eyes, smiling



a small bird with a grey-yellow head and back, a black eye, a bill which is grey on top and yellow on the bottom, and a dull white [mask] and abdomen. GT: breast Top-1 word: breast



This person [mask] mustache, bags under eyes, bushy eyebrows, and big nose. He [mask] young. He [mask] beard.

GT: has, is, has Top-1 word: has, is, has

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



Top-1 retrieved text

a jet black bird with a bright
yellow-orange color on its head and
chest, its beat is short, sharp, and
black



Top-1 retrieved text

The person has mouth slightly open, gray hair, and bags under eyes and is wearing necktie. He is smiling, and chubby.



Top-1 retrieved text

a multicolored bird with an orange
and white base color with black tips
and a red head with a short beak.



Top-1 retrieved text

The man has sideburns, bags under eyes, big nose, and double chin and is wearing necktie. He is smiling.

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
Calab A IIO	Mask-1	99.55
CelebA-HQ	Mask-3	99.24
CLID 200	Mask-1	83.65
CUB-200	Mask-3	80.17







➤ Ablation Study

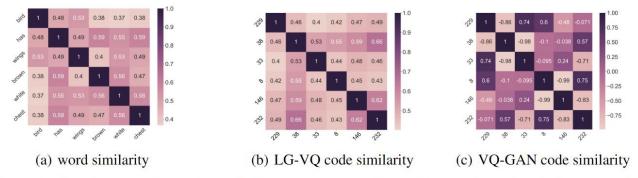


Figure 6: Visualization of words similarity and image codes similarity aligned with the word. We extract some representative words from the text as a demonstration.

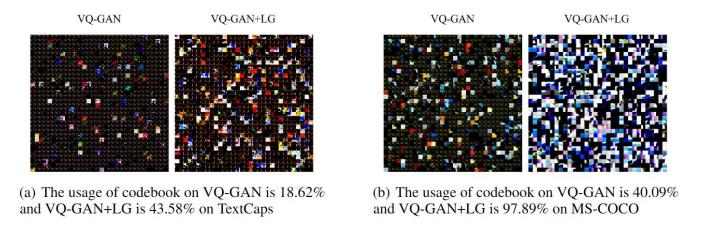


Figure 7: Visualization of the codebook of VQ-GAN and LG-VQ on TextCaps and MS-COCO.

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







> Application



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

Semantic VQ-GAN+LG CVQ+LG Semantic VQ-GAN+LG CVQ+LG Semantic VQ-GAN+LG CVQ+LG

VQ-GAN+LG	VQ-GAN+LG	VQ-GAN+LG	VQ-GAN+LG	VQ-GAN+LG	VQ-GAN+LG
VQ-GAN+LG	VQ-GAN+LG	VQ-GAN+LG	VQ-GAN+LG		
VQ-GAN+LG	VQ-GAN+LG	VQ-GAN+LG			
V					

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

Catting	VQ-V	/AE	VQ-C	JAN	CV	Q
Setting	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







> Application



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

Mask Image VQ-GAN+LG CVQ+LG Mask Image VQ-GAN+LG CVQ+LG Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+LG CVQ+LG

Mask Image VQ-GAN+L

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%







> Application



VQ-GAN: this bird is black and white, and has a gray crown and black breast.

Our: this bird has a white crown and white throat, brown back, and black feet.



VQ-GAN: this bird has a white light red body, a white or brown blue crown and is a brown throat, and also has a very short crenshaw.

Our: this little bird has a yellow belly, wings, brown eyes, long black beak, brown crown and beak.



VQ-GAN: this bird has wings that are black and white with black spots.

Our: a small bird with a blue head, and long, pointed wing bar.



VQ-GAN: this small pter is brown in color and has a red crown.

Our: this bird has wings that are brown with a red crown and a short point orange bill.

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



Q: what is the person with a cowboy hat riding trying to get a cow? GT: horse

VQ-GAN: motorcycle
Our: horse



Q: what are sitting on the counter in different stages of cutting with a knife? GT: carrots

VQ-GAN: pizzas
Our: vegetables



Q: what is the man swinging at a ball on a tennis court? GT: racquet

VQ-GAN: bat
Our: racquet



Q: what are rubbing their heads and necks together?

GT: giraffes

VQ-GAN: elephants

Our: giraffes



Q: what is the color of the dog?

GT: white VQ-GAN: brown Our: white



Q: how many small children that are enjoying a small snack? GT: four VO-GAN: three

Our: four

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

Madal	I	mage Caption	ning	
Model	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.

Catting	V	QA
Setting	Accuracy [↑]	WUPS ↑
VQ-GAN	37.82 ± 0.97	82.06 ± 0.54
VQ-GAN+LG	40.97 ± 0.13	83.56 ± 0.11

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







> Application

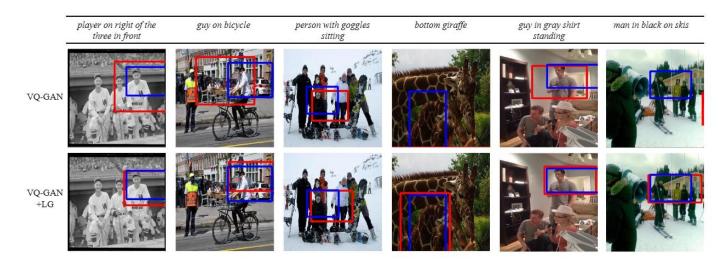


Figure C: Examples of **visual grounding** on refcoco. Blue boxes are the ground-truth, red boxes are the predictions of the model.

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

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

Thanks for Listening!