Towards Improved Text-Aligned Codebook Learning: Multi-Hierarchical Codebook-Text Alignment with Long Text

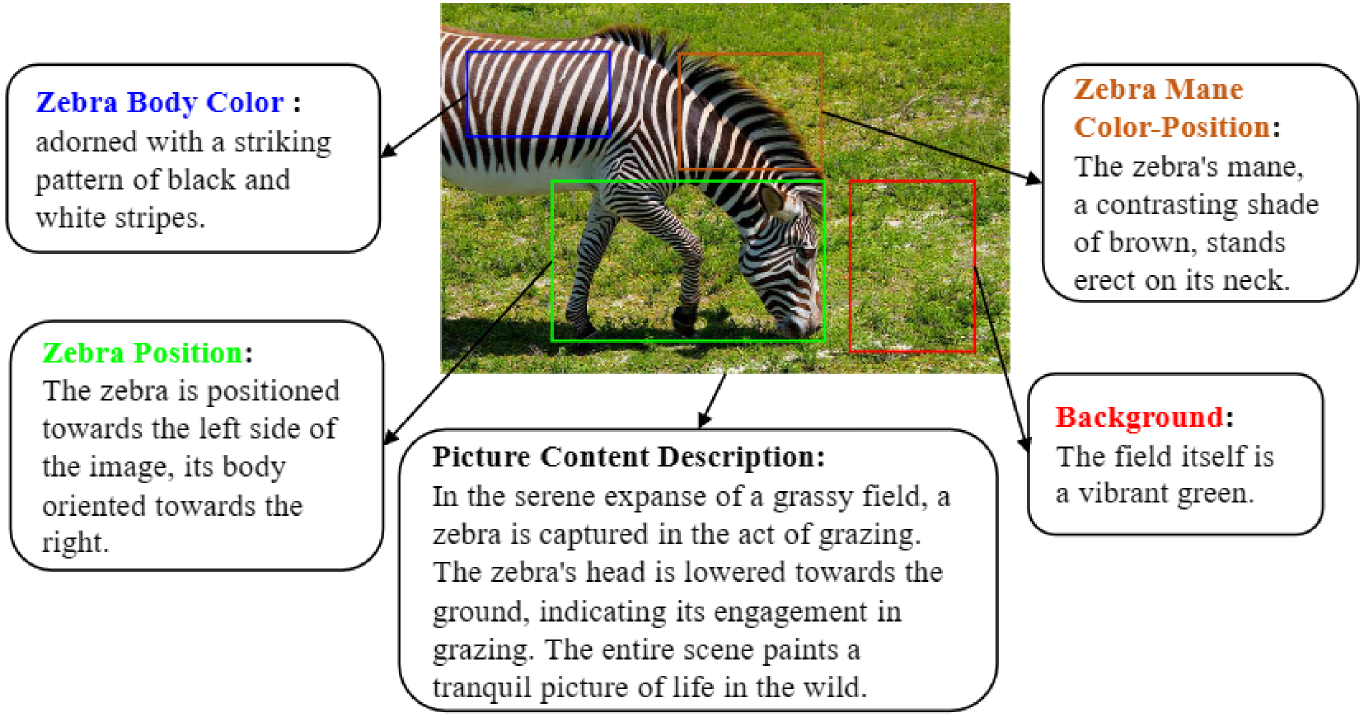
Guotao Liang[1,2], Baoquan Zhang[1]*, Zhiyuan Wen[2], Junteng Zhao[1], Yunming Ye[1], Kola Ye[3], Yao He[3]
[1]Harbin Institute of Technology, Shenzhen [2]Peng Cheng Laboratory, [3]SiFar Company
lianggt@pcl.ac.cn, {baoquanzhang, yeyunming}@hit.edu.cn, 23S051042@stu.hit.edu.cn,
{wenzhiyuan2012, kolaygm, heyao18818}@gmail.com

Highlight

CVPR Nashville JUNE 11-15, 2025

Motivation

Original Caption: A zebra grazing on lush green grass in a field.

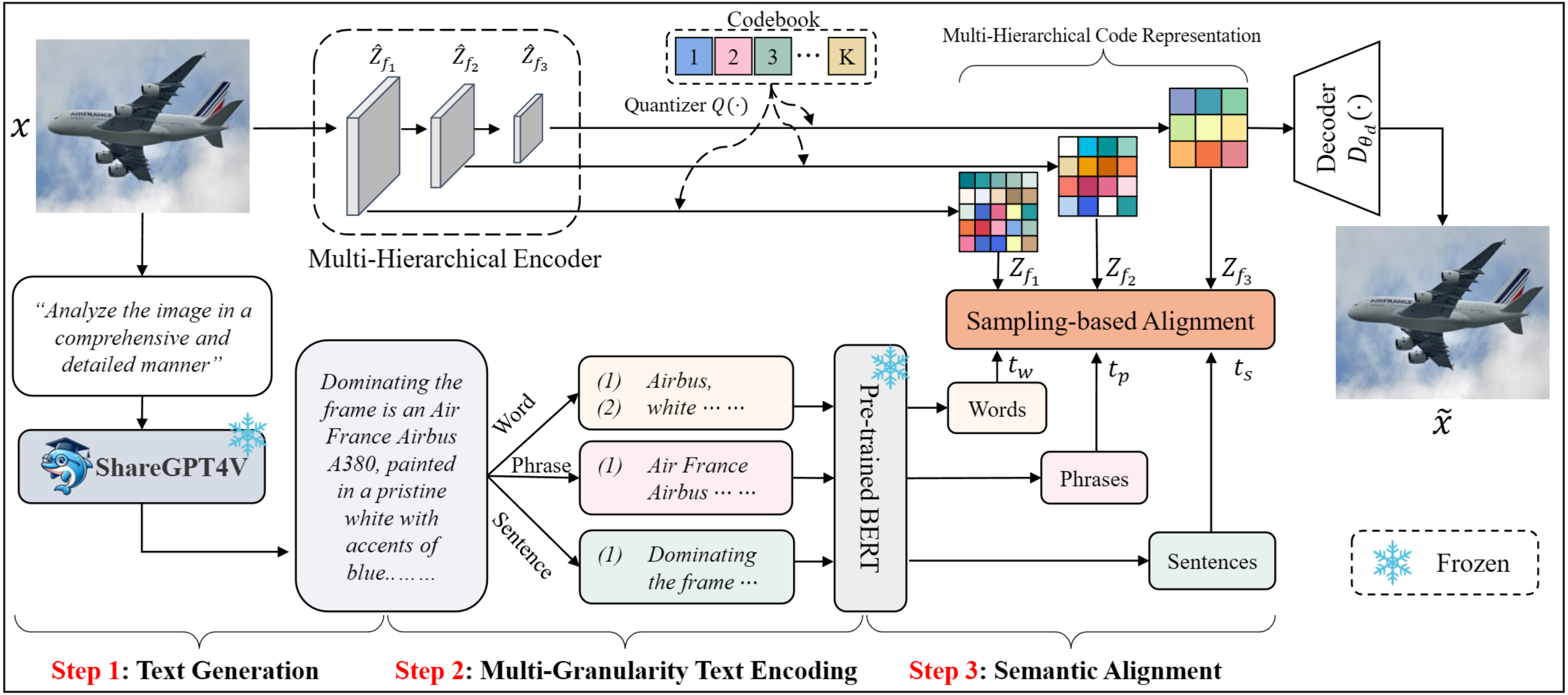


Long Text: In the serene expanse of a grassy field, a zebra, adorned with a striking pattern of black and white stripes, is captured in the act of grazing. The zebra's mane, a contrasting shade of brown, stands erect on its neck, adding to its majestic appearance. The zebra is positioned towards the left side of the image, its body oriented towards the right, as if gazing into the distance. The field itself is a vibrant green.

➢ Challenges

- How to encode long text
- How to align codebook and text

Proposed Method: Text-Augmented VQ



Step 1: Text Generation **Step 2: Multi-Granularity Text Encoding** **Step 3: Semantic Alignment**

Proposed Method: Text-Augmented VQ

- Challenges: $|Z_f| \neq |t|$

We formulate the semantic alignment problem as an ***optimal transport problem*** : Wasserstein distance

Theorem 1 *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous with respect to the Lebesgue measure with Radon–Nikodym density $\rho(x)$. Let $\nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}$ for some $\{y_j\}_{j=1}^{n} \subset \mathbb{R}^d$, $\nu_j \geq 0$ and $\sum_{j=1}^{n} \nu_j = 1$, where δ is Dirac delta function. Then, for any $\epsilon > 0$, there exists a fully connected deep neural network $u(\cdot) : \mathbb{R}^d \to \mathbb{R}$ with sufficiently large width and depth (depending on $\epsilon > 0$) such that the Wasserstein distance between $\nabla u(\mu)$ and ν is less than ϵ, where $\nabla u(\cdot)$ is gradient of $u(\cdot)$.*

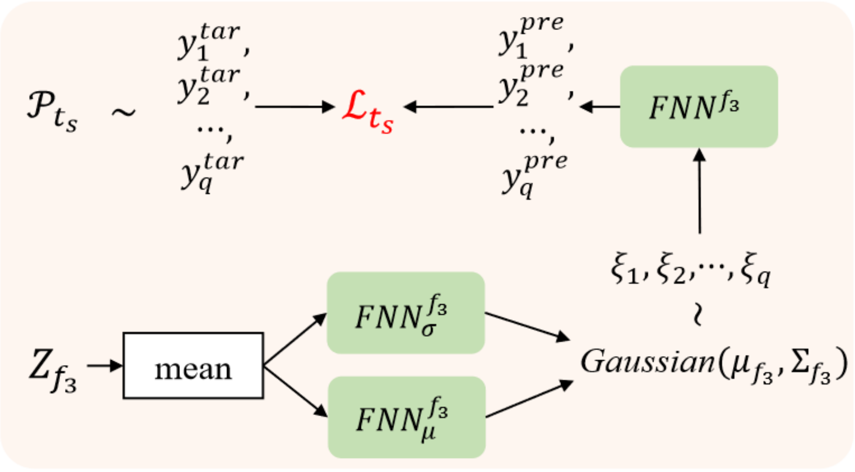
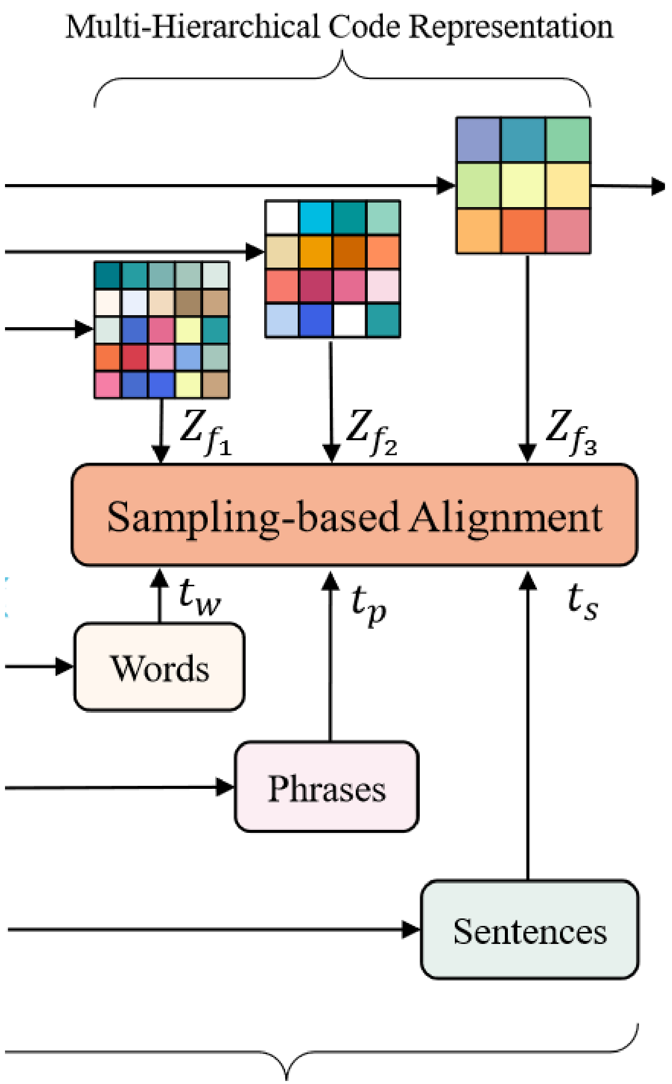


Figure 3. Illustration of the Sampling-based Alignment Strategy.



Step 3: Semantic Alignment

Experiments

Models	Codebook Size	#Tokens	CelebA-HQ FID↓	CUB-200 FID↓	MS-COCO FID↓
VQCT [60]	6207	512	5.02	4.52	9.82
VQ-GAN [13]	1024	256	5.66	5.31	14.45
LG-VQ [29]	1024	256	5.34	4.74	10.72
TA-VQ (Ours)	1024	256	5.03	4.60	10.32
CVQ [65]	1024	256	5.19	4.64	9.94
LG-CVQ [29]	1024	256	4.90	4.40	9.69
TA-CVQ (Ours)	1024	256	**4.71**	**4.03**	**9.65**

Table 1. Results (FID↓) of image reconstruction on CelebA-HQ, CUB-200, and MS-COCO. The best results are highlighted in bold.

Experiments

Model	Visual Grounding Accuracy(0.5)↑
VQ-GAN [13]	9.14
VQCT [60]	9.46
LG-VQ [29]	9.62
TA-VQ	**10.17**

Table 9. Result of visual grounding on refcoco dataset [59].

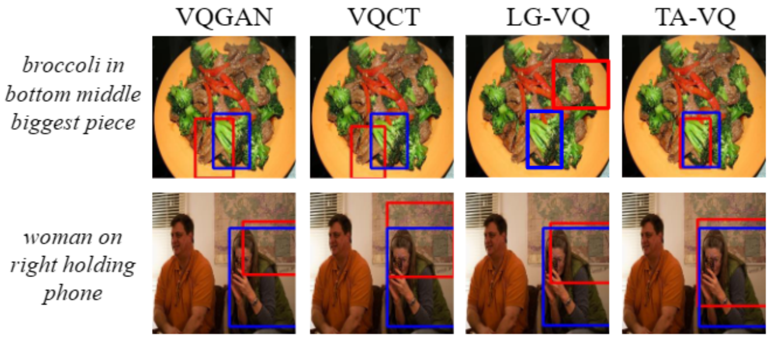


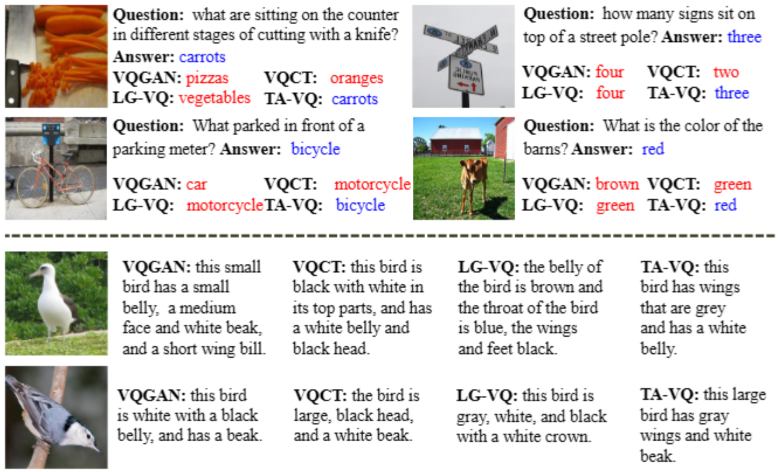
Figure 6. Visualizations for visual grounding. The blue boxes are the ground-truth, red boxes are the model predictions. More Examples are provided in supplementary materials.

Model	Image Captioning			
	BLEU4↑	ROUGE-L↑	METEOR↑	CIDEr-D↑
VQ-GAN [13]	1.29	33.40	24.47	93.62
VQCT [60]	1.38	26.50	24.63	98.22
LG-VQ [29]	1.69	34.73	25.78	102.77
TA-VQ	**1.90**	**35.50**	**27.61**	**109.42**

Table 11. Results of image captioning on CUB-200 datasets.

Setting	VQA	
	Accuracy↑	WUPS↑[55]
VQCT [60]	40.42	82.06
VQ-GAN [13]	37.82	83.22
LG-VQ [29]	40.97	83.56
TA-VQ	**41.56**	**83.77**

Table 12. Results of VQA on COCO-QA [46] dataset.



Experiments

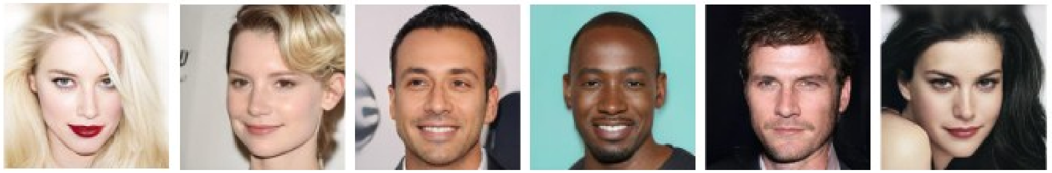


Figure 4. Examples of unconditional generation on CelebA-HQ. More examples are provided in supplementary materials.

Figure 5. Examples of semantic synthesis (row 1), text-to-image (row 2), and image completion (row 3). More examples are provided in supplementary materials.

Model	Image Completion FID↓
VQ-GAN	9.02
LG-VQ	8.14
TA-VQ	**8.04**

Table 6. Result (FID↓) of image completion on CelebA-HQ.

Model	Text-to-Image FID↓
Corgi [67]	19.74
LAFITE [66]	12.54
VQ-GAN [13]	15.29
CVQ [65]	13.23
LG-VQ [29]	12.61
TA-VQ	**11.97**

Table 10. Results (FID↓) of text-to-image on CelebA-HQ.

Model	Unconditional Generation FID↓
DC-VAE [42]	15.8
VQ-GAN [13]	10.2
LG-VQ [29]	9.1
TA-VQ	**8.8**

Table 7. Result (FID↓) of unconditional image generation on CelebA-HQ.

Model	Semantic Synthesis FID↓
VQCT [60]	14.47
VQ-GAN [13]	11.53
LG-VQ [29]	11.46
TA-VQ	**10.74**

Table 8. Result (FID↓) of semantic synthesis on CelebA-HQ.

Thanks for Listening !