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

Supplementary Material

7. Experiment Details

Image Generation: For semantic image synthesis, unconditional generation, and image completion, we follow the default setting of VQ-GAN-Transformer[1]. We use a 16-layer transformer as the backbone network with the head of 16 and dimension of 1024. More settings can be found in [13]. For the text-to-image task, we follow the default setting of VQ-Diffusion[2]. We use a 19-layer transformer as the backbone network with the head of 16 and dimension of 1024. The diffusion step is 100. More settings can be found in [17].

Visual Grounding: The RefCOCO dataset [59] comprises 19,994 images with 50,000 referred objects, each associated with multiple referring expressions, totaling 142,210 expressions. Following [29], we adopt the RefCOCOgumd [40] protocol to partition the dataset, resulting in 42,404 expressions for training, 3,811 for validation, and 3,785 for testing. We follow the default setting of LG-VQ [29], we first apply the VQ model to quantize the image into discrete token representations, which are then processed by a learnable adapter network (*e.g.*, a 2-layer MLP). We use a pre-trained CLIP model [43] to encode the object descriptions, and then concatenate the image token representations with the text embeddings. These concatenated features are fed into a trainable transformer, and the last hidden embeddings are used to predict the object box. We use a 3-layer transformer as the backbone network with the head of 16 and dimension of 512. The learnable adapter network is 2-layer MLP with ReLU activation function. The prediction network of the object box is 3-layer MLP with ReLU activation function.

Visual Text Reasoning: The COCO-QA dataset [46] is automatically generated from captions in the Microsoft COCO dataset [30] and contains 78,736 training questions and 38,948 test questions, based on 8,000 and 4,000 images, respectively. The dataset comprises four types of questions: object (70%), number (7%), color (17%), and location (6%). Each answer is a single-word. We follow the default setting of LG-VQ [29]. we first use the VQ model to quantize the image into discrete token representations, and then feed it into a learnable Adapter network (*e.g.*, 2-layer MLP). Following this, we concatenate image token representation with the text embedding and feed them into a pre-trained language model (*i.e.*, GPT2 [44]). We adjust the output of the language model to adapt to different tasks. The learnable adapter network is 2-layer MLP with ReLU activation function. For the image captioning task, we generate captions by predicting the next token in an autoregressive manner. For the VQA task, we feed the last hidden embedding into 2-layer MLP to predict the answer.

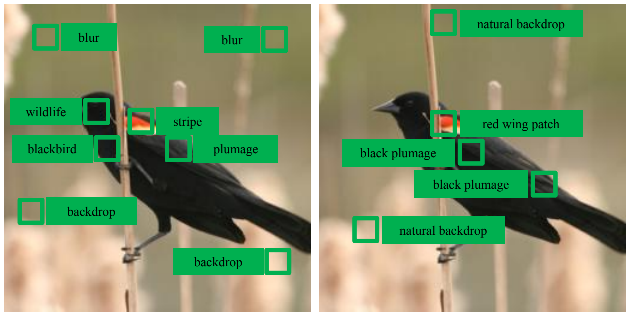


Figure 8. Visualization of codebook-word-phrase alignment.

Table C. Ablation study of f_j on CUB.

	[8, 16, 32]	[2, 8, 16]	[2, 4, 16]	[16, 16, 16]	[4, 8, 16]
FID↓	7.60	4.70	5.96	5.11	**4.60**

8. More Ablation Study

Can our codebook be effectively aligned with the text? In Fig. 8, we visualize an example to demonstrate that our method achieves satisfactory alignment.

Varying the hierarchical grid features f_j. We provide the impact of different f_j on performance in Tab. C.

9. Examples of Origin Caption and Long Text

We provide some examples of origin caption and long text in Fig. 9 for MS-COCO dataset, in Fig. 10 for CelebA-HQ dataset, and in Fig. 11 for CUB-200 dataset.

10. More Examples and Qualitative Results

We provide qualitative comparison of image reconstruction in Fig. 12, unconditional generation in Fig. 13, image completion in Fig. 14, semantic image synthesis in Fig. 15, text-to-image in Fig. 16, visual grounding in Fig. 17, and VQA in Fig. 18.

[1] https://github.com/CompVis/taming-transformers
[2] https://github.com/microsoft/VQ-Diffusion

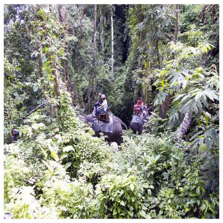
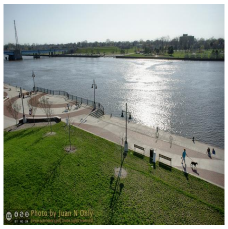
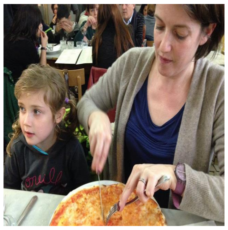
	Origin Caption	Long Text
	SOME PEOPLE IN THE WOODS RIDING TWO ELEPHANTS. They are brave for riding in the jungle on those elephants. Some people who are riding on top of elephants. there are people riding elephants in the middle of a forest. Several elephants in the jungle carrying people on their backs.	In the heart of a dense jungle, two majestic gray elephants are making their way through the lush green foliage. Each elephant carries a passenger on its back, comfortably seated on blue blankets that contrast beautifully with the elephants' gray skin. The elephant leading the way is carrying a passenger dressed in a vibrant red shirt, while the one following closely is carrying a passenger in a cool blue shirt. Their journey through the jungle is framed by an array of trees and bushes, creating a canopy of green above them. The scene is a harmonious blend of nature and human interaction, as these magnificent creatures carry their passengers through their natural habitat.
	A skate park next to a body of water and green park. Sky view of a walkway along the shore of a river. A wide walkway allows for walks along the river. A waterfront walkway and garden area next to a river. some grass and a person in a blue shirt walking a dog and water	The image captures a serene scene of a riverfront park on a sunny day. The park is nestled on the bank of a calm river, its surface reflecting the clear blue sky above. A concrete walkway meanders along the river, inviting visitors for a leisurely stroll. Scattered throughout the park are several trees, their leaves rustling gently in the breeze. A few benches are strategically placed along the walkway, offering rest to weary walkers. A lamppost stands sentinel near the walkway, ready to bathe the park in a warm glow as dusk falls. In the distance, a bridge arches over the river, connecting two parts of the city. The bridge's reflection dances on the water's surface, creating a mirror image that adds to the tranquility of the scene. The sun shines brightly overhead, casting long shadows and bathing everything in a golden light. It's a perfect day for a walk in the park by the river.
	Woman cutting pizza with fork and knife sitting next to young girl A woman and child sitting at a table with a pizza in front of them. A lady and a child are sitting. The lady is cutting pizza pieces. A woman cutting a pizza with a fork and knife.	In the heart of a bustling restaurant, a woman and a young girl share a moment over a shared meal. The woman, clad in a cozy gray cardigan, is engaged in the act of cutting into her pizza with a fork and knife. Her companion, a young girl with blonde hair adorned with a pink bow, mirrors her actions, also using a fork and knife to partake in the feast. The table they're seated at is draped in a pristine white tablecloth, adding an air of elegance to their dining experience. In front of them is a large pizza, generously topped with red sauce and cheese, its aroma wafting through the air. The restaurant around them is alive with activity. Other diners can be seen in the background, each engrossed in their own conversations and meals. Yet, amidst the hustle and bustle, the woman and girl seem to have found a moment of connection over their shared meal. It's a snapshot of everyday life, captured in vivid detail.

Figure 9. Examples of origin caption and long text on MS-COCO dataset.

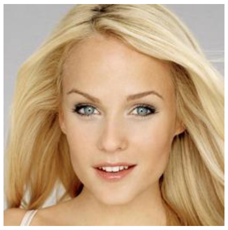
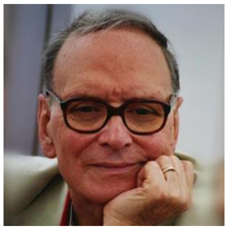
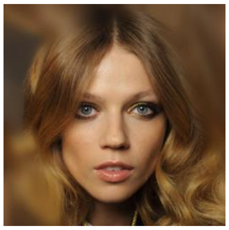
	Origin Caption	Long Text
	She has arched eyebrows. She is smiling, and young and is wearing lipstick. This woman is wearing heavy makeup. She has wavy hair, and mouth slightly open. She has mouth slightly open, arched eyebrows, and wavy hair and is wearing heavy makeup. She is smiling. This person is attractive and has blond hair, mouth slightly open, and arched eyebrows. This person has arched eyebrows, wavy hair, and mouth slightly open. She wears lipstick. She is attractive.	The image captures a close-up of a young woman with striking blue eyes and blonde hair. Her hair, styled in loose waves, cascades down her shoulders, adding a touch of elegance to her appearance. She is wearing a white tank top, which contrasts beautifully with her blonde hair. Her gaze is directed straight at the camera, creating a sense of connection with the viewer. A slight smile graces her face, adding a warm and friendly aura to the overall image. The background is a solid light gray color, which ensures that the focus remains solely on the woman. The simplicity of the background further accentuates the details of the woman's face and hair. The image does not contain any text or other discernible objects. The relative position of the woman to the background suggests she is the main subject of this image. The image does not provide any information about the location or setting. The image is a portrait, focusing on the woman's face and upper body.
	This person has receding hairline. The man has narrow eyes, receding hairline, and eyeglasses. This man has narrow eyes, big nose, receding hairline, and gray hair. The man has big nose, eyeglasses, and narrow eyes. The person has receding hairline.	In the image, there's an older man who is the main subject. He has a bald head and is wearing glasses. His chin is resting on his hand, suggesting a moment of contemplation or deep thought. He is dressed in a beige jacket, which adds a professional or formal tone to the image. The background is a blurred gray color, which puts the focus entirely on the man. There are no other discernible objects or texts in the image. The man's position relative to the background suggests he is standing in front of it. The image does not provide any information about the location or setting. It's a simple yet intriguing portrait of an older man, captured in a moment of quiet reflection.
	She is wearing lipstick. She is young and has brown hair, and wavy hair. This woman is wearing lipstick. She has arched eyebrows. This woman is young and has wavy hair, and brown hair. This young woman has mouth slightly open, and wavy hair. The woman has arched eyebrows. She is young. She is wearing lipstick.	The image captures a close-up of a woman's face, her gaze directed straight at the camera. Her hair, a vibrant shade of blonde, frames her face, adding a touch of warmth to the overall composition. Her eyes, a striking shade of blue, stand out against her complexion, and her lips, painted a soft pink, add a subtle contrast. The background, though blurred, gives the impression of a room with a window, suggesting an indoor setting. The focus on the woman's face and the blurred background create a depth of field effect, drawing attention to her expressive features. There are no discernible texts or other objects in the image. The relative position of the woman to the background suggests she is in the foreground of the scene. The image does not provide any information about the actions of the objects or their precise locations. The image is devoid of any aesthetic descriptions, focusing solely on the factual elements present.

Figure 10. Examples of origin caption and long text on CelebA-HQ dataset.

	Origin Caption	Long Text
	this is a bird with a yellow breast, black face and a small beak. this bird has a white crown, a yellow breast, and a black bill this bird has wings that are brown and has a yellow chest this bird has a bright yellow stomach and a black mark on its head. the bird has a yellow breast, black and white crown, and brown back.	In the heart of a verdant forest, a small bird of vibrant yellow and black hues has found solace on a branch. The bird, with its wings neatly folded at its sides, is perched on the left side of the branch, facing towards the right side of the image. It appears to be in the midst of a meal, holding a small insect delicately in its beak. The branch on which it rests is adorned with green leaves, adding a touch of nature's artistry to the scene. The background is a soft blur of more branches and leaves, creating a sense of depth and continuity within the forest. This tranquil moment captured in time paints a vivid picture of life in the woods.
	the bird is plump with a white chest and has a black bill. this bird is off white with a brown crown and a black eyebrow. this bird is white and brown in color, with a small beak. this small bird has fluffy feathers and coffee colored wings. this bird is white with brown and has a very short beak.	In the serene expanse of a clear blue sky, a Cedar Waxwing bird has found solace on a branch adorned with vibrant red berries. The bird, a picture of tranquility, is facing to the right, its body slightly turned towards us, as if aware of our gaze yet unperturbed by it. The branch it's perched on is laden with red berries, their bright color contrasting beautifully with the bird's brown and green plumage. The waxwing's distinctive black tail feathers are clearly visible, adding to the bird's charm. The image captures a moment of peace and beauty in nature, with the bird and its surroundings in perfect harmony.
	the blue bird is small with long blue and black tail. this bird is mostly blue with a black superciliary and primaries. this bird is blue with black and has a long, pointy beak. this is a small blue bird that has gray primaries and a pointed beak this plump little bird is vibrant blue with black wingbars.	In the image, a vibrant blue bird is the main subject. It's standing on a wooden deck, its body oriented towards the left side of the frame. The bird's beak is open, as if it's about to take a bite, and it's holding a small piece of food in its beak, suggesting it might have been feeding or about to feed itself. The wooden deck beneath the bird adds a rustic charm to the scene. It's not just a plain deck though - there are small rocks scattered around, adding a touch of nature to the man-made structure. The background of the image is blurred, drawing focus to the bird. It appears to be a garden or yard, but the details are indistinct due to the focus on the bird in the foreground. This effect also creates a sense of depth in the image, further emphasizing the blue bird as the focal point. Overall, the image captures a simple yet captivating moment in nature, with the blue bird as its star.

Figure 11. Examples of origin caption and long text on CUB-200 dataset.

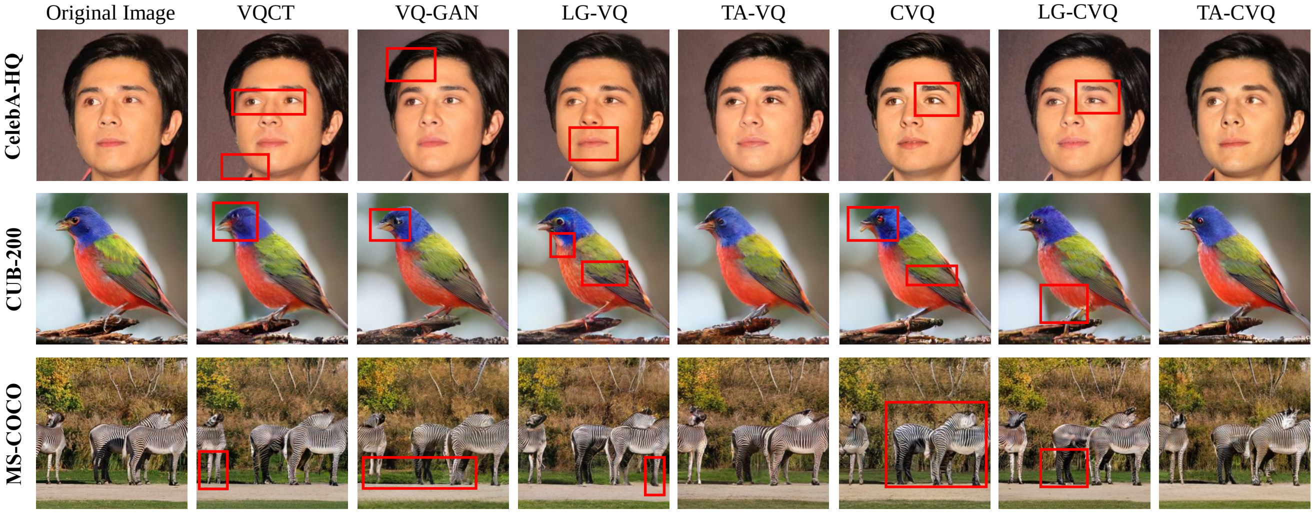


Figure 12. **Reconstructions from different models**. The red-color boxes highlight reconstruction details.

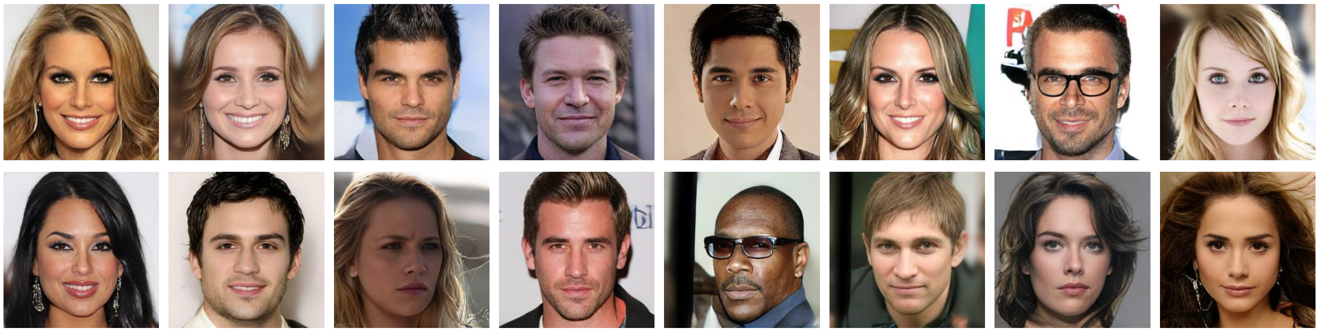


Figure 13. Visualizations for **unconditional generation** on CelebA-HQ.

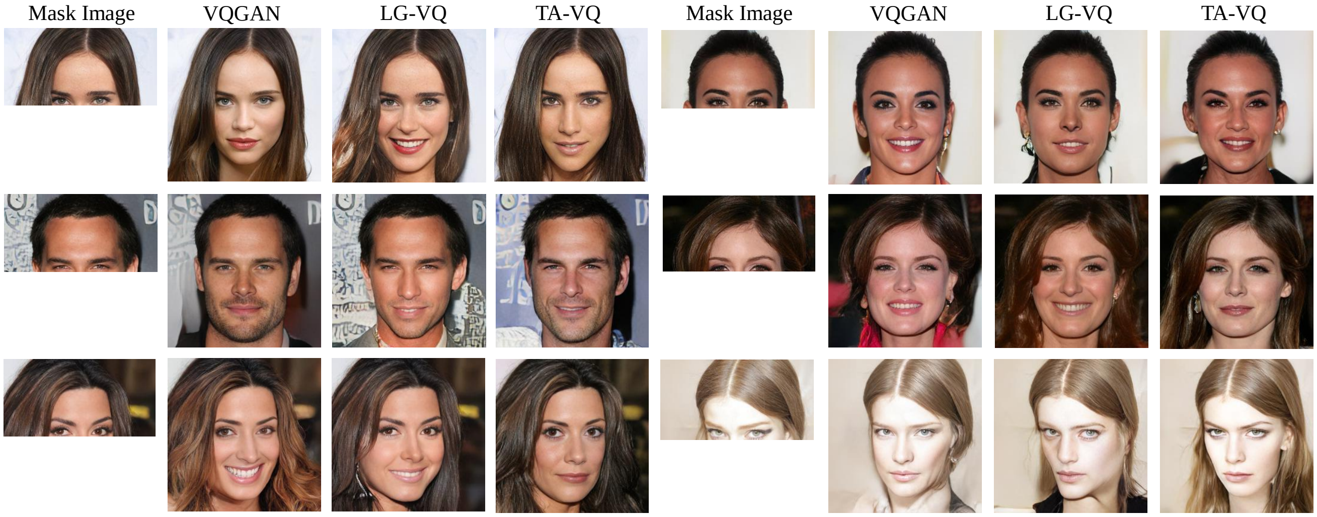


Figure 14. Visualizations for **image completion** on CelebA-HQ.

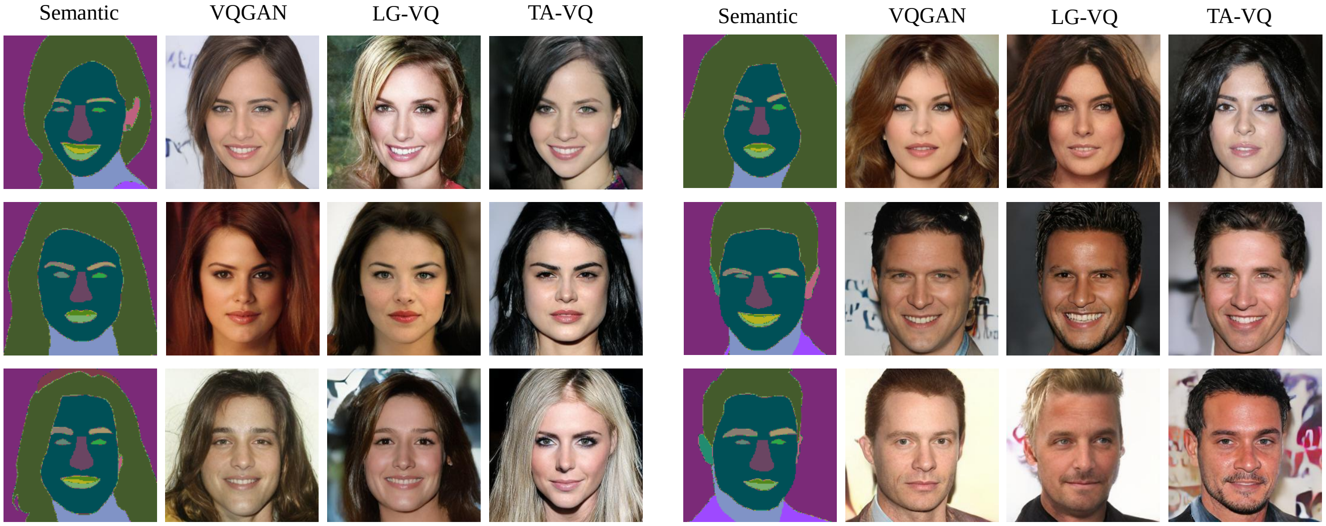
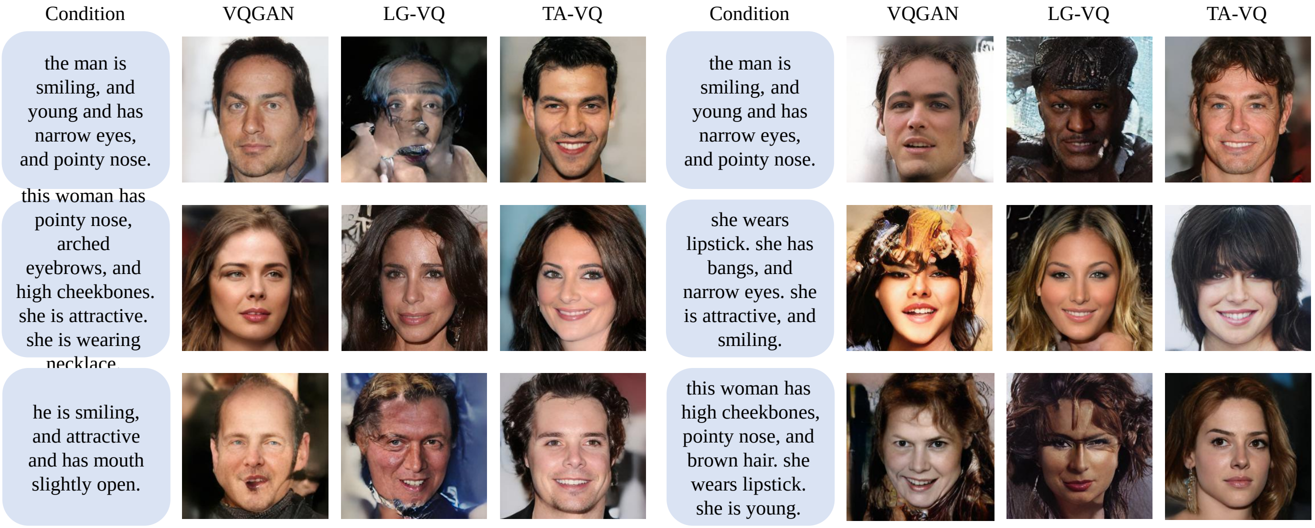


Figure 15. Visualizations for **semantic synthesis** on CelebA-HQ.



the man is smiling, and young and has narrow eyes, and pointy nose.

the man is smiling, and young and has narrow eyes, and pointy nose.

this woman has pointy nose, arched eyebrows, and high cheekbones. she is attractive. she is wearing necklace

she wears lipstick. she has bangs, and narrow eyes. she is attractive, and smiling.

he is smiling, and attractive and has mouth slightly open.

this woman has high cheekbones, pointy nose, and brown hair. she wears lipstick. she is young.

Figure 16. Visualizations for **text-to-image** on CelebA-HQ.

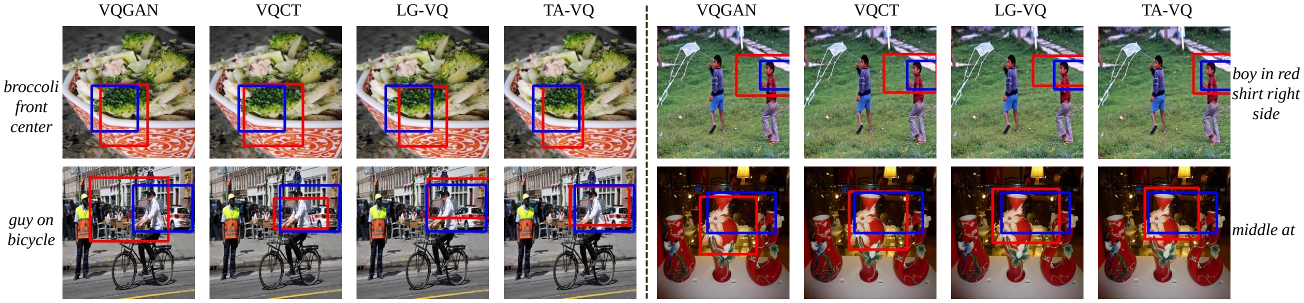
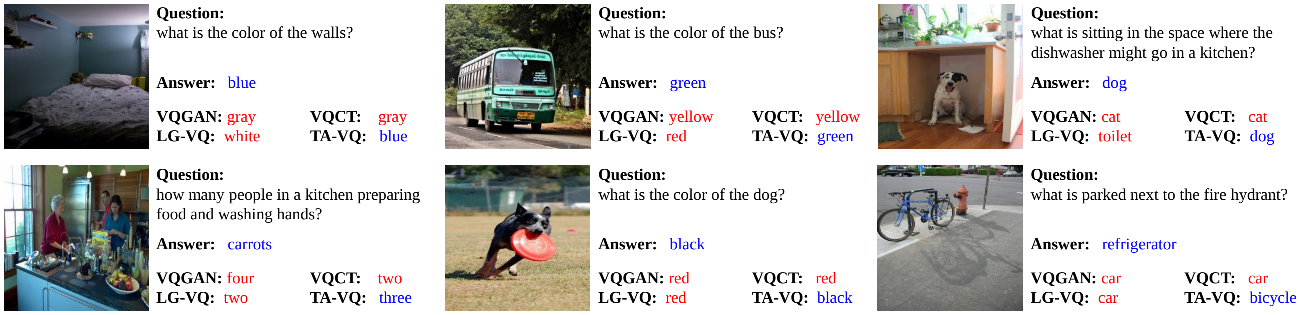


Figure 17. Visualizations for **visual grounding**. Blue boxes are the ground-truth, red boxes are the model predictions.



Question: what is the color of the walls?
Answer: blue
VQGAN: gray **VQCT:** gray
LG-VQ: white **TA-VQ:** blue

Question: what is the color of the bus?
Answer: green
VQGAN: yellow **VQCT:** yellow
LG-VQ: red **TA-VQ:** green

Question: what is sitting in the space where the dishwasher might go in a kitchen?
Answer: dog
VQGAN: cat **VQCT:** cat
LG-VQ: toilet **TA-VQ:** dog

Question: how many people in a kitchen preparing food and washing hands?
Answer: carrots
VQGAN: four **VQCT:** two
LG-VQ: two **TA-VQ:** three

Question: what is the color of the dog?
Answer: black
VQGAN: red **VQCT:** red
LG-VQ: red **TA-VQ:** black

Question: what is parked next to the fire hydrant?
Answer: refrigerator
VQGAN: car **VQCT:** car
LG-VQ: car **TA-VQ:** bicycle

Figure 18. Visualizations for **visual question answering** (VQA).