

A APPENDIX

A.1 COGNITION ALIGNMENT TEMPLATES

- Provide the VQ indices that would represent the essential features of this image, capturing its key visual elements.
- Generate the VQ indices corresponding to the overall appearance and characteristics of this image.
- Imagine a hypothetical encoding process for this image. What would the resulting VQ indices look like, summarizing its content?
- Derive the VQ indices that capture the diversity and complexity of this image’s visual information.
- Formulate the VQ indices for this image, focusing on the patterns and structures that define its visual identity.
- Describe the VQ indices that effectively represent the visual essence of this image.
- In an abstract sense, what would the VQ indices of this image be, encapsulating its visual properties and characteristics?
- Assuming a VQ-based representation, what are the indices that would summarize the key aspects of this image?
- Create the VQ indices for this image, focusing on the fundamental elements that make it unique and recognizable.
- What are the VQ indices that would allow for the reconstruction or recognition of this image, based on its most salient features?
- Construct the VQ indices that would encapsulate the emotional tone and atmosphere conveyed by this image, beyond its literal visual content.
- Imagine the VQ indices as a codebook of visual elements, and devise the sequence that best represents the narrative or story within this image.
- Generate the VQ indices that capture the texture, color palette, and composition of this image, creating a digital fingerprint of its visual style.
- What are the VQ indices that summarize the most prominent shapes, lines, and forms found within this image, forming a visual abstract?
- Derive the VQ indices that would allow for efficient compression and reconstruction of this image, while preserving its key visual details.
- Formulate the VQ indices as a set of vectors that represent the hierarchical organization of visual information in this image, from general to specific.
- Create the VQ indices that reflect the balance of light and shadow, contrast, and saturation in this image, capturing its overall visual mood.
- Assuming a VQ-VAE model, what would the latent space representations (VQ indices) of this image look like, encoding its generative factors?
- Imagine the VQ indices as a set of instructions for a neural renderer to reconstruct this image, specifying its visual features at different levels of abstraction.
- Describe the VQ indices that would enable a machine learning model to recognize or classify this image based on its visual characteristics and patterns.

Table 4: The list of instructions for generating VQ indices of images.

- For each pixel in this image, please provide the corresponding RGB value.
- I would like to know the exact RGB value of every pixel depicted in this image.
- Could you please list the RGB values for all pixels in this image?
- Please give me a detailed breakdown of the RGB values for each pixel in this image.
- I need the RGB value of each and every pixel in this image, can you help?
- Would it be possible to obtain the RGB values for all the pixels in this image?
- Please provide a comprehensive report on the RGB values of each pixel in this image.
- I'm looking for the RGB value of each pixel in this image, could you assist?
- Can you please specify the RGB value for every single pixel in this image?
- I require the RGB values of all pixels in this image, please provide them.
- Please tell me the RGB value of each pixel in this image, from top left to bottom right.
- I'm interested in the RGB values of each pixel, can you give me that information for this image?
- Could you please share the RGB values for each pixel in this image with me?
- I would appreciate it if you could tell me the RGB value of each pixel in this image.
- Please provide the RGB values for all pixels in this image, in a readable format.
- I'm seeking the RGB value of each pixel in this image, can you help me out?
- Would you be able to give me the RGB values for each pixel in this image?
- I need to know the RGB value of every pixel in this image, please provide the details.
- Please tell me the RGB value of each and every pixel that makes up this image.
- I'm looking to obtain the RGB values for all pixels in this image, can you assist me?

Table 5: The list of instructions for predicting pixel RGB values of images.

We report lists of instructions used to predict high-level and low-level visual semantics in Tab 4 and Tab 5. They present the same meaning with natural language variance. For each input image-conversations pair in the 790K visual instruction tuning dataset from LLaVA-Next Liu et al. (2024b), we randomly sample a template from each list and expand it with information on the current image’s dimensional and the region we care about. For templates predicting high-level visual semantics, we also provide a hint regarding the length of the targeted VQ indices based on the specific VQ-VAE used. The expanded templates are then combined with $Target_{VQ}$ and $Target_{PX}$ in Sec 3.2 to create additional conversations, which are inserted at random positions within the original conversations. We report one example used in visual instruction tuning with cognition alignment in Tab 6.

- from: **Human**
<image> What are the colors of the bus in the image? Answer the question with gpt-T-COCO format.
- from: **GPT**
The bus in the image is white and red.

- from: **Human**
What feature can be seen on the back of the bus?
- from: **GPT**
The back of the bus features an advertisement.

- from: **Human**

Imagine a hypothetical encoding process for the upper left quarter of this image. What would the resulting VQ indices look like, summarizing its content? Note that the image resolution is [3, 672, 672], so you should generate 1764 indices

- from: **GPT**

```
<vq_start>3492, 3492, 532, 532, 532, 532, 532, 532, 532, 532, 532, 532, 532,  
532, 532, 532, 532, 532, 532, 532, 532, 532, ....., 2054, 3764, 4015, 570, 2675, 344,  
3369, 580, 2413, 282, 1653, 2826, 1593, 2253, 1988, 3105, 3488, 1455, 1945, 3069,  
1118, 1130, 1130, 2571, 2883, 3160, 1433, 3428, 815, 901<vq_end>
```

- from: **Human**
Is the bus driving down the street or pulled off to the side?
- from: **GPT**
The bus is driving down the street, which is crowded with people and other vehicles.

- from: **Human**
Please tell me the RGB value of each and every pixel that makes up the upper left quarter of this image. Note that the image resolution is [3, 672, 672].
- from: **GPT**
<rgb_start>[[205, 203, 191], [203, 201, 189], [202, 200, 188], [183, 183, 171], [208, 209, 195], [186, 183, 166], ..., [205, 203, 191], [204, 202, 190], [203, 201, 189]]<rgb_end>

17

Stage	I	II	III
Training Data	LLaVA-Next-558k-Pretrain	LLaVAR-442K-OCR	LLaVA-Next-790k-Finetune
Batch Size	256	64	32
Warm-up Ratio	0.03	0.03	0.03
Weight Decay	0.00	0.00	0.00
Trainable Module (learning rate)	perceiver(1e-3), Epigone(1e-3)	perceiver(1e-3) Epigone(1e-3) Denoising Transformer(1e-3) Vision Encoder(2e-6)	perceiver(1e-5) Epigone(1e-5) Denoising Transformer(1e-5) Vision Encoder(2e-6) Language Model(1e-5)

Table 7: Statistic of training data, trainable modules, and hyper-parameters.

A.2 TRAINING STRATEGY.

The training of VLSA consists of three stages.

In Stage I, the focus is on the optimization of the SA-perceiver and the Epigone module, while maintaining the constancy of other parameters. The dataset employed for this initial phase aligns with the 558K instances previously utilized in the pre-training of LLaVA-Next.

Transitioning to Stage II, we further enable the updating of both the denoising transformer and the vision encoder. Given the intrinsic limitations of diffusion models in generating small text in the picture, there exists a risk of the framework neglecting nuanced details (such as tiny textual elements) after the reconstructive training. To counterbalance this shortcoming, the framework undergoes training utilizing the LLaVAR Zhang et al. (2023c) dataset, comprising 422K instructions explicitly designed to bolster OCR capabilities within text-rich images.

Finally, Stage III entails a comprehensive optimization of all parameters, utilizing a fine-tuning dataset containing 790K instructions as leveraged by LLaVA-Next. To facilitate fairness comparisons, the performance of LLaVA-Next pre-trained on 980K instances, including LLaVAR, is also reported.

We train all competitive models on 8 NVIDIA A100 GPUs for 1 epoch in each stage. For stage I, the warm-up ratio, batch size, learning rate, and weight decay are set to 0.03, 256, 1e-3, and 0.00, respectively. For stage II, the warm-up ratio, batch size, and weight decay are changed to 0.03, 64, and 0.00, respectively. The learning rate is set to 1e-3 for the SA-perceiver, the Epigone, and the denoising transformer, and 2e-6 for the vision encoder. As for Stage III, the warm-up ratio, batch size, and weight decay are set to 0.03, 32, and 0.00 separately. The learning rate is set to 1e-5 for the LLM, the SA-perceiver, the Epigone, and the denoising transformer, and 2e-6 for the vision encoder.

We simultaneously leverage the reconstruction loss \mathcal{L}_{rec} and LLM’s autoregressive loss \mathcal{L} in all training phases. The final loss is the summation of \mathcal{L}_{rec} and \mathcal{L} .

A.3 MORE CASE STUDIES

We provide more qualitative comparisons between the original LLaVA-Next and our VLSA on visual writing tasks in Fig 4, and Fig 5.

User:

Examine the image closely and identify as many details as you can. Pay attention to landmarks, architectural features, natural elements, signs, and any other clues that might help determine the location. Based on these details, tell me where this place is.



LLaVA-Next

The image shows a large, multi-story building with a traditional architectural style, featuring multiple levels and a flat roof. The building is illuminated, suggesting it is a significant structure, possibly a landmark or a place of interest. It is situated on the edge of a body of water, which reflects the building's lights and the surrounding landscape. The reflection is visible on the water's surface, indicating that the water is calm. In the background, there are snow-capped mountains, suggesting a mountainous region. The sky is dark, which could indicate either dusk or dawn, or it could be a cloudy day. There are no visible texts or inscriptions in the image. The style of the image is a photograph with a focus on the building and its surroundings."

VLSA

The image depicts the Potala Palace, a renowned landmark located in Lhasa, Tibet. The palace is a large, multi-tiered structure with a distinctive red and white color scheme. The building has a series of rectangular sections with flat roofs, and the facade appears to be made of stone or concrete. The structure is situated on a hillside, and there are trees in the foreground, reflecting on the water below. It is surrounded by a serene lake and snow-capped mountains in the background. The palace is illuminated at night, casting a warm glow that contrasts with the cool blue of the sky. The image is a photograph, and there are no visible texts or distinctive brands that provide additional information about the location. The Potala Palace is a significant cultural and historical site, known for its grandeur and architectural beauty.

Figure 4: VLSA provides more accurate object recognition and detailed description.

User:

Compose an engaging story about this picture, explaining where this place is and highlighting the animals and their activity.



LLaVA-Next

In the vast expanse of the African savanna, a young leopard is on the hunt. The leopard, with its agile body and sharp senses, is crouched low to the ground, stalking its prey. It is in pursuit of a group of three antelopes.

The presence of the leopard indicates that this area is rich in prey, making it a suitable habitat for predators like the leopard. The animals' activity of running and chasing each other showcases the natural behavior and interactions between these species in their shared environment.

VLSA

In the vast expanse of the African savanna, a group of three antelopes are captured in a moment of pure exhilaration. They are running across a grassy field, their hooves pounding the earth as they are chased by a cheetah. The cheetah, with its distinctive black and yellow spots, is in the foreground of the image, its body low to the ground as it dashes towards the right side of the frame. The antelopes, with their long, curved horns, are in hot pursuit. They are spread out across the field, their bodies leaning forward in an attempt to escape from the cheetah. The savanna stretches out behind them, a sea of green grass under a clear blue sky. The scene is a snapshot of life in the wild, a testament to the beauty and power of nature.

Figure 5: VLS provides more detailed, visually-grounded responses than LLaVA-Next.