A. Implementation Details

A.1. Model Configuration

We use Llama-3.1-8B [10] as our base language model due to its strong performance on language tasks and efficient architecture. For visual quantization, we employ a pretrained VQ-GAN model with a codebook size of 8192, which provides sufficient granularity for capturing visual details while maintaining computational efficiency. When applying our visual BPE tokenization, we experiment with extended vocabulary sizes ranging from 4K, 8K to 16K additional tokens to investigate the trade-off between representational capacity and learning efficiency as discussed in Section 4.3.

For priority-guided encoding, we set spatial consistency weight $\alpha = 0.3$ and scaling parameter $\sigma = 2.0$ by default, determined through ablation studies on a validation set. These parameters balance the influence of frequency and spatial consistency in token pair selection.

A.2. Hyperparameters for the VQ-GAN model

The hyperparameters for the VQ-GAN model used in our experiments are shown in Table 5. The embedding dimension of 256 and codebook size of 8192 were chosen to provide sufficient representational capacity while maintaining computational efficiency. The input resolution of 512 allows for capturing fine-grained visual details without excessive memory requirements. We disabled dropout to preserve maximum visual information during the quantization process.

Hyperparameter	Value
embedding dimension	256
codebook size	8192
z_channels	256
resolution	512
dropout	0

Table 5. Hyperparameters for the VQ-GAN model

A.3. Hyperparameters for multi-stage training

The hyperparameters for multi-stage training are shown in Table 6. We carefully designed these parameters to align with the objectives of each training stage. In Stage 1, we use a higher learning rate (1e-3) to efficiently align the newly initialized visual token embeddings. For Stages 2 and 3, we reduce the learning rate (3e-5 and 5e-5 respectively) to prevent catastrophic forgetting while enabling meaningful updates to transformer layers. We increase gradient accumulation in later stages to effectively handle more complex data types. All stages use a cosine learning rate schedule with a 3% warmup period to stabilize training.

Hyperparameter	Stage 1	Stage 2	Stage 3
batch size	1	1	1
gradient accumulation	2	4	4
learning rate	1e-3	3e-5	5e-5
learning schedule	cosine	cosine	cosine
warmup ratio	0.03	0.03	0.03
weight decay	0	0	0
epoch	2	3	3
optimizer	AdamW	AdamW	AdamW
deepspeed stage	2	2	2

Table 6. Hyperparameters for multi-stage training

A.4. Data Details

Following our categorization in Section 3.6.1, we use a diverse set of datasets for different training stages.

- **Foundation Data (FD)**: We use 595K images from CC-3M [50], 558K from LCS [33], and a subset of LAION-2B-en [47] for basic image-caption alignment.
- **Perception Data (PD)**: We incorporate 50.6K samples from RefCOCO [24] and 66.2K from AOKVQA [48] to enhance detailed visual perception.
- **Reasoning Data (RD)**: We utilize 504K general QA entries and 343K reasoning-focused entries from the LLaVA-OneVision Dataset [26].
- **Instruction Data (ID)**: We include 57.3K entries from ShareGPT4o [8], 70K from ALLaVA Inst [7], 180K OCR-related entries from LLaVA-OneVision, and 100K from Infinity-MM [18].

Our curriculum learning approach relies on carefully designed data composition ratios that shift across training stages:

- **Stage 1 (Embedding Alignment)**: $R_{FD} \gg R_{PD} > R_{RD} = R_{ID} = 0$, focusing primarily on foundation data with some perception data
- **Stage 2 (Selective Fine-tuning)**: $R_{FD} \approx R_{PD} > R_{RD} > R_{ID}$, with increased emphasis on perception and reasoning data
- **Stage 3 (Full Fine-tuning)**: $R_{ID} > R_{RD} > R_{FD} \approx R_{PD}$, prioritizing instruction and reasoning data

Specifically, table 7 presents the specific percentage breakdown for each data type across the three training stages.

Training Stage	FD	PD	RD	ID
Stage 1 (Embedding Alignment)	80%	20%	0%	0%
Stage 2 (Selective Fine-tuning)	40%	30%	20%	10%
Stage 3 (Full Fine-tuning)	15%	15%	30%	40%

Table 7. Data composition ratios (%) across training stages

A.5. More analysis on inactive tokens.

To further investigate the token activation mechanisms in Section 4.3, we analyzed token usage across vocabulary sizes and training strategies. Results Table 8 show that less efficient training (where data composition and weight updating misalign with model progression) worsen the issue. This suggests a potential optimization direction: training strategies that properly align with the model's capability progression can better utilize the BPE vocabulary.

	4K vocab	8K vocab	16K vocab
Standard	0.7	3.1	7.9
Reverse curriculum	2.4	7.8	12.3
2-stage training	3.2	9.1	15.6

Table 8. Token usage across vocabulary sizes. The value represent the percentage of inactive tokens.

A.6. Computational Analysis

We've quantified the computational overhead as shown in Table 9. The values are normalized and relative to baseline. For inference, larger vocabularies have minimal impact since token mapping is negligible once established. For training, larger vocabularies require more iterations for mapping learning and embedding expansions, increasing training cost.

	Memory (×base)	**Training** (×base)	**Inference** (×base)
0 (base)	1.00	1.00	1.00
4K	1.15	1.09	1.01
8K	1.27	1.21	1.02
16K	1.57	1.45	1.05

Table 9. Quantified analysis of computational overhead.

A.7. Categorical Comparison

We conducted additional evaluation on *MLLM-Bench*. The results in Table 10 explicitly show how our performances vary across different task categories. While our model shows some decreases in tasks requiring fine-grained details, it demonstrates significant improvements in tasks requiring general comprehension and reasoning.

Category	**Being-VL-0.5**$^+$	**LLaVA-Next**	***tie*** *(equal perf)*
Perception	28	**32**	10
Understanding	**55**	32	23
Applying	**29**	16	15
Analyzing	**53**	38	29
Evaluation	14	**19**	7
Creation	**11**	6	3

Table 10. Evaluation results on *MLLM-Bench*.

B. Case Study

Case 1

Question: Please give me a detailed description of this image.

Being-VL-0.5: In this image, a man wearing a turquoise t-shirt and navy shorts is sitting on a gray bollard with red stripes. To his left, a brown and white dog sits on the brick pavement, connected to the man by a leash. The dog appears calm while the man sits with one leg extended. They are on a brick sidewalk in front of a building with display windows.

Being-VL-0: There is a person sitting with a dog beside him. The person is wearing a blue shirt and shorts. The dog is brown and white. They are in front of a building with windows. The person is sitting on a post, and the dog is sitting on the ground. This appears to be in a city, with a brick floor.

Key Differences: Our method captures more precise visual details (specific colors and features) and clearer spatial relationships between the man and dog in the scene.

Case 2

Question: Please narrate what you observe in this image.

Being-VL-0.5: The image features a white seagull standing on wooden deck planks in the foreground. The bird is positioned in profile, facing right, with its yellow beak and pink feet contrasting against its white and light gray feathers. In the background is an orange and black tour boat navigating through brownish water. The boat has a covered upper area and appears to be carrying passengers. The wooden deck seems to be part of a wharf or landing platform.

Being-VL-0: In the image, there's a seagull standing on a wooden platform. The seagull is white with a yellow beak. In the background, there's a boat on brown water. The seagull is looking in one direction while standing on wooden boards. The boat is orange and black with some people on it.

Key Differences: Our method provides more detailed descriptions of visual elements (seagull's features) and better spatial understanding of the scene components.

 # Case 3

Question: Could you describe what you see in this image?

Being-VL-0.5: This image shows a European urban transportation scene. In the foreground, two cyclists are riding on a dedicated bike lane that runs parallel to tram tracks. The leftmost cyclist wears a checkered shirt and has a bag on his bicycle rack. The scene displays structured transportation zones: pedestrian sidewalk, cycling path, and tram tracks integrated together. In the background stands a large historic building with distinctive architecture. The sky is blue with white clouds.

Being-VL-0: The image shows a city street with people riding bicycles. There are tram tracks on the ground and large buildings in the background. On the left, a person with a checkered shirt is riding a bicycle. The sky is blue with some clouds. This appears to be a European city based on the architecture and transportation setup.

Key Differences: Our method provides more detailed description of key elements and offers clearer understanding of the transportation infrastructure organization.

In this section, we present qualitative examples to demonstrate the enhanced visual understanding capabilities of our approach compared to the frequency-only BPE baseline (Being-VL-0). These cases demonstrate several key advantages of our approach:

- **Semantic Integrity**: Our priority-guided encoding better preserves complete semantic entities (people, animals, vehicles) as coherent token groups, enabling more accurate descriptions of subjects.
- **Spatial Relationship Understanding**: By incorporating spatial consistency in our encoding strategy, our model shows enhanced ability to describe relative positioning of elements within the scene.
- **Fine-grained Visual Detail Recognition**: Our approach better captures small but significant visual details, including colors, patterns, and distinctive features.
- **Structural Pattern Recognition**: The unified token created by our method facilitates stronger recognition of functional structures and their relationships within the scene.

C. Broader Impact

This work advances multimodal understanding through a unified token-based approach, with several potential societal implications. On the positive side, improved visual-language integration could enhance accessibility technologies for visually impaired users, enable more natural human-computer interaction, and support educational applications through better comprehension of multimodal learning materials. Our method's unified representation strategy may also lead to more computationally efficient models, potentially reducing the environmental footprint of multimodal AI systems.

However, like other powerful visual-language models, our approach could be misused to generate misleading content if deployed without proper safeguards. Models with enhanced visual understanding may also inherit or amplify biases present in training data. We encourage thoughtful consideration of these risks in downstream applications, including implementing appropriate content filtering, conducting fairness evaluations across diverse demographics, and establishing clear guidelines for responsible deployment. Furthermore, the growing computational requirements for training such models raise sustainability concerns that should be addressed through efficiency optimizations and responsible resource use.

D. Detail of Priority-Guided Encoding

Algorithm 2 presents the complete version of our priority-guided encoding process (Algorithm 1) in the main manuscript. The key extensions compared to the simplified version include:

- Comprehensive processing of both horizontal and vertical. adjacencies in two-dimensional visual data.
- Detailed calculation procedures for spatial consistency metrics
- Implementation of the diversity filtering mechanism to ensure vocabulary coverage.

E. Licenses

In our code, we have used the following libraries which are covered by the corresponding licenses:

- Numpy (BSD-3-Clause license)
- PyTorch (BSD-3-Clause license)
- Transformers (Apache license)
- Numba (BSD-2-Clause license)

Algorithm 2 Priority-Guided Encoding (Detailed Version)

Input: Quantized training data $\mathcal{C}$, initial vocabulary V, target vocabulary size N_{vocab}, spatial weight α, filtering threshold τ
Output: Extended vocabulary D
$D \leftarrow V$ ▷ Initialize with base vocabulary
while $N_{\text{vocab}} <$ target size **do**
  $P \leftarrow \emptyset$ ▷ Priority scores for token pairs
  for each image I in $\mathcal{C}$ **do**
    for each position (i, j) in I **do**
      Consider horizontal pair $(I_{i,j}, I_{i,j+1})$ if valid
      Consider vertical pair $(I_{i,j}, I_{i+1,j})$ if valid
      Update frequency counts for all considered pairs
    end for
  end for
  for each token pair (a, b) with nonzero frequency **do**
    $F(a, b) \leftarrow \text{count}(a, b) / \sum_{x,y} \text{count}(x, y)$ ▷ Normalized frequency
    $\bar{u}(a, b) \leftarrow (0, 0)$ ▷ Initialize average relative position
    for each occurrence of pair (a, b) in position (i, j, d) **do**
      $u_i(a, b) \leftarrow (0, 1)$ if d is horizontal, $(1, 0)$ if d is vertical
      $\bar{u}(a, b) \leftarrow \bar{u}(a, b) + u_i(a, b)$
    end for
    $\bar{u}(a, b) \leftarrow \bar{u}(a, b)/N_{a,b}$ ▷ Average relative position
    $S(a, b) \leftarrow 0$ ▷ Initialize spatial consistency
    for each occurrence of pair (a, b) with position $u_i(a, b)$ **do**
      $d(u_i, \bar{u}) \leftarrow \exp(-\|u_i - \bar{u}\|^2/2\sigma^2)$ ▷ Spatial similarity
      $S(a, b) \leftarrow S(a, b) + d(u_i, \bar{u})$
    end for
    $S(a, b) \leftarrow S(a, b)/N_{a,b}$ ▷ Average spatial consistency
    $P(a, b) \leftarrow F(a, b) + \alpha \cdot S(a, b)$ ▷ Combined priority score
  end for
  Select top-k pairs by priority: $\{(a_1, b_1), \ldots, (a_k, b_k)\}$
  Filter out pairs with similarity $> \tau$ to existing tokens
  $(a^*, b^*) \leftarrow \arg\max_{i \in \{1,\ldots,k\}} P(a_i, b_i)$
  Create new token $c = (a^*, b^*)$
  $D \leftarrow D \cup \{c\}$
  Update $\mathcal{C}$ by replacing all adjacent occurrences of (a^*, b^*) with c
end while
return D