This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unified Multimodal Understanding via Byte-Pair Visual Encoding

Wanpeng Zhang[1,4] Yicheng Feng[1,4] Hao Luo[1,4] Yijiang Li[2] Zihao Yue[3]
Sipeng Zheng[4] Zongqing Lu[1,4†]

[1]School of Computer Science, Peking University [2]University of California, San Diego
[3]School of Information, Renmin University of China [4]BeingBeyond

Abstract

Multimodal large language models (MLLMs) have made significant progress in vision-language understanding, yet effectively aligning different modalities remains a fundamental challenge. We present a framework that unifies multimodal understanding by applying byte-pair encoding to visual tokens. Unlike conventional approaches that rely on modality-specific encoders, our method directly incorporates structural information into visual tokens, mirroring successful tokenization strategies in text-only language models. We introduce the priority-guided encoding that considers both frequency and spatial consistency, coupled with a multi-stage training procedure based on curriculum-driven data composition. These enhancements enable the transformer model to better capture cross-modal relationships and reason with visual information. Comprehensive experiments demonstrate improved performance across diverse vision-language tasks. By bridging the gap between visual and textual representations, our approach contributes to the advancement of more capable multimodal foundation models. For further details, visit our website: https://beingbeyond.github.io/Being-VL-0.5.

1. Introduction

Multimodal Large Language Models (MLLMs) have made significant progress in integrating information from various modalities[33, 61, 65]. Nevertheless, effectively representing visual information and aligning it seamlessly with texts still remains a core challenge [4, 51, 66]. Conventional approaches primarily generate visual embeddings using modality-specific encoders [12, 23, 69] (such as CLIP [42]), yet recent research suggests that directly discretizing visual information can facilitate a more unified token representation [36, 37, 54, 73]. In a recent study, researchers introduced a visual encoding scheme to vision modality [72], analogous to the byte-pair encoding (BPE) tokenization [25, 41, 49] in LLMs. Previous work [72] theoretically demonstrated that structurally integrating discrete tokens enhances information preservation, thereby enabling the Transformer models [56] to more effectively learn from two-dimensional visual data. Although the preliminary results are promising, translating this theoretical framework into practical multimodal models poses significant challenges: (1) devising sophisticated token encoding strategies beyond simple frequency counting; (2) identifying data composition strategies that maximize visual-language alignment; and (3) designing effective training procedures to fully leverage these BPE-encoded visual tokens. In this paper, we address these challenges by exploring practical implementations and optimization strategies for BPE-based visual tokenization. Our comprehensive analysis highlights key factors that significantly influence performance across modalities and proposes corresponding optimization.

Specifically, we introduce a priority-guided encoding scheme that accounts for both co-occurrence frequency and spatial consistency of visual patterns. This novel approach generates semantically enriched visual tokens, thereby preserving the inherent structural information in images more effectively. Moreover, we propose data composition strategies that leverage a curriculum-based approach tailored for optimizing multimodal learning, explicitly designed to complement the visual BPE tokenization process. Unlike conventional visual encoders [42, 67], which extract visual features through a single forward pass [60, 68], visual BPE tokenization generates hierarchical representations where encoded tokens progressively capture increasingly intricate visual patterns. Given this hierarchical property, an aligned learning curriculum is necessary; tokens must first acquire foundational semantic meanings through elementary visual-text associations before advancing toward more sophisticated visual reasoning tasks. Consequently, a multi-stage training procedure is further derived with stage-specific parameter freezing strategies: (1) embedding alignment with frozen language model parameters, (2) selective fine-tuning

†Correspondence to Zongqing Lu <zongqing.lu@pku.edu.cn>.

of early transformer layers, and (3) full model fine-tuning. This incremental training approach facilitates more effective cross-modal knowledge transfer while maintaining the linguistic capabilities inherent to the foundation model.

We conduct experiments to demonstrate that our proposed optimization achieve substantial performance improvements compared to baselines. In addition to quantitative improvements, we provide detailed analyses of token patterns, computational efficiency, and qualitative improvements in visual comprehension capabilities. Our contributions are summarized as follows:

- A discrete tokenization approach that preserves structural information in images through priority-guided encoding, providing unified visual representation for MLLMs.
- A progressive training strategy that combines stage-specific parameter freezing, allowing our trained model **Being-VL-0.5** to build better visual understanding.
- Empirical evidence across multiple benchmarks and extensive analysis that validates the effectiveness of our unified encoding methods and elucidates the reason why it advances visual-language understanding.

2. Related Work

2.1. Representation Methods in MLLMs

The core challenge of MLLMs lies in effectively representing and integrating information from different modalities. The research community has primarily developed two representative approaches: methods based on continuous representations and methods based on discrete tokens.

Continuous token representation methods have emerged as the mainstream paradigm, with representative works such as LLaVA [32, 33], Emu [52, 58], DeepSeek-VL [35, 59], and Qwen-VL [1, 2, 57]. These models use pre-trained visual encoders (*e.g.*, CLIP [42]) to map images into continuous representations, which are then aligned with language models through projection layers [15, 70]. Although this approach has achieved significant results, it faces two fundamental challenges: first, the modality gap between encoders and language models [5, 22], wherein the high-dimensional continuous features provided by visual encoders differ from the discrete token representations expected by language models; second, the information bottleneck problem, wherein images compressed through visual encoders may lose substantial detailed information, particularly affecting low-frequency visual patterns [43]. Recent analysis [3, 21, 62] indicates that this architectural design makes models prone to visual hallucinations, generating descriptions even when relevant visual information is absent.

Discrete token representation methods have gained increasing attention in recent years [36, 37, 54, 73]. These methods typically use vector quantization models (such as VQ-GAN [11] or VQVAE [45, 55]) to discretize images into a series of tokens, thereby enabling visual information to be processed in a form similar to text. The advantage of this approach lies in its ability to process multimodal inputs in a unified manner, thus reducing inter-modal differences. However, existing discrete token methods also have significant limitations: most works simply quantize images into fixed-size tokens without considering the semantic structure of visual content; information loss during quantization may lead to the loss of fine-grained visual details; and the unbalanced representation of key visual concepts in the token space can result in limited recognition capability for certain visual patterns.

These methods differ significantly from tokenization strategies used in natural language processing. In the NLP domain, algorithms such as BPE form semantically meaningful tokens by combining frequently co-occurring character sequences [14], a strategy that has proven beneficial for Transformer learning [39, 40]. Research by [44] demonstrates that for Markov sequence data, even smaller Transformer models can significantly reduce prediction loss when using appropriate tokenization strategies.

2.2. BPE Tokenization for Visual Data

Extending BPE tokenization to the visual domain is an emerging research direction [6, 46]. The core idea is to form a more semantical visual vocabulary by combining frequently co-occurring patterns based on quantized images. Being-VL-0 [72] first proposed a theoretical framework for applying BPE to quantized visual data, demonstrating that for two-dimensional data conforming to specific generative processes, using visual BPE can significantly reduce model prediction loss. Their analysis indicates that compared to unigram models based on independent and identically distributed assumptions, BPE tokenization can more effectively capture structured patterns in visual data.

However, significant challenges remain in the transition from theory to practice. The implementation of Being-VL-0 [72] uses primarily frequency-based encoding without considering spatial relationships that are crucial in visual data. The relationship between vocabulary size and performance, as well as strategies that balance frequency with spatial consistency, remain underexplored. Besides, designing effective training procedures that leverage the unique properties of BPE visual tokens is also important. Research on parameter freezing strategies [17, 53] suggests that different settings at various training stages can impact the performance, but these findings have not been systematically applied to visual BPE training. Our work addresses these gaps by exploring enhancement strategies for visual BPE tokenization across vocabulary construction, data composition, and training procedures. Through these contributions, we aim to advance the practical application of BPE visual tokenization in multimodal foundation models.

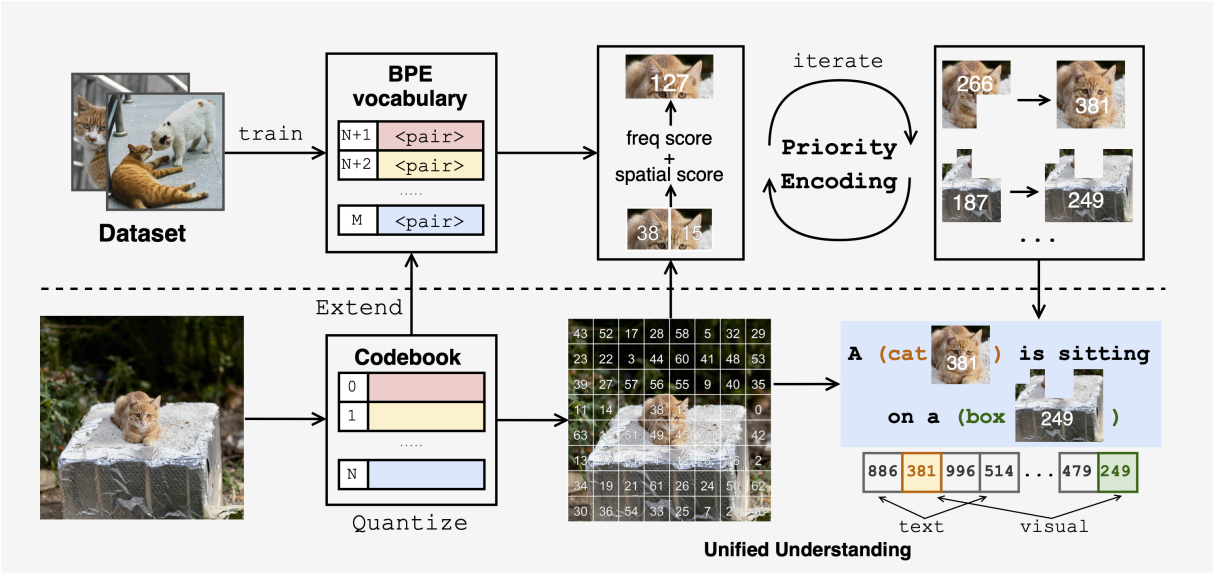

Figure 1. Overview of our framework. The upper part shows the BPE vocabulary construction process: starting from a training dataset, we identify candidate token pairs and apply our priority scoring mechanism (combining frequency and spatial consistency) to iteratively extend the vocabulary with new encoded tokens. The lower part illustrates the MLLM pipeline: an input image is first quantized using a VQ-GAN codebook, then encoded using the trained vocabulary. The resulting visual tokens (*e.g.*, token `381` representing the whole "cat" and `249` representing the whole "box") are seamlessly integrated with text tokens to form a unified sequence for multimodal understanding.

3. Method

In this section, we first introduce key concepts and notation for our visual tokenization framework. Then we present the detailed training pipeline for our **Being-VL-0.5** model.

3.1. Visual Tokenization Pipeline

The core of our approach involves transforming visual information into token sequences that can be processed by language models. This transformation occurs through a pipeline with the following key steps:

Vector Quantization (VQ). Given an image I divided into patches, we first apply a VQ model with a codebook $V = \{v_1, v_2, ..., v_K\}$ containing K entries. This process maps each image patch to its closest codebook entry, producing a grid of discrete indices $Q(I) = \{q_{i,j}\}_{i,j=1}^{h,w}$, where h and w represent the height and width of the patch grid.

BPE Visual Tokenization. The BPE visual tokenization process, denoted as $T(Q(I))$, identifies frequently co-occurring patterns in the quantized representation and encodes them into new tokens, creating an extended vocabulary $D = V \cup V_{ext}$, where V_{ext} represents the set of encoded tokens. Formally, the tokenization function $T : Q(I) \rightarrow (t_1, t_2, ..., t_n)$ maps the quantized image to a sequence of tokens $t_i \in D$, where $n \leq h \times w$ due to the encoding of multiple indices into single tokens.

This approach parallels BPE tokenization in text processing but adapts it for two-dimensional visual data, incorporating spatial relationships that are unique to the visual modality.

3.2. Token-Based MLLM

Unified Token Sequence. We formalize our token-based multimodal language model as a conditional probability distribution over token sequences. Given a text input $X = (x_1, x_2, ..., x_m)$ and an image input I, the model processes the combined sequence and generates output tokens autoregressively. The combined input consists of special tokens marking the beginning and end of the tokenized image representation:

$$S = (x_1, \ldots, x_m, [\text{BOI}], t_1, \ldots, t_n, [\text{EOI}], \ldots), \quad (1)$$

where [BOI] and [EOI] are special tokens indicating the image boundary in the sequence.

Autoregressive Modeling. The model defines a probability distribution over the next token given all previous tokens:

$$p(s_i|s_{<i}) = \text{Softmax}(f_\theta(E(s_1), E(s_2), ..., E(s_{i-1}))), \quad (2)$$

where f_θ represents the transformer-based language model with parameters θ, and $E(s_i)$ represents the embedding of token s_i. For model generation, tokens are sampled sequentially:

$$\hat{s}_i \sim p(s_i|\hat{s}_1, \hat{s}_2, ..., \hat{s}_{i-1}, x_1, x_2, ..., x_m, t_1, t_2, ..., t_n). \quad (3)$$

Training Objective. The model is trained to minimize the negative log-likelihood of the target sequence $Y = (y_1, y_2, ..., y_l)$ conditioned on the input:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(X,I,Y)\sim\mathcal{D}} \left[\sum_{i=1}^{|Y|} \log p_\theta(y_i | y_{<i}, X, T(Q(I))) \right], \tag{4}$$

where $\mathcal{D}$ represents the training dataset containing input-output triplets.

3.3. Framework Overview

As formalized in the preliminaries above, our token-based multimodal model processes visual information through a pipeline that includes vector quantization $Q(\cdot)$ and visual tokenization $T(\cdot)$ before integration with text tokens in a language model. Our optimizations target each component of this pipeline while considering their interdependencies.

Specifically, our framework enhances this pipeline through: (1) Enhanced Vocabulary Construction; (2) Optimized Data Composition Strategies; (3) Efficient Multi-Stage Training. These components work together to improve the model's ability to process visual information, align visual and textual representations, and generate accurate responses to multimodal inputs.

Figure 1 illustrates the overall framework, which operates in two phases. In the vocabulary construction phase (upper part), we train a BPE vocabulary using our priority-guided encoding scheme that considers both co-occurrence frequency and spatial consistency. This approach creates an extended vocabulary that effectively captures meaningful visual patterns beyond what frequency-only methods can achieve. In the application phase (lower part), input images are first quantized through a VQ-GAN model, then encoded using the trained vocabulary. The resulting visual tokens are integrated with text tokens to form a unified sequence for the language model.

3.4. Vocabulary Construction

The effectiveness of token-based representations hinges on creating a vocabulary that efficiently captures meaningful patterns in the data. Our enhanced vocabulary construction method extends beyond simple frequency-based encoding to incorporate structural information specific to visual data, while relying exclusively on image statistics without reference to textual information.

In standard BPE tokenization, token pairs are encoded based solely on their co-occurrence frequency. While effective for text, this can be suboptimal for visual data where spatial relationships play crucial roles. We introduce a priority function $P(a, b)$ that guides the encoding process with a more comprehensive evaluation of candidate token pairs:

$$P(a, b) = F(a, b) + \alpha \cdot S(a, b). \tag{5}$$

Algorithm 1 Priority-Guided Encoding

Input: Quantized training data $\mathcal{C}$, initial vocabulary V, target vocabulary size N_{vocab}
$D \leftarrow V$
while $N_{\text{vocab}} <$ target size **do**
  Compute $F(a, b)$ and $S(a, b)$ for all adjacent token pairs in $\mathcal{C}$
  Calculate priority $P(a, b) = F(a, b) + \alpha \cdot S(a, b)$ and select top-k pairs by priority: $\{(a_1, b_1), \ldots, (a_k, b_k)\}$.
  $(a^*, b^*) \leftarrow \arg\max_{i\in\{1,\ldots,k\}} P(a_i, b_i)$
  Create new token $c = (a^*, b^*)$
  $D \leftarrow D \cup \{c\}$
  Update $\mathcal{C}$ by replacing all adjacent occurrences of (a^*, b^*) with c (horizontally or vertically).
end while
return D

This function combines two key factors:

- **Co-occurrence frequency** $F(a, b)$: The raw frequency with which tokens a and b appear adjacent to each other, normalized across the training corpus of quantized images.
- **Spatial consistency** $S(a, b)$: A measure of how consistently the token pair maintains spatial relationships across different images, calculated as:

$$S(a, b) = \frac{1}{N_{a,b}} \sum_{i=1}^{N_{a,b}} d(u_i(a, b), \bar{u}(a, b)). \tag{6}$$

Here, $u_i(a, b)$ represents the relative positioning of tokens a and b in instance i, $\bar{u}(a, b)$ is the average relative positioning across all instances, $N_{a,b}$ is the number of co-occurrences, and the spatial distance function is defined as:

$$d(u_1, u_2) = \exp\left(-\frac{\|u_1 - u_2\|^2}{2\sigma^2}\right), \tag{7}$$

where $\|u_1 - u_2\|$ is the Euclidean distance between the two relative position vectors, and σ is a scaling parameter that controls how quickly similarity decreases with distance. This formulation ensures that token pairs with consistent spatial relationships across images receive higher scores.

The weighting parameter α controls the relative importance of each factor and is determined through cross-validation on a development set. This dual-factor approach ensures that encoded tokens represent not only frequent patterns but also spatially consistent visual structures. It's important to emphasize that this entire process relies exclusively on image data, with no dependency on textual information or captions. The algorithm identifies patterns based solely on the statistical and structural properties of the visual content. Algorithm 1 outlines our priority-guided encoding process with diversity filtering. For complete details, please refer to Algorithm 2 in the Appendix.

3.5. Model Expanding

Once we have constructed our enhanced visual vocabulary D, we need to expand the pre-trained language model to accommodate these new tokens. This expansion process serves as the bridge between our tokenization approach and the subsequent model training. Given a pre-trained language model M with a text vocabulary V_{text} and corresponding embedding weights $E_{\text{text}} \in \mathbb{R}^{|V_{\text{text}}| \times d}$ (where d is the embedding dimension), we expand the embedding layer to incorporate the visual vocabulary D:

$$E_{\text{expanded}} = \left[E_{\text{text}}, E_{\text{visual}}\right] \in \mathbb{R}^{(|V_{\text{text}}|+|D|) \times d}, \tag{8}$$

where $E_{\text{visual}} \in \mathbb{R}^{|D| \times d}$ represents the newly added embeddings for visual tokens. This expansion increases the model's vocabulary size from $|V_{\text{text}}|$ to $|V_{\text{text}}| + |D|$.

For the initialization of the newly added visual token embeddings, we adopt He initialization [20], *i.e.*, $E_{\text{visual}}[i, j] \sim \mathcal{N}(0, \sqrt{2/d_l})$, where d_l is the embedding dimension. This initialization helps maintain proper signal magnitude throughout the network, preventing vanishing or exploding gradients during the early stages of training.

The process is illustrated in the top part of Figure 2. We first extend a VQ token embedding corresponding to the VQ-GAN codebook, and then extend a BPE token embedding corresponding to the BPE vocabulary. The extended weights from these two components constitute the initialization of the visual part. By default, we standardize both dimensions of them to 8K, corresponding to an 8K VQ codebook and an 8K BPE vocabulary, respectively.

3.6. Multi-Stage Training

With the expanded model supporting both text and visual tokens, we now introduce our multi-stage training pipeline. Our approach combines strategic data composition with stage-specific parameter freezing to maximize the effectiveness of visual tokenization. Figure 2 illustrates the overall process.

3.6.1. Curriculum-Based Data Composition

We develop a curriculum-based data composition strategy, which is designed to complement the BPE visual tokenization. Unlike traditional visual encoders that extract features in a single forward pass, BPE visual tokenization creates a hierarchical representation where encoded tokens incrementally capture more complex visual patterns. This hierarchical nature of BPE tokens requires a matching learning curriculum: tokens must first establish basic semantic meanings through simple visual-text associations before they can be effectively utilized for complex reasoning.

Specifically, we categorize our training data into four types based on complexity: Foundation Data (basic image-caption pairs), Perception Data (detailed visual attributes), Reasoning Data (complex visual QA), and Instruction Data

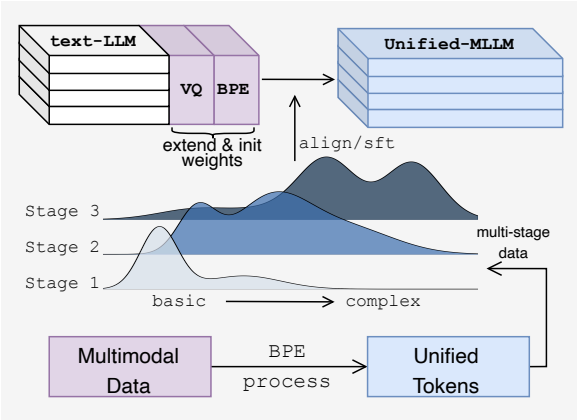

Figure 2. Illustration of our multi-stage training strategy. The bottom part shows multimodal data being processed into unified tokens. The middle curves represent the data distribution across training stages, showing a gradual shift from basic concepts (Stage 1) to more balanced (Stage 2) and finally complex reasoning-focused data (Stage 3). The top part illustrates how we expand a text-only LLM to incorporate visual capabilities through weight initialization and alignment, resulting in a unified multimodal model. This curriculum approach mirrors the hierarchical nature of BPE visual tokens, which progressively capture more complex visual patterns throughout the learning process.

(multi-turn interactions). Details of all types of datasets can be found in the Appendix A.4.

We employ a curriculum that progressively shifts emphasis across training stages, formalized as a composition ratio $R_i(s) = w_i(s)/\sum_j w_j(s)$, where $R_i(s)$ represents the proportion of data type i at stage s with corresponding weight $w_i(s)$. Each stage employs distinct composition ratios, transitioning from foundation-heavy in early stages to instruction-focused in later stages. This curriculum ensures that visual tokens first establish basic meanings before tackling complex reasoning tasks.

This stage-specific curriculum ensures that the model first develops strong foundational visual understanding before focusing on more complex tasks, aligning with the parameter freezing strategy of each stage.

3.6.2. Progressive Parameter Unfreezing

Complementing our data curriculum, we implement stage-specific parameter freezing to control which components of the model are updated at each stage. The parameter update can be expressed as:

$$\theta_{s+1} = \theta_s - \eta \cdot M_s \odot \nabla_\theta \mathcal{L}, \tag{9}$$

where θ_s represents the model parameters at stage s, η is the learning rate, $\nabla_\theta \mathcal{L}$ is the loss gradient, $\odot$ denotes element-wise multiplication, and M_s is a binary mask that determines which parameters are updated during stage s.

Model	Params	Tokenizer	Benchmark					
			VQAv2	MMBench	MME-P	SciQA-IMG	POPE	VizWiz
Continuous Embedding-Based Models								
InstructBLIP	7B	Continuous	75.2	38.3	1212.8	60.5	81.5	34.5
LLaVA-1.5	7B	Continuous	78.5	64.3	1510.7	66.8	85.9	50.0
mPLUG-Owl2	8B	Continuous	79.4	64.5	1450.2	–	86.1	54.5
HyperLLaVA	7B	Continuous	79.1	65.9	1481.2	**70.4**	86.3	51.9
ShareGPT4V	7B	Continuous	80.6	68.8	**1567.4**	68.4	–	57.2
VILA-1.5	8B	Continuous	80.9	**72.3**	–	–	84.4	**58.7**
LLaVA-Next	7B	Continuous	**81.8**	67.4	1519.0	70.1	**86.5**	57.6
Discrete Token-Based Models								
Chameleon	7B	Discrete	56.2	37.3	1297.5	–	78.2	46.4
Being-VL-0	8B	Discrete	60.6	44.0	1316.2	64.3	81.3	48.2
Unified-IO-2	7B	Discrete	79.4	71.5	–	**86.2**	**87.7**	–
Being-VL-0.5 *w/o BPE*	8B	Discrete	54.3	38.2	1301.2	57.8	76.1	45.0
Being-VL-0.5	8B	Discrete	80.2	71.8	1525.8	70.3	84.3	57.4
Being-VL-0.5$^+$	8B	Discrete	**80.6**	**72.1**	**1536.3**	69.0	86.0	**57.8**

Table 1. Performance comparison of different visual-language models across standard benchmarks. Models are grouped by tokenizer type (Continuous / Discrete). The best performance in each group is **bolded**, while second-best result in each group is underlined.

Our training pipeline consists of three stages with different masks and objectives:

Stage 1: Embedding Alignment focuses exclusively on training the newly extended visual token embeddings while freezing the rest of the model. The mask is defined as $M_1[i] = 1$ if parameter i belongs to new embeddings, 0 otherwise. This alignment establishes basic visual-linguistic associations without disrupting pre-trained language capabilities.

Stage 2: Selective Fine-tuning unfreezes early transformer layers while keeping later layers frozen, with $M_2[i] = 1$ if parameter i is in layers 1 to k (25% of total layers in our default implementation), 0 otherwise. This enhances cross-modal interactions in the feature integration layers using a broader range of data types.

Stage 3: Full Fine-tuning unfreezes all parameters ($M_3[i] = 1, \forall i$) and emphasizes complex reasoning and instruction-following data. This stage refines the model's advanced capabilities while building upon the foundation established in earlier stages.

4. Experiments

In this section, we empirically evaluate our designed framework and training pipelines. Based on the methods described in Section 3, we train our model **Being-VL-0.5** with an 8K extended vocabulary and a variant **Being-VL-0.5**$^+$ with a 16K extended vocabulary. We first describe our experimental setup, then present comprehensive results demonstrating the effectiveness of our approach compared to baselines. We further conduct analyses of each component to validate our design choices. Considering the limited space, please refer to the Appendix A for hyperparameter choices and further analyses of our experiments.

4.1. Experiments Setup

Evaluation Benchmarks: We conduct evaluation on widely-used benchmarks [27, 28] that assess different aspects of visual understanding: VQAv2 [16] and VizWiz [19] for general visual question answering, MMBench [34] for comprehensive multimodal understanding, MME-P [13] for perception capabilities, SciQA-IMG [38] for scientific reasoning, and POPE [29] for hallucination evaluation.

Baselines: We primarily focus on open-source models with comparable parameters and similar training data scales to ensure that performance differences reflect architectural choices rather than scale advantages. We compare our **Being-VL-0.5** against two kinds of baselines: continuous embedding-based models and discrete token-based models. Continuous embedding-based models include InstructBLIP [9], LLaVA-1.5/LLaVA-Next [31, 32], mPLUG-Owl2 [63, 64], HyperLLaVA [71], ShareGPT4V [8], and VILA-1.5 [30]. Discrete token-based models include Chameleon [54], Being-VL-0 [72], and Unified-IO-2 [37]. We also compare '**Being-VL-0.5** *w/o BPE*', which is a variant that completely removes the BPE part and retains only the base VQ token embedding. This is used to investigate the effectiveness of our framework's core BPE design.

4.2. General Visual Understanding

Table 1 presents our models' performance compared to baselines. Our framework effectively narrows the performance gap between discrete token-based models and continuous embedding-based models. Both **Being-VL-0.5** and **Being-VL-0.5**$^+$ achieve competitive performance across all benchmarks, with **Being-VL-0.5**$^+$ reaching 80.6 on VQAv2 and 72.1 on MMBench, comparable to the best continuous embedding-based models like VILA-1.5 (80.9 and

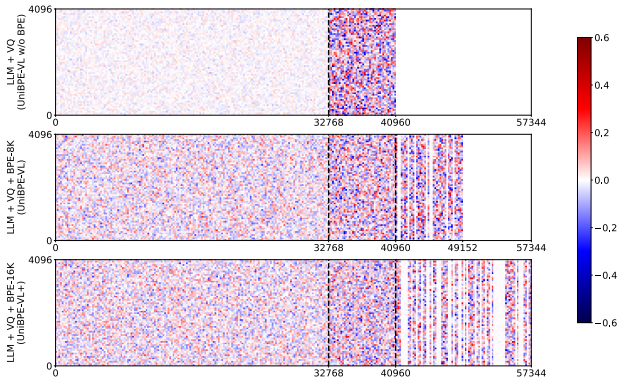

Figure 3. Visualization of embedding weight distributions across three model variants. The heatmaps represent sampled weights from the embedding matrices (4096 dimensions), wherein color intensity indicates weight magnitude (blue for negative, red for positive). The depth of the color indicates the activation magnitude of the corresponding token embedding during model inference. *Top*: LLM+VQ (**Being-VL-0.5** *w/o BPE*) with text tokens (32K) + VQ tokens (8K). *Middle*: LLM+VQ+BPE-8K (**Being-VL-0.5**) with text tokens (32K) + VQ tokens (8K) + BPE tokens (8K). *Bottom*: LLM+VQ+BPE-16K (**Being-VL-0.5**$^+$) with text tokens (32K) + VQ tokens (8K) + BPE tokens (16K).

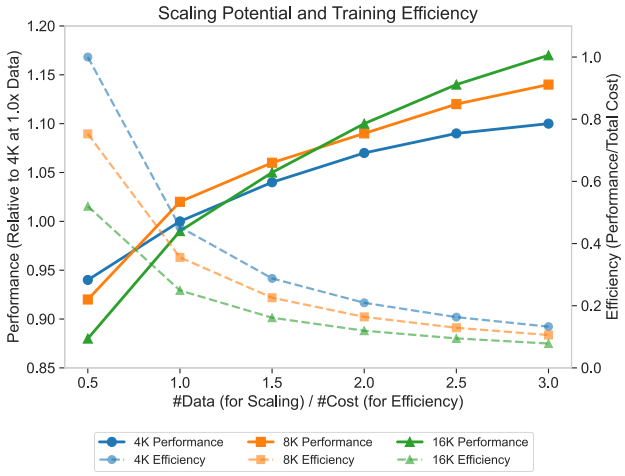

Figure 4. Scaling potential and training efficiency comparison for different vocabulary sizes (4K, 8K, 16K). Performance curves (solid lines) show relative performance gains as training data increases, while efficiency curves (dashed lines) illustrate the performance-to-cost ratio. All metrics are normalized relative to the 4K vocabulary model at standard training data amount (1.0×).

72.3 respectively). **Being-VL-0.5**$^+$ demonstrates strong perception capabilities with 1536.3 on MME-P, outperforming many continuous embedding models.

Among discrete token-based approaches, Being-VL-0 represents the work most closely aligned with our method, as both implement BPE-based visual tokenization. Our method outperforms Being-VL-0 across all benchmarks, with notable gains on MMBench (71.8 *vs*. 44.0) and MME-P (1525.8 *vs*. 1316.2). These improvements are attributable to our enhanced vocabulary construction and optimized training strategies. Unified-IO-2 achieves stronger performance on SciQA-IMG (86.2 *vs*. our 70.3), but it is worth noting that our models evaluate this benchmark in a zero-shot manner without task-specific fine-tuning.

Additionally, our method achieves an optimal balance between capturing low-level visual details and modeling high-level semantic concepts, with reduced hallucination on POPE (84.3 for **Being-VL-0.5**, 86.0 for **Being-VL-0.5**$^+$) compared to earlier discrete token approaches. When evaluating the overall performance across all benchmarks, our method demonstrates competitive results while maintaining the advantages of unified token representation.

Furthermore, the significant performance drop of '**Being-VL-0.5** *w/o BPE*' on all benchmarks also highlights the necessity of the BPE vocabulary as a core part of our framework. We also conducted a qualitative case study, which provides additional evidence of **Being-VL-0.5**'s comprehensive visual understanding. For detailed results, please refer to Section B.

4.3. BPE Token Activation Mechanism

Figure 3 visualizes embedding weight distributions across our model variants, revealing key insights into how BPE tokenization helps unifying multimodal knowledge. In the base LLM+VQ model (top), which is a variant removing BPE vocabulary, we observe a contrast between text (0-32K) and visual tokens (32K-40K). Text tokens exhibit predominantly low-magnitude weights (lighter colors), while visual tokens exhibit significantly higher magnitudes (deeper colors). This imbalance suggests that the model struggles to establish a unified representation space across modalities. With BPE tokenization (middle and bottom panels), this pattern changes substantially. These two models show more balanced weight distributions, suggesting BPE helps bridge the representation gap between modalities, creating a more unified semantic space for visual-textual interaction. This visualization elucidates why models with improved BPE patterns yield superior performance, thereby outperforming Being-VL-0.

However, some BPE tokens exhibit minimal activation (near-zero weights), appearing as white vertical stripes. This pattern is more evident in the 16K variant, where more embeddings remain unused. Intuitively, we hypothesize that this is because the current training does not fully utilize these 16K tokens, which may explain why **Being-VL-0.5**$^+$ underperforms **Being-VL-0.5** on certain tasks. Concurrently, this observation inspires us to examine the trade-offs between representational capacity and learning efficiency.

Configuration	Curriculum	Progressive	Avg. Score	
			Perception	Reasoning
Standard	✓	✓	**80.3**	**71.1**
Progressive only	–	✓	74.9	65.1
Curriculum only	✓	–	76.8	67.5
Single stage	–	–	71.2	62.3

Table 2. Ablation study on different components in training. Configurations marked with "–" use a fixed data mixture ratio throughout training or employ a standard full fine-tuning approach.

Curriculum Strategy	Avg. Score	
	Perception	Reasoning
Progressive (FD → PD → RD → ID)	**80.3**	**71.1**
Reverse (ID → RD → PD → FD)	73.4	64.0
Random	77.3	68.4

Table 3. Analysis of different curriculum strategies. FD, PD, RD, and ID refer to Foundation Data, Perception Data, Reasoning Data, and Instruction Data as described in Section 3.6.1.

Unfreezing Strategy	Avg. Score	
	Perception	Reasoning
3-stage (align → selective → full)	**80.3**	**71.1**
2-stage (align → full)	78.2	69.3
Reverse order	72.1	63.0

Table 4. Analysis of different parameter unfreezing strategies.

4.4. Scaling Potential and Training Efficiency

To further investigate the trade-offs between vocabulary size, performance scaling, and training efficiency, we conducted extended experiments using the same checkpoint. Figure 4 presents our findings comparing three vocabulary sizes (4K/8K/16K) across varying data scales. For performance evaluation, we computed the normalized mean scores across four benchmarks (MME, POPE, VQAv2, and MMBench) and normalized them relative to the 4K vocabulary model at standard training data ($1.0\times$), providing an intuitive comparison of relative performance gains.

The horizontal axis serves dual purposes: for performance curves, it represents the relative amount of training data; for efficiency curves, it represents the total cost factor, incorporating computational resources $\times$ training time. Specifically, total cost increases with both vocabulary size and data amount. Our results reveal an important trade-off: while the 16K vocabulary shows superior scaling potential, eventually outperforming both 4K and 8K models as training data increases beyond $1.5\times$, its efficiency (performance-to-cost ratio) remains consistently lower than smaller vocabulary models. In contrast, the 8K vocabulary achieves a better balance, showing strong performance scaling while maintaining reasonable efficiency.

4.5. Analysis of Multi-Stage Training

We conduct ablation studies to investigate the impact of the multi-stage training. Table 2 shows that our standard approach achieves clear improvements over the single-stage training (12.8% and 14.1% higher on perception and reasoning tasks respectively), demonstrating the effectiveness of the overall strategy. The results reveal that both curriculum-based data composition and progressive parameter unfreezing contribute to the performance, with curriculum data composition showing a stronger impact.

Table 3 shows that the shifting of data types matters significantly. Our curriculum design outperforms both reverse ordering and random mixing. This aligns with our design principle that visual BPE tokens should first establish basic semantic meanings before tackling complex reasoning tasks. Similarly, Table 4 highlights the value of our three-stage progressive approach for parameter unfreezing. The simplified two-stage method shows modest performance differences (2.7% lower on perception tasks), while reverse unfreezing leads to more notable performance drops (11.4% lower on perception tasks), suggesting that the order of parameter updates affects how well the model integrates visual and textual information.

5. Conclusions, Limitations, and Future Work

This paper presents a unified visual tokenization framework that enhances multimodal understanding through three key innovations: priority-guided vocabulary construction that considers both frequency and spatial patterns, curriculum-based data composition that aligns with token complexity progression, and a specialized multi-stage training procedure. Our comprehensive experiments across diverse benchmarks demonstrate that this approach effectively bridges the gap between visual and textual modalities, achieving competitive performance compared to continuous embedding-based methods while maintaining the advantages of a unified token representation.

Our current study focused on models with 8B parameters due to computational constraints. While these models already show promising results, our scaling analysis indicates potential for further improvements with more training data. The observed relationship between vocabulary size and performance suggests that further scaling could achieve even stronger multimodal understanding capabilities.

A natural extension of our unified token-based approach is multimodal generation. Unlike models using separate encoders for different modalities, our framework enables the model to generate visual tokens in the same way it generates text tokens, establishing a truly unified representation paradigm. Although the current work concentrates on understanding tasks, extending our approach to incorporate generation capabilities constitutes a promising direction for future research.

Acknowledgments

This work was supported by NSFC under Grant 62450001. The authors would like to thank the anonymous reviewers for their valuable comments and advice.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 2

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1

[5] Arnab Barua, Mobyen Uddin Ahmed, and Shahina Begum. A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions. *Ieee access*, 11:14804–14831, 2023. 2

[6] David M Chan, Rodolfo Corona, Joonyong Park, Cheol Jun Cho, Yutong Bai, and Trevor Darrell. Analyzing the language of visual tokens. *arXiv preprint arXiv:2411.05001*, 2024. 2

[7] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024. 12

[8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 6, 12

[9] W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. arxiv. *Preprint posted online on June*, 15:2023, 2023. 6

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 12

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[12] Yicheng Feng, Yijiang Li, Wanpeng Zhang, Hao Luo, Zihao Yue, Sipeng Zheng, and Zongqing Lu. Videoorion: Tokenizing object dynamics in videos. *arXiv preprint arXiv:2411.16156*, 2024. 1

[13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 6

[14] Matthias Gallé. Investigating the effectiveness of bpe: The power of shorter sequences. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1375–1381, 2019. 2

[15] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 6

[17] Jian Gu, Aldeida Aleti, Chunyang Chen, and Hongyu Zhang. A semantic-based layer freezing approach to efficient fine-tuning of language models. *arXiv preprint arXiv:2406.11753*, 2024. 2

[18] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024. 12

[19] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 5

[21] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*, 2024. 2

[22] Shixin Jiang, Jiafeng Liang, Ming Liu, and Bing Qin. From specific-mllm to omni-mllm: A survey about the mllms alligned with multi-modality. *arXiv preprint arXiv:2412.11694*, 2024. 2

[23] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified language-vision pretraining

with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023. 1

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 12

[25] László Kozma and Johannes Voderholzer. Theoretical analysis of byte-pair encoding. *arXiv preprint arXiv:2411.08671*, 2024. 1

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 12

[27] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024. 6

[28] Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*, 2024. 6

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 6

[30] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 6

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 6

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 6

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 12

[34] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 6

[35] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 2

[36] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 1, 2

[37] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 1, 2, 6

[38] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 6

[39] Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*, 2024. 2

[40] Mike A Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. *arXiv preprint arXiv:2404.11757*, 2024. 2

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[43] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34, 2024. 2

[44] Nived Rajaraman, Jiantao Jiao, and Kannan Ramchandran. Toward a theory of tokenization in llms. *arXiv preprint arXiv:2404.08335*, 2024. 2

[45] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[46] Anton Razzhigaev, Anton Voronov, Andrey Kaznacheev, Andrey Kuznetsov, Denis Dimitrov, and Alexander Panchenko. Pixel-level bpe for auto-regressive image generation. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 26–30, 2022. 2

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 12

[48] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 12

[49] Rico Sennrich. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 1

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 12

[51] Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, and Weimin Zhang. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.07594*, 2023. 1

[52] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2

[53] Yuchang Sun, Yuexiang Xie, Bolin Ding, Yaliang Li, and Jun Zhang. Exploring selective layer fine-tuning in federated learning. *arXiv preprint arXiv:2408.15600*, 2024. 2

[54] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 2, 6

[55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[56] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[57] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[58] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2

[59] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 2

[60] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[61] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 1

[62] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023. 2

[63] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 6

[64] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023. 6

[65] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1

[66] Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. Can mllms perform text-to-image in-context learning? *arXiv preprint arXiv:2402.01293*, 2024. 1

[67] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[68] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, pages 1–22, 2023. 1

[69] Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[70] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2

[71] Wenqiao Zhang, Tianwei Lin, Jiang Liu, Fangxun Shu, Haoyuan Li, Lei Zhang, He Wanggui, Hao Zhou, Zheqi Lv, Hao Jiang, Juncheng Li, Siliang Tang, and Yueting Zhuang. Hyperllava: Dynamic visual and language expert tuning for multimodal large language models, 2024. 6

[72] Wanpeng Zhang, Zilong Xie, Yicheng Feng, Yijiang Li, Xingrun Xing, Sipeng Zheng, and Zongqing Lu. From pixels to tokens: Byte-pair encoding on quantized visual modalities. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 6

[73] Sipeng Zheng, Bohan Zhou, Yicheng Feng, Ye Wang, and Zongqing Lu. Unicode: Learning a unified codebook for multimodal large language models. *arXiv preprint arXiv:2403.09072*, 2024. 1, 2