

REVIEW

A Review of Vision Language Models for Architectures, Training Methods, Datasets, Evaluation Metrics, Results, and Fine-Tuning Techniques for Vietnamese

Van-Thuan Nguyen^{1,2}, Van-Nui Nguyen² and Van-Hung Le^{3,*}

¹Faculty of Engineering Technology, Hung Vuong University, Nong Trang, Phu Tho, Vietnam

²Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam

³Information Technology Department, Tan Trao University, Minh Xuan, Tuyen Quang, Vietnam

*Corresponding Author: Van-Hung Le. Email: lvhung@tqu.edu.vn

Received: 27 February 2026; Accepted: 14 May 2026; Published: 13 August 2026

ABSTRACT: The vision-language models (VLM) combine the image and text to solve practical applications. Specifically, VLM leverages the results of computer vision in conjunction with natural language processing (NLP), like a large language model (LLM), to address real-world problems such as automating and improving the quality of medical examinations and treatments in healthcare, building autonomous driving systems, image captioning, and generating automated chatbots. To understand the development and application of VLM, we surveyed VLM, classifying it according to model architecture, learning methods, evaluation measures, datasets, challenges, and future development directions of VLM based on the model architecture. Simultaneously, to experiment with the VLM model, focusing on LLM, we collected the VQA-TQU1 dataset with multimodal information streams: images, text descriptions, and audio data of Tan Trao University from 2024 to 2026. The VQA-TQU1 dataset was fine-tuned on the Vintern-1B model for generating visual question answering (VQA) (the results on the *BLEU*, *ROUGE-L*, *METEOR*, *F1*, *Char-F1* measures were 0.5677, 0.739, 0.7532, 0.7435, 0.7461, respectively and compare it with state-of-the-art methods), generating automated responses about Tan Trao University's admissions, and adjusting the data based on the LLM fine-tuned in the Vintern-1B model. We also tested the online video captioning problem on the VQA-TQU1 dataset with two videos based on the SmolVLM2 and Wav2vec 2.0 models, with results showing *RTF* = 0.24 and *RTF* = 0.36, respectively.

KEYWORDS: Vision-language models (VLMs); VLMs classification; embedding-based VLM; generative VLM; cross-modal transformer; VLM + LLM; embodied VLM; fine-tuning VLM for Vietnamese

1 Introduction

Previously, there were two approaches to solving real-world problems: first, vision-based approaches using prominent models (such as CNN, ViT); second, text-only approaches using prominent models (such as BERT, GPT). VLM learns to associate visual content with linguistic semantics. This allows the model to understand both the content in images and the descriptions in language. Recent development of VLM has expanded to include visual input processing capabilities; however, the connection between language and images remains incomplete. Current models often struggle with understanding spatial relationships, counting objects, and are prone to optical illusions, making improved reliability a key research goal [1]. VLMs are divided into five approaches. Instead of relying on costly manual labeling data like traditional DNNs [2],

VLMs leverage web-scale image-text data, enabling efficient zero-shot prediction across a wide range of tasks [1].

Building upon the transformer platform, the research focus has shifted to adapting pre-training for the multimodal domain [3]. MVLM models such as CLIP [4], or GPT-4V [5] have demonstrated superior inference capabilities on both image and text data compared to classical methods [6].

To provide a structured classification system, a robust analytical basis and based on research by Zhang et al. [1], we classify VLMs into five main categories (Embedding-based VLM, Generative VLM, Cross-modal Transformer, VLM + LLM, Embodied VLM, as illustrated in Fig. 1) based on three core criteria: (1) Representational learning objectives (models focused on contrast alignment or generative modeling), (2) Multimodal interaction mechanisms (how image and text features are combined, such as through shared embedding spaces, cross-attention layers, or adapters connected to the LLM), and (3) Inference models (outputs used to retrieve, generate text, or perform actions in the real environment).

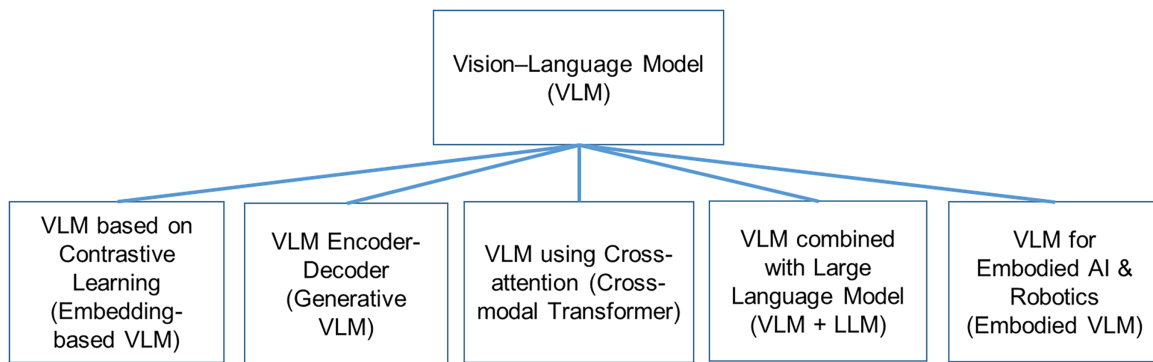


Figure 1: Illustration categorizes VLM models by architecture and application.

Based on these criteria, the boundaries between the categories become clearer: Embedding-based VLM [7] is a group of joint embedding space learning models for images and text, instead of generating text token by token. Embedding-based VLM focuses entirely on aligning methods within a shared space without generating text. Generative VLM [8] is a group of models that combine images and text to generate a sequence of output tokens, typically answers, descriptions, arguments, or actions. Generating VLMs utilizes a traditional Encoder-Decoder architecture to generate token-based content. Cross-modal Transformer VLM [9] is a group of learning models that directly interact with images and text using attention, rather than just matching embeddings or generating tokens from only one side. In short, it “sees” the image and “reads” the question simultaneously to understand the details of each image region related to each word. Cross-modal Transformers (Cross-modal Transformers) emphasize deep and early semantic association through cross-modal mechanisms. VLM + LLM [10] is an architecture that combines the visual-linguistic model (VLM) with the big language model (LLM) to create a system of seeing, understanding, reasoning, and responding in natural language. In short, VLM helps with “seeing,” and LLM helps with “thinking and speaking”. VLM + LLM utilizes pre-trained LLMs as giant reasoning brains linked via lightweight transducers. Embodied VLM [11] is a group of models that combine vision, language, and action, allowing an agent to see the environment -> understand commands -> plan -> perform actions in the physical or simulated world. Thus, Embodied VLM is seeing + understanding + acting. A comprehensive comparison of these five groups, highlighting architectural differences, advantages, limitations, and typical tasks, is presented in Table 1.

Table 1: Differences in architecture, advantages, limitations, and typical tasks of the five types of VLM models.

VLM Models	Interaction Mechanism & Learning Objectives	Advantages	Limitations	Tasks and Typical Model
Embedding-Based VLM	Mechanism: Mapping images and text into a shared embedded space. Objective: Learning through contrast (bringing matching pairs closer together, pushing mismatched pairs further apart).	High computational efficiency for zero-shot retrieval and learning; excellent general representation extraction; easily scalable.	It's not possible to generate new text (only selection/matching is possible); it requires a huge amount of image-text data for training.	Image-text retrieval, zero-shot classification. (Examples: CLIP, ALIGN, SigLIP)
Generative VLM	Mechanism: The Encoder-Decoder architecture utilizes a self-regression or diffusion mechanism. Objective: To model a multimodal distribution to generate output tokens.	Capable of generating entirely new data (text or images); high flexibility for multimodal synthesis.	The decoding process is computationally intensive, prone to hallucinations, and has a slow generation rate.	Create images from text (Text-to-Image), and add image captions (Image Captioning). (Examples: Coca-Cola, CM3Leon)
Cross-Modal Transformer	Mechanism: Early and deep integration, where text and image tokens interact directly with each other through cross-attention layers. Objective: Masked signal modeling and image-text matching.	Deep understanding of the semantics of visual grounding; sensitivity to syntactic structures.	Inference speed is slower due to early method interactions; less efficient when scaling to long text generation tasks.	Answering questions using visual questionnaires (VQA), understanding reference expressions (REC). (Examples: VisualBERT, ViLBERT)
VLM + LLM	Mechanism: Connects the image encoder to a pre-trained/weighted LLM via transducers (e.g., Q-Former, Linear Projection). Objective: Refine multimodal directives and generate autoregressive text.	Inheriting strong logical reasoning abilities and vast knowledge from LLM, superior zero-shot and few-shot conversational skills.	High VRAM consumption for LLM; potential to inherit illusions/biases from the original LLM; risk of losing fine image details.	Multimodal conversational assistants, complex visual inference. (Examples: LLaVA, Flamingo, BLIP-2, GPT-4V)
Embodied VLM	Mechanism: Integrates the visual space, language, and actions of the robot/agent, typically handling all methods as identical tokens. Objective: To model the Visual-Language-Action (VLA).	Allows interaction with physical/simulated environments in the real world; transforms multimodal understanding into concrete actions.	The training process is extremely complex, heavily reliant on scarce robot interaction data, and requires extremely low latency to ensure real-time safety.	Robot navigation, Robot operation planning. (Examples: PaLM-E, RT-2)

Another issue is the rapidly developing AI application systems for automatically responding to user requests, which can include image or voice information. Solving these problems often requires using VLM models to combine both image and voice information into multimodal data to meet user requests. This study focuses on responding to user requests in Vietnamese regarding university admissions at Tan Trao University. The main contributions to this overview and problem-solving are as follows:

- We conducted a fairly comprehensive survey of five types of VLM models: Embedding-based VLM, Generative VLM, Cross-modal Transformer, VLM + LLM, and Embodied VLM. Simultaneously, the VLM models were examined in terms of methodology, datasets, evaluation metrics, results, and challenges.
- We collected and prepared the VQA-TQU1 dataset with multimodal data, including images, text descriptions, and audio data describing the admissions information activities of Tan Trao University from 2024 to 2026. VQA-TQU1 includes 180 color images and 5 question-and-answer pairs, each question containing approximately 100 words, with 1080 interactions in pure Vietnamese, helping the model learn to express itself naturally, intimately, and accurately in everyday language.
- We performed fine-tuning of the Vintern-1B model using input data consisting of images with Vietnamese labels and audio from the VQA-TQU1 dataset to generate question-based descriptions in Vietnamese about TQU's admissions process. Simultaneously, we also tested the online video captioning problem on the VQA-TQU1 dataset with two videos based on the SmolVLM2 and Wav2vec 2.0 models.

The structure of our paper is organized as follows: Related studies are presented in [Section 2](#). [Section 3](#) provides an overview of VLM (from classification, methods, datasets, evaluation metrics, results, and challenges). [Section 4](#) presents the problems in training the VLM. [Section 5](#) presents the Vintern-1B model and the fine-tuning process of the model based on the VQA-TQU1 dataset for building an automated admissions response system. Finally, conclusions and future research are presented in [Section 6](#).

2 Related Works

Currently, research on the application of VLM models has received significant attention over the past half-decade. The VLM model combines image and text to solve practical applications. VLM is applied in many real-world applications, such as in healthcare to automate and improve the quality of medical examinations and treatments, in building autonomous driving systems, image labeling, creating automated chatboxes, etc. Kim et al. [8] conducted a comprehensive survey on the application of VLM models in healthcare to support diagnosis and treatment. The study examined the methods and architectures of VLM models, the metrics used to evaluate VLM models, and the datasets used for fine-tuning and evaluating VLM models. This focuses on two issues: generating medical reports and medical VQA.

Zhou et al. [12] conducted a comprehensive survey of studies using VLM models for autonomous driving, applying VLM models to solve problems such as perception and understanding, navigation and planning, decision-making and control, end-to-end autonomous driving, and data generation. For each problem, the methods, metrics (Top-N Accuracy, BLEU, Recall@K, SPD, APE, FD, nDTW, etc.), and datasets (Caltech Ped Det, Cityscapes, CityPersons, KITTI, CityFlow, nuScenes, BDD100K, ONCE, SUP-AD, Rank2Tell, DRAMA, Reason2Drive, Waymo, etc.) were thoroughly examined. Simultaneously, the advantages and disadvantages of each method, as well as the challenges, were also investigated in detail.

Chen et al. [13] conducted a survey study on several issues related to feature extraction, model architecture, pre-training objectives, pre-training datasets, and downstream tasks of a VLM model. The study details these issues, with each study including a table summarizing the input data, model architecture, pre-training model datasets, encoder, decoder, etc.

Danish et al. [14] conducted a survey using a classification approach to VLM models for various purposes. The VLM models were divided into three groups: Vision-Language Understanding, Text Generation with Multimodal Input, and Multimodal Output with Multimodal Input. For each group, the methods were examined in detail regarding model architecture, evaluation datasets, and detailed results with the group's objectives based on the MME benchmark, Science-QA, VizWiz, Flickr30K, POPE, VQAv2, GQA, LLaVaBench, Chart-QA, MM-Vet, and ViSiTBench datasets. The development direction of each group was also presented.

Pham et al. [15] surveyed the implementation methods of the VLM model, focusing on two parts: model-based and methods for optimizing image-text integration in the VLM model, from 2018 to 2024. The challenges of implementing image-text integration in the VLM model were also presented. Furthermore, the model's applications in solving three problems (image captioning, visual question answering, and cross-modal retrieval) were detailed.

3 Overview of VLM

3.1 VLM Methods

Based on research by Zhang et al. [1], and related studies, VLM models are divided into 5 types: Embedding-Based VLM, Generative VLM, Cross-Modal Transformer, VLM + LLM, and Embodied VLM. Below, we present details of each approach group.

- 1) **Embedding-Based VLM:** The model maps images and text to the same feature space, so that a correct image description pair is close together, while an incorrect pair is far apart. VLM models following this approach, with features such as not generating text, are often powerful in retrieval and zero-shot learning algorithms, especially during training using contrast loss. VLM models using this approach are commonly applied in image-text retrieval and zero-shot image classification.

AlignVLM [16] focuses on alignment between visual representation space and language to improve multimodal content understanding, especially in tasks involving understanding text within images, such as documents. Align VLM architecture (as illustrated in Fig. 2) consists of the following main components: vision encoder (VE), ALIGN module, and fusion into LLM.

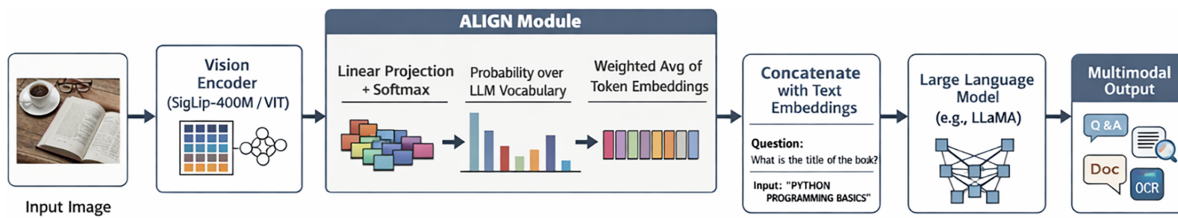


Figure 2: Illustration of the architecture of AlignVLM [16].

SigLIP 2 [17] is a multilingual vision-language encoder model consisting of two main components: a vision transformer (ViT) for extracting image features and a multilingual text encoder transformer for representing descriptions in various languages. These two encoders are trained simultaneously to map images and text to the same joint embedding space. SigLIP 2 is commonly used as a vision-language backbone for VLM models and zero-shot retrieval, VQA, and multimodal understanding tasks. VLM models based on contrast learning are a popular approach, stemming from the idea of the Energy-Based Model (EBM) [18] to assign low energy to matching image text pairs and high energy to mismatched pairs. CLIP [4] shows that with just a simple pre-training task: (i) predicting which caption matches which image; (ii) the model can

learn powerful and scalable visual representations on a dataset of 400 million image-text pairs. After pre-training, the model can use natural language to reference or describe the learned visual concepts, enabling zero-shot execution on many different vision tasks without retraining. Evaluations across more than 30 vision datasets showed the model performed efficiently in most tasks, often achieving results competitive with fully supervised trained models. Inspired by noise-reducing auto-encoders, BERT, MAE, and I-JEPA, this method partially masks the input information. It requires the model to predict the masked portion to learn deeper semantic representations. Within this group, two typical models are frequently mentioned.

The foundation model for visual-linguistics must demonstrate consistent performance across three task groups: pure visual, pure linguistic, and cross-modal/multimodal tasks. Based on this goal, the research team introduces FLAVA [19] as an attempt to build a comprehensive multimodal foundation model. The architecture of FLAVA is presented in Fig. 3. FLAVA was trained on over 70 million image-text pairs. Experimental results showed that the model achieved strong performance on 35 tasks across various paradigms, demonstrating FLAVA's potential as a significant step towards unified models in the field of visual-linguistics.

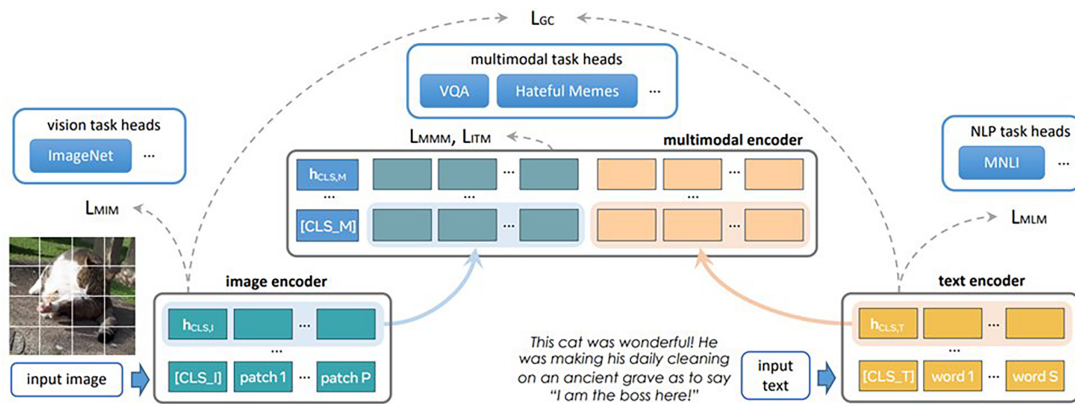


Figure 3: FLAVA [19] integrates image, text, and multi-modal encoders, using MIM/MLM for single-modal pre-training and contrastive/MMM/ITM for image-text pairs. Upon refinement, classifier heads are attached to each encoder to address the respective visual, linguistic, and multimodal tasks.

MaskVLM [20] investigates how to apply masked signal modeling techniques to visual-linguistic representation learning. Instead of developing Masked language modeling (MLM) and masked image modeling (MIM) as two separate tasks, the authors propose a common masking mechanism for both images and text, where the masked signal of one modality is reconstructed by the model based on information from the other modality. Experimental results across multiple V + L tasks show that this method, when combined with standard misalignment functions, achieves leading performance on large-scale pre-training datasets. In particular, the model also demonstrates a clear advantage in constrained data scenarios, outperforming other competing methods. MaskVLM improves upon FLAVA by eliminating intermediate models like dVAE, directly masking image pixels and text tokens. This allows two modalities to provide information to each other to reconstruct the masked portion.

Training VLMs from scratch requires enormous resources (for example, OpenCLIP [21] needs more than 500 GPUs). Therefore, the current trend is to leverage existing LLMs and only train the visual-linguistic connection module. This method saves costs while inheriting the powerful language capabilities of LLMs to extend multimodal capabilities.

Based on the few-shot capabilities of autoregressive LLMs, Frozen [22] transfers this capability to the multimodal domain by training an image encoder to transform images into a continuous sequence of embeddings. These embeddings act as prefixes for a fixed pre-trained LLM. With an input of an interleaved sequence of images and text, the system acts as a multi-modal few-shot learner, capable of performing VQA, learning new objects, and quickly extracting external knowledge through just a few sample examples.

The MiniGPT Series exemplifies the LLM-first approach (learning only the connection module), optimizing resources and leveraging existing LLMs. MiniGPT-4 [11] uses a linear projector to connect the BLIP-2 [23] encoder to the Vicuna LLM, achieving high zero-shot/few-shot efficiency at low cost. MiniGPT-5 [24] expands image generation capabilities through connection with Stable Diffusion. MiniGPT-v2 [25] integrates multiple tasks (including grounding) using task identifiers for precise processing within a single model. Qwen-VL [26] is a successor model to Qwen-7B, using a cross-attention mechanism to compress image features down to 256 tokens. The model applies the Q-Former [27] method (similar to BLIP-2) as a bridge to transform embeddings from ViT to the token space of LLM. This design helps Qwen-VL achieve high computational efficiency and strong capabilities in captioning, VQA, and object localization tasks.

- 2) **Generative VLM:** Unlike performance learning, generative-based VLMs focus on generating new data and modeling the entire multi-modal distribution. In the Encoder–Decoder architecture of Generative VLM, images are fed into an encoder to extract visual features based on CNNs or ViTs to encode the image into visual embeddings. Then, a language decoder uses these features based on Transformer decoders, receiving visual embeddings via cross-attention and generating responses/descriptions using an autoregressive mechanism to produce a text string based on each token. The text generation process occurs token by token until the end of the string. Major development directions include: text-to-image generation (Captioning, VQA), image-to-text generation (such as Diffusion models), and multi-modal autoregressive models that handle both within a unified token space. As such, they have become crucial foundations for modern innovative AI applications.

Show or Tell [28] does not propose a completely new VLM architecture, but focuses on leveraging (prompting) pre-trained Vision-Language Models to solve the semantic segmentation problem. The focus is on comparing and combining two forms of prompting: visual prompts (show) and textual prompts (tell).

BLIP-2 [23] is a vision-language model designed to efficiently connect VEs and large language models (LLMs) without requiring the entire large model to be retrained. Instead of directly fine-tuning the LLM, BLIP-2 introduces a Q-Former as an intermediary bridge to align the two powerful pre-trained models. BLIP-2 consists of three main components: First, the VE (Frozen) uses a pretrained vision model (ViT, CLIP, EVA, etc.) for execution. It receives input images and extracts visual embeddings. The entire VE is parameter frozen to reduce training costs. Second, the Q-Former (Querying Transformer) is a lightweight transformer with learnable query tokens. It performs cross-attention with visual embeddings to select important visual information. The Q-Former acts as an adapter between the VE and the LLM behind it. Thirdly, there is LLM (Frozen), which uses pretrained LLMs (OPT, FLAN-T5, LLaMA,...) for implementation. It takes the output of the Q-Former as input (mapped to the language space) and uses an autoregressive text generation mechanism.

CoCa (Contrastive Captioner) [29] is an efficient model combining contrast learning and caption generation (similar to SimVLM), as presented in Fig. 4. The decoder architecture is separated: the first half processes text unimodally to calculate contrast loss with the image, while the second half performs cross attention to generate autoregressive captions. Thanks to end-to-end training on web-scale data (with every label treated as text), CoCa achieved outstanding performance on recognition tasks (ImageNet,

Kinetics, Moments-in Time, Monfort), retrieval (MSCOCO [30], Lin, Flickr30K, Plummer, MSR-VTT), and description generation (No Caps). In particular, the SOTA model built on ImageNet achieved 91.0% fine-tuning, demonstrating the power of unified contrast learning and generation.

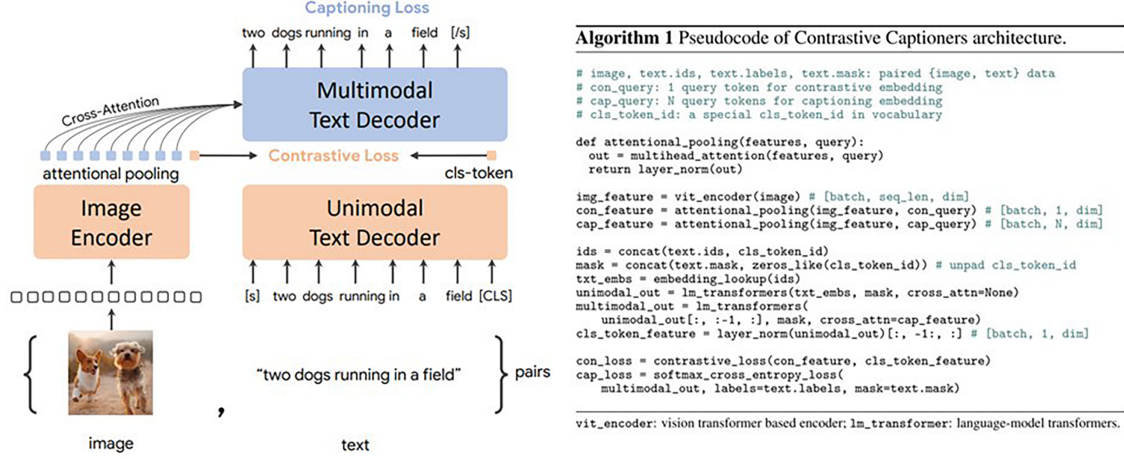


Figure 4: Architectural details of CoCa and corresponding training objectives [29]. On the left is the architecture of the CoCa model, on the right is a pseudocode of the algorithm used by CoCa.

CM3Leon [29] is a token-based decoder-only model with enhanced recall, capable of generating and filling in both images and text. It is the first model to apply the standard LLM training process: large-scale pre-training combined with multi-task supervised fine-tuning (SFT). CM3Leon introduces contrast decoding to optimize quality, achieving SOTA results in image generation (FID 4.88 on MS-COCO zero-shot) at only 1/5 of the computational cost. After SFT, the model demonstrates superior control over image editing and segmentation, proving the effectiveness of transferring the training method from text to multi-modal.

- 3) **Cross-Modal Transformer:** VisualBERT [31] is a simple and flexible model for many vision-language tasks. VisualBERT consists of a stack of Transformer layers capable of implicitly aligning input text components with corresponding image regions through a self-attention mechanism. They also proposed two visually grounded language modeling goals for pre-training VisualBERT on image description data. Experiments on four vision-language tasks, including VQA [32], VCR [33], NLVR2 [34], and Flickr30K [35], showed that VisualBERT outperformed or achieved comparable performance to the most advanced models while maintaining remarkable simplicity. Further analysis reveals that VisualBERT is capable of grounding linguistic elements to visual regions without any explicit oversight, and is even sensitive to syntactic relationships—for example, tracking the connection between verbs and their corresponding visual regions and arguments. ViLBERT [36] extends the traditional BERT [37] architecture into a multi-modal (as in Fig. 5), two-stream model where images and text are processed in two separate streams but interact with each other through co-attentional transformer layers. The model is pre-trained on two representative tasks using the Conceptual Captions [38] dataset, which is automatically collected on a large scale, and then transferred to established visual-linguistic tasks such as visual questioning, visual common sense inference, expression-based localization, and description-based image retrieval.

With minimal adjustments to the original architecture, ViLBERT achieved superior performance and established the best results across all four tasks. This work marks a significant shift: instead of learning the

connection between vision and language in individual problems, “visual grounding” is now viewed as a pre-trained and flexibly reusable capability.

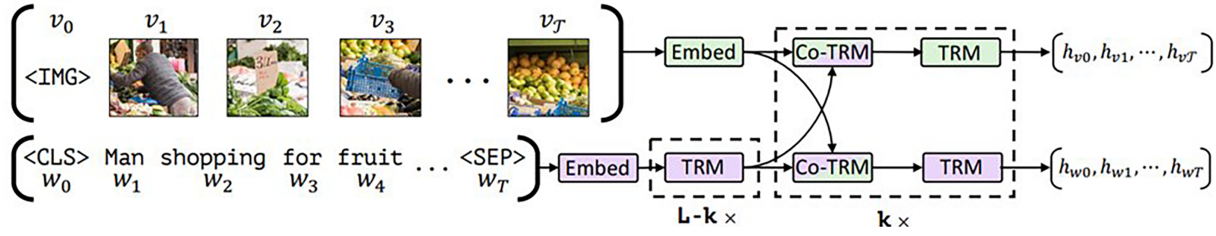


Figure 5: The ViLBERT architecture illustration shows two parallel processing streams for images (blue) and text (purple), interacting with each other via co-attention transformer layers.

In summary, these two models combine text tokens and image tokens (extracted from a CNN or detector), and are trained on two goals: predicting the obscured portion of text or an image, and predicting whether a caption matches the image. Thanks to the attention mechanism, the models learn to associate words with visual signals. This forms the foundation for later VLMs.

- 4) **VLM + LLM:** DinoV2–SigLIP–Phi-3 (LoRA) [39] is a visual-linguistic model built to powerfully combine self-supervised visual representation, visual-linguistic alignment, and efficient language generation. The model leverages LoRA for image feature extraction, SigLIP for visual-linguistic spatial alignment, and Phi-3 as a large language model, with LoRA for efficient parameter fine-tuning. LoRA comprises four components: First, the VE-DINOv2, which uses a ViT pretrained by self-supervised learning to extract semantically rich, dense visual features from input images and exploit the representation of structure, shape, and spatial relationships. Second, the Vision–Language Alignment–SigLIP, which receives visual embeddings from LoRA and text embeddings as input. The first step involves mapping and aligning the two visual-linguistic spaces using contrast/sigmoid-based loss. This creates visual tokens aligned with the linguistic space, suitable for LLM.

The third step is the Projection/Adapter Layer, which converts the aligned visual features into the Phi-3 embedding space, ensuring size and representation distribution compatibility. The fourth step is LLM-Phi-3 + LoRA, where Phi-3 acts as a language-generating decoder, performing multi-modal inference, and LoRA (Low-Rank Adaptation) is applied to the attention/FFN layers.

Flamingo [40] is a VLM proposed by DeepMind, designed to combine the powerful image comprehension capabilities with the reasoning and language generation abilities of LLMs. Flamingo’s key feature is its multi-modal few-shot learning capability, allowing the model to perform many vision-language tasks efficiently with only a few instructional examples. The Flamingo model architecture is built on a VLM + LLM architecture, consisting of the following main components: First, a VE, typically using a pre-trained ViT, which extracts features from images or sequences of frames (video); Second, a Perceiver Resampler, which acts as an intermediary between the VE and the LLM, converting a large number of image tokens into a fixed set of tokens, offloading the LLM and enabling efficient processing of both individual images and videos; Third is the LLM, which relies on a large language model (like Chinchilla) to handle inference, context understanding, and text generation; Fourth are Gated Cross-Attention Layers, which are used to insert into the LLM layers to allow language tokens to directly pay attention to visual information, using gating mechanisms to control the degree to which visual information influences the language generation process.

LLaVA (Large Language and Vision Assistant) [39] is a VLM proposed to effectively combine image comprehension capabilities with the powerful linguistic inference of Large Language Models (LLMs). The

main goal of LLaVA is to build a multimodal assistant capable of conversation and inference based on both visual and natural language information. The LLaVA model architecture is built on a VLM + LLM architecture, comprising three main components: First, the VE, typically using a pre-trained ViT-L/14 CLIP, extracts visual features from the input image; Second, the projection layer (visual projection), used to map image features from the visual space to the LLM embedding space, ensuring compatibility between image representation and language tokens; Thirdly, the LLM, often using LLaMA (7B, 13B), plays a central role in reasoning, understanding instructions, and generating responses. Unlike BLIP-2 or InstructBLIP, LLaVA does not use a Q-Former, but instead uses a simple linear projection layer to connect the VE to the LLM.

InstructBLIP [41] is an advanced VLM proposed to enhance the ability to understand and follow language instructions in vision-language tasks. This model extends BLIP-2 by combining instruction tuning with LLM, significantly improving multimodal inference capabilities. InstructBLIP consists of three main components. LLMs are responsible for inferring and generating responses based on encoded instructions and visual information. Unlike traditional VLMs that only train on image-description pairs, InstructBLIP is fine-tuned on a multimodal instruction dataset, where each sample includes: an image, a natural language instruction (question, command, request), and a corresponding response.

GPT-4V (GPT-4 with Vision) [5] is a multimodal model developed by OpenAI, extending from GPT-4 by integrating image comprehension with natural language processing. Unlike traditional VLMs designed for specific tasks, GPT-4V aims for a general inference model that can flexibly handle many vision-language problems through dialogue. GPT-4V can be considered a hybrid VLM + LLM model, where the architecture of GPT-4V includes two main components. First is the VE, responsible for extracting features from the input image. Visual features are mapped to a representation space suitable for the language model. Second is the LLM (GPT-4), which plays a central role in inference, context understanding, and language generation. GPT-4V simultaneously receives information from images and text to provide a consistent, highly logical response. Details about its internal architecture are not fully disclosed; however, GPT-4V follows an end-to-end multimodal reasoning model, where images are treated as contextual input for the LLM.

- 5) **Embodied VLM:** PaLM-E (PaLM-Embodied) [42] is a large-scale multimodal model proposed by Google Research, aiming to integrate vision, language, and action (Vision–Language–Action) into a unified model. PaLM-E is designed for the Embodied AI context, where the agent not only understands images and language but also interacts and acts within a physical environment, typically a robot. The model architecture of PaLM-E is built upon the LLM PaLM, extended by integrating multimodal features. Unlike many traditional VLMs, PaLM-E represents all modalities as tokens, allowing the LLM to process them consistently.

RT-2 (Robotics Transformer 2) [11] is a Vision-Language Action (VLA) model proposed by Google DeepMind, aiming to directly connect visual-linguistic understanding with robot action. RT-2 inherits and expands on RT-1 by leveraging Large VLM trained on large-scale internet data, thereby increasing semantic inference and generalization capabilities in the robot environment. RT-2 is built on the Vision-Language Transformer, consisting of the following main components: The VE, to extract features from RGB images observing the robot environment, inherited from large VLMs (e.g., PaLM-E-style or Flamingo-style); The Language-Action Transformer; and Action Tokenization. Although VLM architectures are diverse, their performance in specific tasks is strongly influenced by how the internal interacting components are designed. Evaluating the trade-offs between these components is key to understanding why some models excel in certain scenarios: Contrastive Learning (e.g., CLIP, SigLIP): Aligning the embedded space using contrast loss functions is highly effective for zero-shot retrieval and classification tasks. However, because these components compress the entire image into aggregate vectors, they lack the detail required for spatial inference or text generation tasks, forcing their use as a Vision Encoder in larger VLM systems.

The Q-Former module (e.g., BLIP-2, InstructBLIP): The Q-Former acts as an information bottleneck to extract the most important image features before feeding them into the LLM. Its biggest advantage is its computational efficiency and scalability, as it significantly reduces the number of image tokens (e.g., down to 32 tokens). However, the trade-off is that excessive compression makes the model prone to losing small image details or text (as in OCR scenarios), reducing its detail inference capability.

Linear/MLP Projection (e.g., LLaVA, Vintern-1B): In contrast to Q-Former, models like LLaVA or Vintern-1B use a simple projection layer (MLP) to directly convert the entire image token into the LLM space. This design maximizes fine-grained reasoning and is very powerful in text-centric VQA or OCR tasks. However, it consumes a large amount of memory because image tokens occupy too much space in the context window of the LLM. For low-resource applications (such as Vietnamese), this architecture is preferred because the LLM can fully accept visual clues for accurate translation and description (typically the optimal results of Vintern-1B on our VQA-TQU1 dataset).

Gated Cross-Attention (e.g., Flamingo): By directly embedding interleaved cross-attention layers into the core architecture of the LLM, Flamingo cleverly controls the amount of visual information transmitted into the language. This structure proves superior in interleaved multi-modal few-shot scenarios. However, it requires deep intervention into the LLM structure and involves extremely high retraining costs compared to methods that only train the Adapter layer, such as LLaVA.

Instruction Tuning: This is the crucial component that transforms a model from merely “recognizing” to being able to “reason and follow commands” like InstructBLIP, LLaVA, and GPT-4V. In multilingual adaptation scenarios, fine-tuning instructions on high-quality datasets (such as Vietnamese conversations in VQA-TQU1) helps LLMs learn how to align language and avoid the illusion caused by low language resource barriers, solving the semantic compensation problem that purely English-language training models cannot.

3.2 Datasets

To evaluate the performance of machine learning and deep learning (DL) models, standardized datasets are needed to assess and compare peer results. To evaluate the performance of VLM models in solving problems (image classification, robust classification, captioning, VQA, knowledge-based VQA, OCR-based VQA, Multimodal Question Answering, etc.), we surveyed a series of datasets used for training and testing VLM models, as shown in Table 2. In Table 2, we have shown the datasets, including the dataset name, the number of samples/images, the number of object classes, the purpose of the dataset for evaluation, its use for training/validation/testing the model, and the purpose of evaluation in the VLM model.

3.3 Evaluation VLM

Measuring the ability to map text to visual cues is core to VLM assessment. Tasks include image captioning (VQA), text-centric VQA, and zero-shot prediction (such as ImageNet) to test background knowledge. Modern benchmarks like Winoground [43] focus on compositional reasoning. Additionally, assessing bias and hallucinations is also a key consideration.

Table 2: Summary of datasets commonly used for training and evaluation in VLMs.

Dataset	# Images/Samples	# Classes/Labels	Main Task	Usage	Purpose in VLMs
ImageNet-1K	~1.28M images	1000 classes	Classification Image	Training	Visual encoder pretraining

(Continued)

Table 2 (continued)

Dataset	# Images/Samples	# Classes/Labels	Main Task	Usage	Purpose in VLMs
ImageNet-A	~7.5K images	200 classes	Classification Robust	Testing	Adversarial evaluation, Robustness, and
ImageNet-R	~30K images	200 classes	Classification Robust	Testing	Non-natural image Generalization to domains
ImageNet-V2	~10K images	1000 classes	Classification Image	Testing	Distribution shift evaluation
CIFAR-10	60K images	10 classes	Classification Image	Train/Test	Lightweight visual pretraining
CIFAR-100	60K images	100 classes	Classification Image	Train/Test	Representation learning, Fine-grained visual
MSCOCO [30]	~330K images	80 object classes	Captioning, VQA	Training	Alignment and generation Vision-language
Flickr30K [35]	31K images	–	Image Captioning	Train/Test	Image-text alignment learning
VQAv2 [44]	204K images/1.1M QA	Open-ended	Visual Question Answering	Train/Test	Training and evaluation, Visual reasoning
OK-VQA [45]	~14K images/14K QA	Open-ended	Knowledge- based VQA	Testing	Reasoning with the external world knowledge
TextVQA [46]	28,408 images/45,336 QA	Open-ended	OCR-based VQA	Train/Test	Text understanding in natural images
POPE [47]	~9K QA pairs	–	Hallucination Detection	Testing	Hallucination in VLMs: Evaluation of the object
AI2D [46]	~5K images	~20 concept types	Diagram Question Answering	Testing	Diagram and abstract visual reasoning
ScienceQA [48]	~21K samples	Open-ended	Multimodal Question Answering	Train/Test	Scientific reasoning Multi-step

(Continued)

Table 2 (continued)

Dataset	# Images/Samples	# Classes/Labels	Main Task	Usage	Purpose in VLMs
MMBench [49]	~3K questions	–	Multimodal Benchmark	Testing	VLM benchmarking Comprehensive
NoCaps [50]	~15K images	–	Zero-shot Image Captioning	Testing	Out-of-distribution generalization
DocVQA [51]	12,767 images/50,000 QA	–	Document Question Answering	Train/Test	Document understanding and layout reasoning
ChartQA [52]	20,882 images/32,719 QA	–	Chart Question Answering	Train/Test	Numerical and logical reasoning
InfographicVQA [53]	5485 images/30,035 QA	–	Visual + TextQuestion Answering	Testing	Complex multimodal information understanding

In this study, we surveyed a range of evaluation criteria for VLM models with the following objectives: Image–Text Mapping (Core Ability), Image Captioning, Text-Centric VQA, Zero-shot Recognition, Compositional Reasoning, Visual Dialog, REC, Document and Layout Understanding, Video Understanding, Bias Evaluation, Hallucination Detection, and Responsible AI. These objectives are presented in Table 3.

Table 3: Measures for evaluating the VLM model.

Evaluation Aspect	Task/Benchmark	Representative Metrics	Evaluation Focus
Image–Text Mapping (Core Ability) [52]	General VLM Benchmarks	Accuracy, Similarity Scores	Ability to associate textual concepts with visual cues
Image Captioning [54]	MSCOCO Captioning [55]	BLEU, ROUGE, CIDEr, METEOR, F1-Score	Quality, relevance, and completeness of the generated captions
Text-to-Text Image Evaluation [56]	CLIPScore [30]	Cosine Similarity (CLIP Embedding)	Semantic alignment between image and generated text
TIFA [57]	VQA Accuracy	Faithfulness of captions via question answering	–
DSG [58]	Dependency verified Accuracy	Hallucination-aware semantic consistency	–
VPEval [59]	Program Execution Accuracy	Grounded reasoning via visual programs	–

(Continued)

Table 3 (continued)

Evaluation Aspect	Task/Benchmark	Representative Metrics	Evaluation Focus
VQA [12,44,60]	VQAv2 [44], OK-VQA [45], TextVQA [46]	Accuracy, Soft Accuracy	Visual grounding and multimodal reasoning
Text-Centric VQA [60,61]	TextVQA [46], DocVQA [51], ChartQA [52]	Exact Match, ANLS	Understanding text embedded in images
Zero-shot Recognition [62]	ImageNet (Zero-shot) [63]	Top-1/Top-5 Accuracy	Background knowledge and visual–semantic alignment
Compositional Reasoning [64]	Winoground [43]	Pairwise Accuracy	Fine-grained compositional and relational reasoning
Visual Dialog [65]	VisDial [65]	NDCG, MRR, Recall@k	Context-aware multi-turn visual understanding
Referring Expression Comprehension (REC) [66]	RefCOCO/+g [67]	IoU, Accuracy	Object grounding and spatial reasoning
Document & Layout Understanding [68]	DocVQA [51]	ANLS [69]	Structured document reasoning
MMMU [70]	Accuracy	Multidisciplinary inference from complex documents	–
Video Understanding [71]	Video-MME [72]	Accuracy, Reasoning Score	Spatio-temporal and long-horizon reasoning
Bias Evaluation [73]	VL-Bias [74]	Bias Score, Embedding Analysis	Detection of visual–linguistic bias
Social Bias in VLMs [75]	Disparity Metrics	Fairness and social risk assessment	–
Hallucination Detection [76]	POPE, DSG [55]	Hallucination Rate	Faithfulness of generated outputs
Responsible AI [77]	Safety & Robustness Benchmarks	Robustness Drop, Error Rate	Reliability, safety, and deployment readiness

3.4 Results of VLM

Table 3 presents some results evaluating VLM models based on CLIP (ViT-B/16), ALIGN, SigLIP, and CoCa methods in solving the zero-shot classification problem. The evaluation results were obtained from the following datasets: ImageNet-1K, ImageNet-A, ImageNet-R, CIFAR-10, CIFAR-100, ImageNet-V2, with accuracy metrics showing that the results ranged from 68.7% to 96%. The image-text retrieval results of the CLIP, ALIGN, SigLIP, Llip6, MetaCLIP4, and CoCa methods on the MSCOCO [30] and Flickr30K [35] datasets for text-to-image and image-to-text purposes are shown in Table 4. The results were evaluated on the Recall@1 metric.

The results for VQA, multimodal reasoning, image captioning, document, and text-centric VQA based on the latest VLM models are shown in Table 5.

Table 4: Some results evaluating VLM models based on CLIP (ViT-B/16), ALIGN, SigLIP, and CoCa methods in solving the zero-shot classification problem.

Zero-Shot Classification (Accuracy)				
Dataset/Method	CLIP (ViT-B/16) (%) [78,79]	ALIGN (%)	SigLIP (%)	CoCa (%) [29]
ImageNet-1K	68.7	76.4	72.3	86.3
ImageNet-A	50.6	75.8	–	90.2
ImageNet-R	77.7	92.2	–	96.5
CIFAR-10	89.35	–	96.0	–
CIFAR-100	63.29	–	79.0	–
ImageNet-V2	76.2	76.4	–	80.7

Table 5: The results for VQA, multimodal reasoning, image captioning, document, and text-centric VQA based on the latest VLM models.

VQA (Accuracy-Acc)								
Dataset/Method	BLIP [80] (%)	BLIP-2 [23] (%)	Flamingo [81] (%)	PaLI (%) [82]	DinoV2 (%)	LLaVA (1.5) (%)	Kosmos-2 (%) [81]	GPT-4V (%)
VQAv2 [44]	77.39	65.0	56.3	84.3	63.3	78.5	45.6	64.28
OK-VQA [44]	–	55.4	57.8	64.5	61.3	45.6	64.3	–
TextVQA [46]	31.48	–	37.9	80.78	46.8	46.1	43.9	78
POPE [47]	–	–	–	–	86.3	85.9	80	88.0
AI2D [83]	–	–	–	75.2	58.9	–	–	78
Multimodal Reasoning								
Dataset/Method	BLIP-2 (%)	Flamingo (%) [59]	LLaVA (%)	Kosmos-2 (%) [59]	InstructBLIP (%) [59]	Shikra (%) [59]	MiniGPT-4 (%) [59]	PandaGPT (%) [59]
ScienceQA (image) [48]	84.28	79.90	71.6	–	–	–	–	–
MMBench [49]	46.5	6.6	68.2	59.2	36.0	58.8	24.3	33.5
Image Captioning								
Dataset/Method	Metric	BLIP-2 [23]	Flamingo [81]	ZeroCap [84]	CLIPRe [84]	DeCap* [84]	ICSD [84]	CapDec [84]
MSCOCO [30]	BLEU-4	44.3	40	7.0	4.9	25.3	29.9	26.4
MSCOCO [30]	CIDEr	85.69	57.4	34.5	13.6	92.9	96.6	91.8
MSCOCO [30]	METEOR	–	–	15.4	11.4	25.2	25.4	25.1
MSCOCO [30]	ROUGE	–	–	31.8	29	51.1	52.7	51.8
NoCaps [50]	CIDEr	113.2	110	–	28.8	42.7	42.7	–
Flickr30k [35]	BLEU-4	–	–	5.4	5.2	20	25.2	17.7
Flickr30k [35]	CIDEr	–	–	16.8	10	49.3	54.3	39.1
Flickr30k [35]	METEOR	–	–	11.8	11.6	21.7	20.6	20
Flickr30k [35]	ROUGE	–	–	27.3	27.6	46.2	46.7	43.9
Document & Text-Centric VQA								
Dataset/Method	BLIP-2 acc (%)	PaLI [84] acc (%)	Gemini-2.5-Flash [68,85]	MiniCPM o2.6 (8B) [67]	Ovis2 (8B) [67]	InternVL3 (8B) [67]	Qwen2.5-VL (7B) [67]	VideoLLaMA3 Image (7B) [67]
DocVQA [51] (acc)	65.6	80	–	–	–	–	–	–
DocVQA [51] (ANLS)	–	–	93.20	92.38	93.77	92.05	94.04	91.73

(Continued)

Table 5 (continued)

Dataset/Method	VQA (Accuracy-Acc)							
	BLIP [80] (%)	BLIP-2 [23] (%)	Flamingo [81] (%)	PaLI (%) [82]	DinoV2 (%)	LLaVA (1.5) (%)	Kosmos-2 (%) [81]	GPT-4V (%)
DocVQA [51] (LLM)	–	–	93.03	90.87	91.61	90.09	92.84	89.97
ChartQA [52] (acc)	41.0	61	46.48	83.20	82.88	77.44	80.04	67.40
ChartQA [52] (LLM)	–	–	55.86	74.82	76.14	71.32	75.90	63.46
InfographicVQA [53] (ANLS)	–	–	85.32	69.41	78.96	71.80	77.93	69.47
InfographicVQA [53] (LLM)	–	–	84.60	64.99	74.89	66.55	73.88	65.43
TextVQA [45] VQAS)	–	–	76.53	78.35	82.13	78.63	83.03	62.06
TextVQA [45] (LLM)	–	–	82.83	82.24	86.83	82.55	87.41	67.18
ChartQAPro [86] (acc)	–	–	49.99	29.72	32.65	29.24	34.72	28.62
ChartQAPro [86] (LLM)	–	–	58.04	31.66	36.47	30.87	32.40	31.00
VQAv2 [43] (VQAS)	–	–	56.46	77.18	76.85	74.90	80.67	57.85
VQAv2 [43] (LLM)	–	–	62.06	79.38	79.17	76.98	82.63	60.22

Based on the results summarized in [Tables 4–6](#), distinct performance trends are evident that are directly linked to the architectural design of the VLMs. Instead of a single model excelling in every respect, specific architectures exhibit clear superiority depending on the task requirements: (1) Strengths in Zero-Shot Retrieval and Classification: As seen in [Tables 3 and 4](#), embedding-based VLMs (such as CLIP, ALIGN, and SigLIP) are extremely efficient for zero-shot classification and multimodal retrieval. By mapping images and text to a common embedded space using a contrast loss function, they enable cosine similarity matching at extremely high speeds. However, when maximum accuracy is required, hybrid architectures like CoCa (combining contrast learning with autoregressive decoders) have set a new standard (e.g., achieving Recall@1 91.8% on MSCOCO, surpassing CLIP's 88.1%). (2) Dominance in Complex Reasoning and VQA: The results in [Table 5](#) clearly show that for reasoning-heavy tasks (ScienceQA, MMBench) and text-centric VQA (DocVQA, ChartQA), the VLM + LLM architecture is overwhelmingly dominant. First-generation Generative VLMs and Cross-modal Transformers struggle significantly with complex logic. Conversely, models like GPT-4V, Qwen2.5-VL, and LLaVA achieve exceptionally high scores (e.g., ANLS on DocVQA > 90%) because they successfully transfer the cognitive reasoning burden to the billions of pre-trained parameters of the large language model (LLM). (3) Suitability for Multilingual Adaptation and Low-Resource Languages: For low-resource languages like Vietnamese, training a foundational VLM from scratch is extremely expensive and often encounters the “Alignment Tax” problem due to the scarcity of image-text pair data. Trends in the table suggest that a projection-based VLM + LLM architecture (similar to LLaVA or the proposed Vintern-1B model) is the most optimal model. By freezing the image encoder and multilingual LLM (or applying PEFT techniques like QLoRA) and training only a lightweight converter (MLP Projector), these models can efficiently map image tokens into the rich semantic space of the LLM. This helps overcome the barrier of a large-scale Vietnamese multimodal data shortage by leveraging the inherent linguistic fluency of LLM, thereby perfectly justifying our architectural choice in the experimental section ([Section 5](#)).

Table 6: The image-text retrieval results of the CLIP, ALIGN, SigLIP, Llip6, MetaCLIP4, and CoCa methods on the MSCOCO [30] and Flickr30K [35] datasets for text-to-image and image-to-text purposes on the Recall@1 metric.

Dataset/Method	Task	Image-Text Retrieval (Recall@1)					
		CLIP (%) [87]	ALIGN (%) [87]	SigLIP (%) [88]	Llip64 (%) [88]	MetaCLIP4 (%) [88]	CoCa (%) [29]
MSCOCO [30]	Text → Image	72.2	78.6	77.0	79.7	77.0	82.0
MSCOCO [30]	Image → Text	88.1	89.7	89.1	90.3	87.9	91.8
Flickr30K [35]	Text → Image	95.2	96.8	94.1	96.2	94.6	97.7
Flickr30K [35]	Image → Text	99.4	99.7	89.1	99.6	98.9	99.9

3.5 Challenges of VLM

During our research on VLM, we identified several core challenges. These also directly motivated our experimental architectural and design choices in Section 5: (1) Weak Grounding and Image Recognition: This problem arises from multimodal distance, for example, in the image encoder module of the CLIP/SigLIP model, where compressing images into vectors results in the loss of object details and complex relationships between them. When the image signal is weak, the LLMs within the VLM Generation and VLM + LLM architecture tend to rely on “language priors” to fill in information, leading to grounding (this is often revealed through checkers like POPE). Another cause is due to the bias of the dataset (for example, the data often associates a fork with a knife, causing the model to infer a knife even though there is no knife in the image). To directly address this issue in our experiments (Section 5), we constructed the VQA-TQU1 dataset with strictly manually verified specialized descriptions, forcing the model to adhere closely to visual reality rather than guesswork based on hallucinatory linguistic habits. (2) Spatial Inference Limitations: This is mainly due to the limitations of the ViT architecture being used on most Multimodal Transformers and VLM + LLMs. ViT handles flat image arrays and often struggles to understand relative geometric positions (left/right, top/bottom). Because the fixed resolution of standard visual encoders is often quite small (e.g., 224×224 or 336×336), small objects or detailed text will be blurred and grouped into a single array, making models prone to failure in inference-based tests such as ChartQA or DocVQA. This architectural limitation justifies our choice of the Vintern-1B model for the experimental part (Section 5.1). This model employs a “dynamic high-resolution” image cropping mechanism (dividing a large image into multiple 448×448 segments) to effectively preserve the geometric spatial structure, thereby enabling the most accurate OCR. (3) Computational Cost and Token Explosion: High-resolution image processing leads to token explosion, causing the computational cost of the Attention mechanism to increase exponentially to $O(N^2)$. Splitting high-resolution images into small chunks increases the number of image tokens from a few hundred to thousands, quickly leading to VRAM overflow (OOM) and occupying the entire context window of the LLM. This makes the deployment of large-scale Embodied VLMs or huge VLM + LLMs (such as GPT-4V) on edge devices impossible. This hardware bottleneck directly shaped our experimental methodology (Section V.B): instead of fine-tuning all parameters, we used Parameter Effective Fine-tuning (PEFT) through QLoRA (4-bit quantization) on a compact 1 billion parameter model. This allowed us to successfully train the system on a standard consumer GPU (NVIDIA RTX 3060-12 GB VRAM). (4) Multilingual Adaptation and “Alignment Tax”: The performance of models often degrades significantly in low-resource languages like Vietnamese. Because the foundational image encoders (like CLIP) in Embedding-based VLMs are mostly pre-trained with English image-text pairs, attempting to directly map Vietnamese semantics to this English-centric image space leads to a significant loss of cultural nuances (for example, the culturally iconic “Áo Dài” costume is only recognized by the model as a generic “váy dài”). This multilingual barrier is the fundamental motivation behind our efforts to build the VQA-TQU1 dataset. By refining the MLP projection

layer using high-quality, pure Vietnamese multi-turn dialogues, our research setup successfully overcame the “alignment tax,” allowing the model to communicate naturally and accurately within the specific context of Tan Trao University.

4 Problems in Training the VLM

VLM training is a complex process requiring a sophisticated combination of machine learning techniques, architectural design, and multi-modal data processing. Important considerations when training VLMs [1] are presented in Fig. 6. However, this process can be systematized into practical steps to effectively deploy the model. This section provides comprehensive guidance from theory to practice through core components: architecture selection, data pre-processing, training strategy, and performance evaluation methods.

In VLM training, data is the determining factor and is divided into three main groups. (a) Web-scale datasets (LAION [89], CC12M [38]), although noisy, provide a broad knowledge base, helping models learn concepts and generalize. (b) Technically labeled datasets (ImageNet [90], COCO [56], Visual Genome [91]), despite their small scale, are highly accurate and play a key role in improving image grounding and understanding image structure. (c) Synthetic datasets generated by AI help supplement rare cases with high-quality data, thereby enhancing inference capabilities and reducing reliance on real-world data. The question of computational resources is crucial because it determines the budget required to train the model. CLIP [4] and OpenCLIP [21] used over 500 GPUs for training. The cost of this amount of resources on a public cloud is equivalent to hundreds of thousands of USD, making it inaccessible to most companies or academic labs.

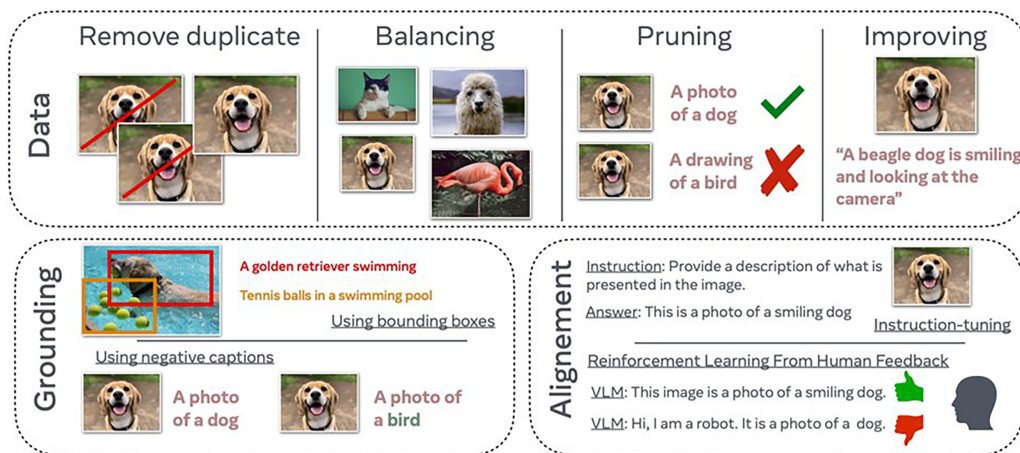


Figure 6: Illustrating the important considerations when training VLMs [1].

Recent software developments, such as ¹ torch.compile from PyTorch [92], significantly speed up training. Additionally, the xformers library [93] uses a more efficient attention mechanism to further accelerate training. A frequently overlooked bottleneck is data loading. When loading large mini-batches, data loading can significantly slow things down. Large-scale data is often stored in compressed tar files, which must be decompressed during loading, further slowing it down. It is recommended to store as many uncompressed files as possible. Furthermore, the FFCV library [94] creates faster-loading data files, significantly speeding up training compared to web datasets. The downside is larger storage space, but this is compensated for by computing power savings. Masking also helps improve training efficiency for large models. When using models with hundreds of millions or billions of parameters, forward-backward integration is costly.

¹Introduction to torch.compile—PyTorch Tutorials 2.11.0+cu130 documentation.

Currently, there are many different approaches to training VLM, each with its own advantages, limitations, and suitable applications. Contrastive models, such as CLIP [4], learn to associate text with images in a shared representation space, allowing for bidirectional data retrieval. Their advantages include creating meaningful representations, providing good support for grounding and zero-shot classification, and forming the basis for complex architectures. However, their disadvantages include the inability to generate new captions (only selection) and the significant data and computational resource requirements. This method is suitable for retrieval tasks and learning general representations. Masking models learn multimodal general distributions by reconstructing masked tokens in images or text, requiring a decoder to calculate the reconstruction loss. The main advantage is the absence of negative examples, reducing batch size and computational cost, and providing good support for generation and inference tasks. However, decoders can create performance bottlenecks, and large-scale training remains resource-intensive. This method is suitable for associative representation learning and reducing reliance on negative patterns.

Generative models (diffusion or autoregressive) learn a common distribution between text and images to generate new data (images from text or *vice versa*). The main advantages are the ability to support complex inference, generate data for interactive systems, and easily visualize learning outcomes. However, this method consumes significantly more computational resources than contrastive or masking models. Main applications include generative research, world model building, and multimodal comprehension assessment. The method using pre-trained encoders (such as Llama, GPT) focuses on learning the mapping between the representation spaces of two models. The main advantage is resource-efficient and easy to implement for alignment studies. However, the model may inherit hallucinations or biases from the existing backbone. This approach is suitable when resources are limited or for building lightweight VLMs. Grounding is a core challenge in VLM and generative models, addressing the issue of models not fully understanding prompts, leading to missed information or the generation of inaccurate content (such as errors in spatial relationships, negation, quantity, or attributes). While there is no definitive solution, the use of annotation bounding boxes is proving to be an effective technique for improving the alignment between text and images.

A prime example of this approach is X-VLM [95], a model that combines bounding box information with box regression and Intersection over Union (IoU) loss functions to accurately locate and align linguistic concepts with their corresponding image regions. Trained on a rich dataset of 16 million images from COCO [56], Visual Genome [91], SBU [96], and Conceptual Captions [38], X-VLM has demonstrated superior performance compared to current methods across a range of tasks such as image-text retrieval, visual inference, visual grounding, and image annotation generation. Inspired by the success of NLP [97] (Non-Linguistic Programming), current visual-linguistic models (VLM) actively apply instruction-based fine-tuning and RLHF to enhance multimodal interaction and ensure outcomes that meet human expectations. Instruction tuning involves refining models on small-scale labeled datasets containing instruction-output structures (ranging from a few thousand to 100,000 samples). Typical models for this approach include LLaVa and InstructBLIP [98]. In parallel, RLHF uses a reward model to adjust output based on human evaluation. A typical example is LLaVARLHF [99], a model that applies the Factually Augmented RLHF algorithm (combining data from GPT-4 and human evaluation) to minimize hallucinations and improve alignment. Experimental results show that LLaVA-RLHF achieves 94% of the performance of GPT-4 on the LLaVA Bench and outperforms baselines on the MMHAL-BENCH by over 60%. Understanding text within images is a key capability of computer vision. While current MLLMs are capable of zero-shot OCR, they often struggle with complex text due to the pre-training data being primarily natural images. To overcome this limitation, specialized solutions have been developed. First, there's instruction tuning with text-rich data, exemplified by LLaVAR [100]. This model combines poster and book cover images with OCR and GPT-4 to generate 16,000 Q&A conversations, significantly improving text-based VQA

capabilities. Second, there's a high-resolution image processing solution like Monkey [101]. This model supports 1344×896 pixels input by splitting images into patches, using resamplers and LoRA to create multi-level descriptions, thereby extracting more detailed information more efficiently. Finally, Lumos [102] proposes a separate architecture where the text recognition module (STR) runs on the device and outputs results to the MLLM in the cloud. This approach allows for processing ultra-large images ($3k \times 4k$) and optimizes performance in complex text tasks.

Due to the ever-increasing size of pre-trained models, making it impossible to fine-tune all parameters, parameter-efficient fine-tuning (PEFT) methods have become essential. This strategy focuses on training only a small portion of parameters to adapt to the downstream task, including the main approaches: LoRA and variants (LoRA [103], QLoRA [104]) using low-rank matrices to optimize costs; prompt-based (CoOp [105]) fine-tuning the input prompt for frozen models; adapter-based (CLIP-Adapter [106]) inserting lightweight modules between network layers; And mapping-based only learns mappings between single-mode modules. Therefore, PEFT allows for efficient optimization on limited resources while maintaining generalizability and high performance on multimodal tasks.

5 Experimental of VLMs

The rapid development of VLM is driving a shift from static image processing to more complex multimedia data. The need to understand motion, temporal interaction, and event sequences demands new, flexible approaches capable of simultaneously processing spatio-temporal information, audio, and dynamic context, rather than being limited to traditional image-text pairs. In this context, besides scaling models to handle complex data, another crucial approach is optimizing the architecture to bring multimedia capabilities to edge devices. Balancing inference performance and hardware resources, while specializing in specific languages (such as Vietnamese), is becoming a practical trend. To build descriptive information for answering VQA-related questions, serving the purpose of developing an AI system that automatically answers user questions. For example, building a system that automatically answers students' questions about admissions at Tan Trao University in Vietnam. In this study, we experimented as shown in Fig. 7.

5.1 Models for VQA

- 1) **Overview of Vintern-1B architecture:** In the context of the Generative AI boom, the hardware cost barrier of large MLLMs has spurred the creation of Vintern-1B [107], a solution developed and optimized in depth by the 5CD-AI team for the Vietnamese market. With a size of approximately 1 billion parameters (1B), Vintern-1B's strategy is not about scale but focuses on maximizing efficiency. The model's architecture is designed to run smoothly on edge devices and personal computers with limited resources, as illustrated in Fig. 8.

Despite its compact size, Vintern-1B ensures superior accuracy in critical tasks such as Vietnamese OCR, document information extraction, and VQA. Vintern-1B employs a cascading (pipeline-based) approach to support image-speech interaction, where speech is not directly modeled in the latent space of the VLM. Instead, speech is processed through text-mediated intermediate modules, including ASR (Automatic Speech Recognition) and TTS (Text-to-Speech), as illustrated in Fig. 9.

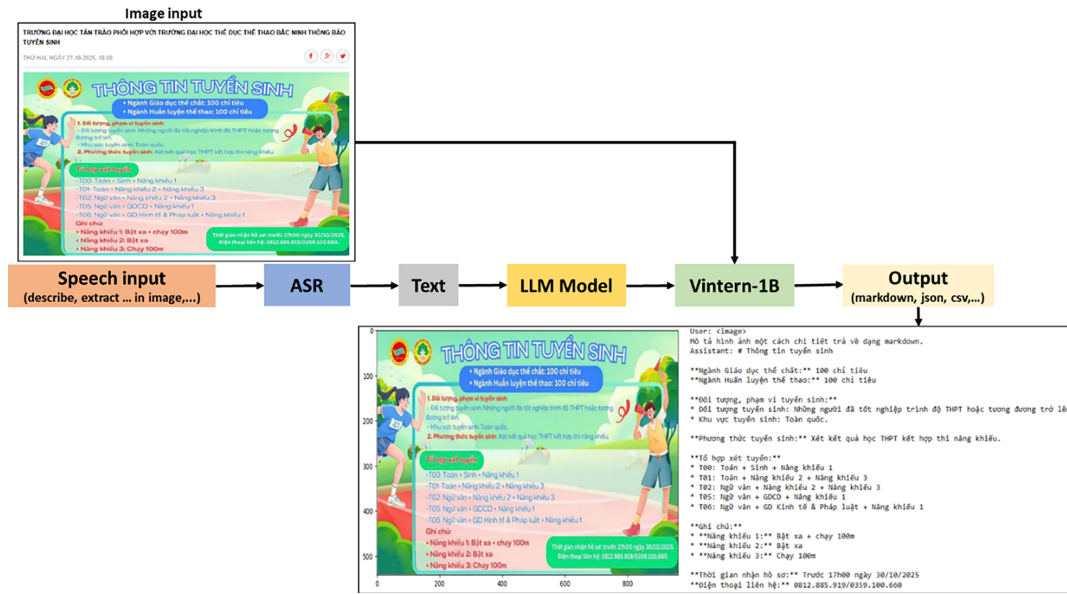


Figure 7: The data processing model is based on Vintern-1B for automated question answering.

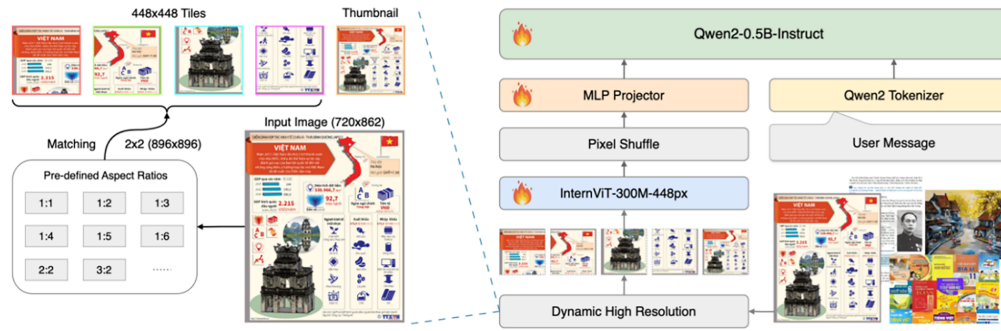


Figure 8: Vintern-1B model architecture.

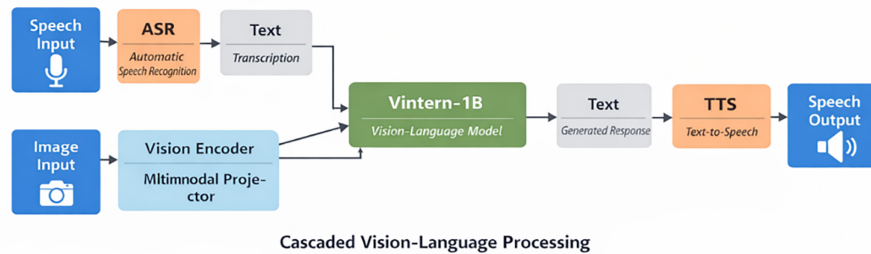


Figure 9: Cascading architecture of the Vintern-1B model.

In Fig. 8, with speech and image as input data, the speed data is converted to text based on the ASR model. Image data features are extracted using a visual encoder; visual tokens are used to convert from the VE to a multimodal projector; the multimodal projector maps image features to the token space of the LLM in the VLM; and the LLM with 1B parameters performs reasoning and text generation, as illustrated in Fig. 10.

The objective function of the text generation process is shown in the [Formula \(1\)](#).

$$P(y_t | y_{<t}, X_{text}, X_{vision}) \quad (1)$$

where X_{vision} is the embedding of the projected image into the LLM space, X_{text} is the token embedding of the prompt/ASR text, and y_t is the token generated at step t .

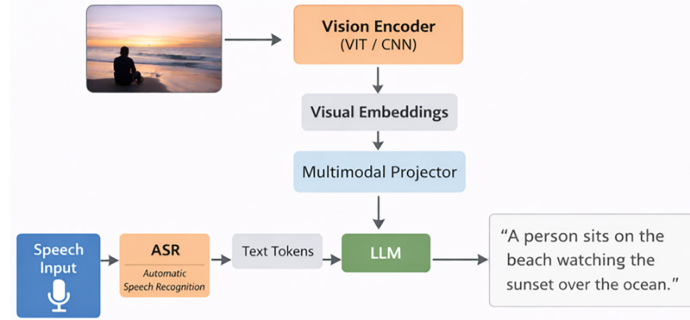


Figure 10: The architecture performs reasoning and text generation of the Vintern-1B model.

In this model, LLM acts as an autoregressive decoder, responsible for combining visual information (from images), combining textual context (prompt/ASR text), and generating output text strings based on each token. Vintern-1B uses autoregressive factorization as a [Formula \(2\)](#).

$$P(Y | X_{image}, X_{text}) = \prod_{t=1}^T P(y_t | y_{<t}, X_{image}, X_{text}) \quad (2)$$

where X_{image} is the visual embeddings (prefix tokens), X_{text} is the text tokens (prompt/ASR output), and y_t is the token generated at the t step.

The detailed architecture of the autoregressive LLM, following the transformer model, includes token embedding, positional encoding, masked self-attention, a feed-forward network, and an output head. The masked self-attention at each layer l is calculated using the [Formula \(3\)](#).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + M\right)V \quad (3)$$

where Q (query) is the query matrix, representing the “what is being searched for” at each location, its size is usually $N \times d$, K (key) is the key matrix, representing features to match the query, with a size of $N \times d$, V (value) is the value matrix, containing information aggregated after attention, with a size of $N \times d_v$, QK^T is the similarity between each Query and all Keys, d is the dimension of the key/query vector, dividing by $\sqrt{d}$ helps stabilize the gradient (scaled dot-product attention), M (mask) is the mask matrix with “causal mask” preventing the model from seeing the “future” (used in the decoder) and Padding mask ignoring token padding, $\text{softmax}(\cdot)$ normalizes the attention weights into a probability distribution.

The loss function for the autoregressive decoder is presented in [Formula \(4\)](#).

$$\mathcal{L}_{\text{VLM}} = - \sum_t \log P(y_t | y_{<t}, X_{image}, X_{text}) \quad (4)$$

where y_t is the token generated at step t , $y_{<t}$ is the previous token string, X_{image} is the image feature (visual embeddings), X_{text} is the input text (prompt/ASR output). Vintern-1B is developed based on the

advanced InternVL architecture, adhering to the “V-L-L” (Vision-Link-Language) design model with three core components working in synchronization. The first component is the VE, using the InternViT-300M-448px model. Acting as the “eyes” of the system, this encoder converts input pixel data into feature vectors (visual embeddings) containing semantic information about shape, color, and text. Next, the MLP Projector acts as an intermediary bridge, aligning the visual feature space to a semantic space that the language model can understand. Finally, the central logic processing unit is the Qwen2-0.5B-Instruct Backbone. This is a small-scale model (under 1 billion parameters) that has been fine-tuned extensively by the 5CD-AI team using Vietnamese data, ensuring powerful inference capabilities and natural, coherent text generation.

The Vintern-1B operating process is a tightly controlled multimodal pipeline, notable for its dynamic resolution mechanism. To address the problem of information loss when processing text-dense documents, the model applies a “tiling” technique: dividing large images into 448×448 pixels tiles to process each part in detail before reassembling them. At the central processing layer, image feature vectors (from the MLP Projector) are concatenated with text vectors from user commands. The LLM’s attention mechanism then analyzes the relationship between the image region and the question to generate the final result, which can be OCR text, a natural answer, or structured data (JSON, Markdown). Vintern-1B is highly modular and easy to upgrade in parts (replacing ASR with a better one doesn’t require retraining the entire VLM model). However, its drawbacks include latency due to the need to wait for ASR to finish before VLM can be processed, error propagation due to incorrect ASR (e.g., “cat” -> “cap”) leading to completely incorrect VLM, and loss of non-verbal information such as emotions and intonation. A major advantage of Vintern-1B is its high optimization for mainstream hardware. Thanks to its total size of only about 1 billion parameters, the model has extremely low resource consumption: only about 4–6 GB of VRAM when operating in fp16 accuracy mode. This allows for real-time model deployment on mid-range gaming graphics cards (such as NVIDIA RTX 2060, 3060, 4060) or edge devices like Jetson Orin, eliminating reliance on expensive server infrastructures like Tesla A100. The code of Vintern-1B is shown in the path².

- 2) **ASR Model:** The increasing demand for voice interaction has revealed the limitations of traditional ASR solutions in the face of the complex six tones and dialects of the Vietnamese. To address this issue, the technological trend has shifted strongly towards end-to-end DL architecture, with Wav2Vec 2.0 (Meta AI) being a prime example. The core strength of Wav2Vec 2.0 lies in its self-supervised learning mechanism, which allows the learning model to represent context from large amounts of unlabeled audio data, thereby significantly reducing reliance on scarce and expensive labeled data sources. To optimize the practical performance for Vietnamese, the study [108] was pre-trained on over 13,000 h of unlabeled Vietnamese audio from YouTube and used the standard VLSP ASR³ dataset (250 h of labeled data) sampled at 16 kHz for fine-tuning. Instead of costly initial training, the strategy employed was to fine-tune the pre-trained weights in combination with the Connectionist Temporal Classification (CTC) loss function. This approach helped the model achieve a dual objective: rapid convergence speed and high accuracy even when operating with limited training data. The Wav2Vec 2.0 model employs an end-to-end architecture, allowing for the direct conversion of raw audio signals (waveforms) into text without intermediate steps such as manual phoneme extraction. The processing flow begins with a Feature Encoder built from a multi-layer convolutional neural network. This block’s task is to compress the time-amplitude sequence into latent speech representations, while simultaneously performing downsampling and background noise filtering to optimize the input.

²https://drive.google.com/drive/folders/1dfgyNGnRtyXsdMegnhq1NU7FnPbH_tvl

³<https://vlsp.org.vn/vlsp2020/eval/asr>

At the heart of the architecture is the Context Network, which includes Transformer Encoder classes using a self-attention mechanism similar to BERT. This block is responsible for learning global contextual relationships along the time axis, allowing the model to capture dependencies between phonemes even when they are far apart. Finally, the output is passed through the Linear Projection and CTC head layer. Here, the CTC loss function solves the asynchronous problem between the length of the audio sequence and the character, automatically aligning and removing empty tokens to produce a complete text sequence.

- 3) ***SmolVLM2 Model for VQA:*** The architecture of SmolVLM2 [109] is not simply a scaled-down version, but the result of sophisticated engineering coordination between three core components. (a) The VE uses the SigLIP architecture instead of the standard CLIP. By eliminating the costly softmax function in the pre-training process and replacing it with pairwise Sigmoid loss, SigLIP allows the model to learn more “dense” image features at lower computational cost. This improvement significantly enhances the model’s ability to recognize fine details in video without massively increasing parameter sizes. (b) To address the memory limitations when processing long video sequences, the Modality Connector employs an advanced token compression mechanism called Pixel Shufe. Instead of loading all tokens from the VE into the LLM—which can easily cause memory overflow when processing up to 64 frames—the model combines adjacent image patches to reduce the sequence length by four times or more. This technique allows for the inclusion of a larger amount of information into the context window while preserving the semantic integrity of the input data. (c) Acting as the “brain” for language processing is the Text Decoder based on the SmolLM2 backbone (a minor variant). This component is trained on high-quality synthetic datasets such as Cosmopedia v2 and FineWeb Edu, equipping the model with strong logical reasoning and instruction-following capabilities despite operating at an extremely small parameter scale. This is key to enabling SmolLM2 not only to understand visual content but also to generate coherent, natural-sounding descriptive passages instead of disjointed words.

5.2 Data Collection for Fine-Tuning the VLM Model of Vietnamese

In this study, we collected a dataset of the admissions activities of Tan Trao University, Tuyen Quang province, Vietnam, called VQA-TQU1⁴. To ensure the reproducibility and validity of the refinement process, the construction and evaluation protocols for the VQA-TQU1 dataset were carefully designed as follows: (1) Dataset Construction and Quality Control: The raw image and text data of the VQA-TQU1 dataset were collected from the official website of Tan Trao University (2024–2026). Initially, the raw collection contained more than 184 records. Then, we applied a rigorous quality control process with human participation: removing duplicate IDs, eliminating low-resolution images, and refining ambiguous texts. The final refined dataset includes exactly 180 high-quality color images closely related to the university’s landscape, architecture, and admissions activities. (2) Annotation Process and Interaction Definition: To simulate real-world human-machine dialogues, the VQA-TQU1 dataset was manually annotated in a multi-turn dialogue format (JSONL). We explicitly defined an “interaction” as a single multi-turn exchange. For each of the 180 images, the annotation included 1 general description statement (image caption) and an average of 5 specific question-answer pairs simulating student queries. This resulted in exactly $180 \times (1 + 5) = 1080$ Vietnamese interactions. The questions showed high diversity, covering many areas such as admission criteria, school vision, organizational structure and logo identity, with each answer containing approximately 100 words in pure and natural Vietnamese, the processing structure of the dataset is illustrated in Fig. 11. (3) Training/Validation/Testing Split Strategy: To rigorously evaluate the model’s generalization ability, the VQA-TQU1 dataset was partitioned at the image level instead of the interaction level to prevent data leakage. Specifically, 180 images were randomly split into: 150 images for training (~83%), 15 images for validation

⁴https://drive.google.com/drive/folders/1dfgyNGnRtyXsdMegnh1NU7FnPbH_tv1

(~8.5%), and 15 images for testing (~8.5%). This rigorous partitioning at the image level ensures that the model is evaluated on image contexts that have never appeared during the testing phase (avoiding model memorization). (4) Argument on dataset size: Although 180 images might seem like a limited number for training a model from scratch, the VQA-TQU1 dataset is sufficiently large for Parameter Effectiveness Fine-Tune (PEFT) on a robust platform model. By using QLoRA techniques on the Vintern-1B model, the goal was not to teach the model general Vietnamese language or basic visual perception (which it already possessed), but to fine-tune subject knowledge and stylistic appropriateness to the context of Tan Trao University. High-quality, densely annotated data (1080 rich interactions) is of superior value compared to the large but noisy dataset in this specialized fine-tuning scenario.

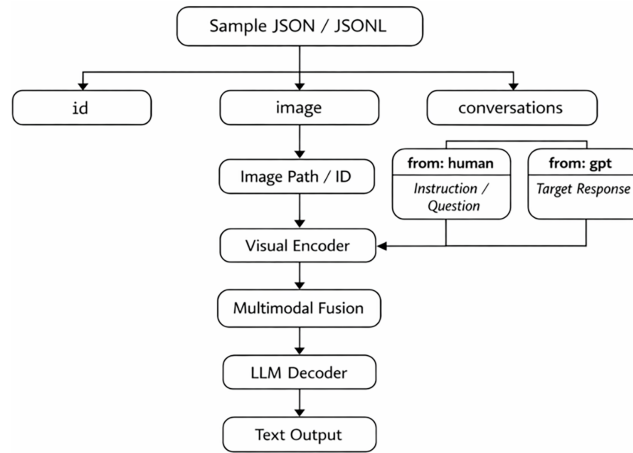


Figure 11: The processing structure of the VQA-TQU1 dataset.

1) *Fine-Tuning VLM Model of Vietnamese*

To train the Vintern-1B model (consisting of Qwen2-0.5B + InternViT-300M) on an RTX 3060 (12 GB VRAM), we used QLoRA (4-bit) instead of bfloat16 to save VRAM. The parameter “bnb_4bit_quant_type: nf4” sets the normal float to 4 for optimal weight distribution in the LLM. Next, “bnb_4bit_compute_dtype: torch.bfloat16” means the computation is performed in 16-bit to maintain accuracy, but stored in 4-bit to save VRAM. The parameter “bnb_4bit_use_double_quant: True” saves an additional 0.4 bits/parameter. In the LoRA Config (adapter configuration), the following is set: r is 128, a high level to help the model learn fine details well; “lora_alpha” is 256, usually set to double the rank to stabilize the gradient; “lora_dropout” is 0.05; the training process should train all linear classes of the language module (Qwen2) with “target_modules: [‘q_proj’, ‘k_proj’, ‘v_proj’, ‘o_proj’, ‘gate_proj’, ‘up_proj’, ‘down_proj’]” for best performance instead of just training q_proj and v_proj. The notebook file contains the “dynamic_preprocess” function, which is key to this model with a dynamic high-resolution mechanism, meaning the input image is not roughly resized to 224×224 . Instead, the image is divided into small tiles of size 448×448 . The parameters in the file are “min_num = 1, max_num = 12”. For example, a long image can be cropped into up to 12 448×448 segments, which allows the model to read small text (OCR) extremely well. However, the challenge is that with an RTX 3060 GPU, 12 image segments will generate approximately 3000+ visual tokens, plus the text context, making it very easy to cause an overflow (OOM) of the GPU’s 12 GB. Therefore, if an OOM error occurs, reduce “max_num” to 6 or 8 in the “dynamic_preprocess” function. The hyperparameters of the training process are: batch size: 1; “gradient_accumulation_steps”: 16, Learning Rate: $2e-4$, Optim: “paged_adamw_32bit”, Gradient Check pointing is True to trade speed for VRAM and reduce memory usage by 50%–60%, epochs: 3.

2) *Evaluations, Results, and Discussions*

In this study, we evaluated several metrics as follows:

- Bilingual Evaluation Understudy (BLEU) measure: BLEU is a method for measuring the degree of overlap of n-gram clusters between a predicted text and a reference text. Although initially designed for machine translation, BLEU is now widely used in OCR research to assess the structural fidelity of text. The BLEU calculation formula is based on a combination of adjusted n-gram precision and a shortness penalty (BP) to avoid unreasonably high scores for excessively short responses, as the [Formula \(5\)](#).

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^k w_n \log p_n \right) \quad (5)$$

where p_n is the corrected n-gram accuracy, w_n is the weight for each n-gram dimension, and BP is the penalty factor if the predicted text length is shorter than the original text.

- Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) measure: ROUGE-L assesses the similarity between two texts through the Longest Common Sub-sequence (LCS). By combining both Precision and Recall metrics, ROUGE-L can account for the order of words in the sequence. This characteristic makes ROUGE-L particularly effective when evaluating texts with specific structures, such as ingredient lists. The F1 score of ROUGE-L is defined as the [Formula \(6\)](#).

$$\text{ROUGE-L}_{F_1} = \frac{(1 + \beta^2) R \cdot P}{R + \beta^2 P} \quad (6)$$

where R represents coverage, P represents accuracy, and β is typically set to 1 to balance the weights between these two metrics.

- F1 Score (F1): F1 is calculated at the word level, and is the average of Precision and Recall, emphasizing the exact word-by-word matching between two texts. F1 plays a crucial complementary role to BLEU and ROUGE-L by focusing on the close matching of lexical tokens. This is extremely essential when processing structured data fields requiring absolute precision, such as nutritional information tables. The formula for calculating the F1 is computed in [Formula \(7\)](#).

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

- METEOR is a word-level metric. It matches tokens between PRED and GT and calculates them using a combination of precision and recall. It may perform better than BLEU in many cases because it's slightly more flexible with word order and some variations in wording. However, it still relies heavily on tokenization, and the metric's native language is English, so in Vietnamese, it should only be considered a reference indicator and not the primary metric.
- Char-F1 is F1 at the character level. The calculation is similar to regular F1, but instead of separating by word, it compares each character after normalization. Because Vietnamese has accents, spaces, and word separation is not entirely consistent, Char-F1 usually reflects the degree of similarity ("na ná") between two sentences more accurately than word-level F1.

The testing process was conducted on a mid-range hardware configuration (NVIDIA RTX 3060-12 GB VRAM) with input data consisting of a joint admissions announcement from Tan Trao University and Bac Ninh University of Physical Education and Sports, and the requirement to output in Markdown format. The results showed that Vintern-1B's text extraction capability (High Accuracy OCR) achieved absolute accuracy

(100%) on key information fields. The most valuable aspect was its perfect handling of Vietnamese language features, eliminating font errors or character confusion often seen in other multilingual models.

a) *Vintern-1B results*

To rigorously validate the effectiveness of the proposed refinement method, we expanded the evaluation scope by comparing the refined Vintern-1B with three unrefined (zero-shot) base models: InternVL-1B, Mini-InternVL-2B, and the Vintern-1B Base platform model. We also incorporated METEOR and Char-F1 indices to better assess sentence fluency and character-level OCR accuracy. The average results from the VQA-TQU1 dataset when evaluating the BLEU, ROUGE-L, and F1-Score measures are shown in Table 7. The results on the BLEU, ROUGE-L, METEOR, F1, and Char-F1 measures were 0.5677, 0.739, 0.7532, 0.7435, and 0.7461, respectively. These results were all higher than those of the following methods: InternVL-1B (Zero-shot) [110], Mini-InternVL-2B (Zero-shot) [111], Vintern-1B Base (Zero-shot). However, these results are not high, but they are more favorable than the results in the study by Nagayi et al. [112] (BLEU = 0.245, ROUGE-L = 0.391, F1-Score = 0.345) with the Tesseract model, or in the study by Rahimi [113] (BLEU-L = 0.453, ROUGE-L = 0.196, F1 = 0.212).

Table 7: Average evaluation results on the VQA-TQU1 dataset.

Metric	InternVL-1B (Zero-Shot) [110]	Mini-InternVL-2B (Zero-Shot) [111]	Vintern-1B Base (Zero-Shot)	Vintern-1B Fine-Tuned (Ours)
BLEU	0.0838	0.0607	0.3183	0.5677
ROUGE-L	0.331	0.3353	0.5718	0.7397
METEOR	0.2932	0.2482	0.5885	0.7532
F1	0.2799	0.2659	0.5737	0.7435
Char-F1	0.5011	0.6184	0.7196	0.7461

The results in Table 7 reveal significant performance gaps. The generalized language-visual models InternVL-1B and Mini-InternVL-2B struggled significantly with the VQA-TQU1 dataset, achieving extremely low BLEU scores of 0.0838 and 0.0607, respectively. A detailed error analysis indicated that these failures were primarily due to the limited Vietnamese language range. Specifically, these base models frequently produced incorrect Vietnamese diacritics, misspellings, or failed to generate coherent syntactic structures when extracting text from Vietnamese admissions documents. Their very low Char-F1 scores (0.5011 and 0.6184) confirm their inability to maintain character-level accuracy in low-resource languages.

The Vintern-1B Base model, having been pre-trained on more Vietnamese datasets, showed a significantly stronger foundation (Char-F1: 0.7196, BLEU: 0.3183) but still lacked the specialized inference capabilities necessary for the context of Tan Trao University. Following targeted refinement, our proposed model achieved a significant performance leap, reaching a BLEU score of 0.5677 and a METEOR score of 0.7532. The refinements thoroughly addressed the Vietnamese diacritics and vocabulary errors encountered in the base models. Through analysis of the rare instances of incorrect predictions in the refined model, we found that the errors were no longer due to limitations in the Vietnamese language. Instead, they stemmed from weak grounding and hallucinations when encountering highly stylized, ultra-small, or blurry text in complex infographics. In these extreme cases, the model occasionally relied on linguistic preferences (guessing) instead of following clear visual evidence. Overall, these results strongly support the need for localized instruction tuning when deploying VLMs in low-resource language environments.

Table 8 shows illustrative results in a case study with input image data (first column), input text (second column), and output results (third column) with question (Q), original data (GT), and predicted response (PRED).

Table 8: An example in the VQA-TQU1 dataset testing on a model.

Image	Description	Conversations
	Hình ảnh là một tài liệu giới thiệu chi tiết về Trường Đại học Tân Trào (Tân Trao university-TTrU). Góc trên bên trái là ảnh chụp thực tế cổng trường với dòng chữ "ĐẠI HỌC TÂN TRẢO" màu đỏ nổi bật và khẩu hiệu "VÌ LỢI ÍCH TRẦM NĂM TRÔNG NGƯỜI". Phần văn bản bên phải và phía dưới trình bày các nội dung cốt lõi: Sự mạng (đào tạo, nghiên cứu khoa học, hợp tác quốc tế), Tâm nhìn đến năm 2030 (trở thành trung tâm đào tạo nguồn nhân lực chất lượng cao, nghiên cứu khoa học và chuyển giao công nghệ của tỉnh và khu vực), Chính sách chất lượng. Giá trị cốt lõi ("Chất lượng đào tạo là giá trị cốt lõi để tồn tại và phát triển bền vững"). Tiếp theo là mục Mục tiêu giáo dục mô tả định hướng, đào tạo người học toàn diện về kiến thức, kỹ năng và phẩm chất chính trị. Phần cuối là mục "Giới thiệu chung về nhà trường" với lịch sử hình thành từ năm 1959 đến năm 2013.	Q: Sự mạng và tâm nhìn đến năm 2030 của Trường Đại học Tân Trào được nêu trong tài liệu là gì? GT: Theo tài liệu, sứ mạng của trường là cơ sở đào tạo, nghiên cứu khoa học, hợp tác quốc tế và cung cấp các sản phẩm, dịch vụ nâng cao chất lượng cuộc sống. Tâm nhìn đến năm 2030 phấn đấu trở thành trung tâm đào tạo nguồn nhân lực chất lượng cao, nghiên cứu khoa học và chuyển giao công nghệ của tỉnh và khu vực. PRED: Trường Đại học Tân Trào là cơ sở đào tạo, nghiên cứu khoa học, hợp tác quốc tế và cung cấp các sản phẩm, dịch vụ nâng cao chất lượng cuộc sống. Tâm nhìn đến năm 2030 là trở thành trung tâm đào tạo nguồn nhân lực chất lượng cao, nghiên cứu khoa học và chuyển giao công nghệ của tỉnh và khu vực. Q: Giá trị cốt lõi mà nhà trường cam kết là gì? GT: "Chất lượng đào tạo là giá trị cốt lõi để tồn tại và phát triển bền vững." PRED: Giá trị cốt lõi được xác định là chất lượng đào tạo nhằm đảm bảo sự tồn tại và phát triển bền vững của nhà trường. Q: Bạn có thể tóm tắt lịch sử hình thành và phát triển của trường dựa trên văn bản không? GT: Trường được thành lập từ Trường Sơ cấp sư phạm năm 1959, sau đó nâng cấp thành Trường Trung cấp Sư phạm (1969), Cao đẳng Sư phạm Tuyên Quang (1999), Cao đẳng Tuyên Quang (2011) và chính thức trở thành Trường Đại học Tân Trào năm 2013. PRED: Trường Đại học Tân Trào có lịch sử phát triển từ năm 1959, trải qua nhiều giai đoạn nâng cấp và chính thức trở thành đại học vào năm 2013. Q: Mục tiêu giáo dục của trường hướng đến việc đào tạo người học như thế nào? GT: Mục tiêu giáo dục là đào tạo người học có kiến thức và kỹ năng chuyên môn vững chắc, thực tiễn, có phẩm chất chính trị và đạo đức tốt, có khả năng tự học suốt đời, thích ứng với sự phát triển của khoa học công nghệ và thị trường lao động. PRED: Trường hướng đến đào tạo người học toàn diện về kiến thức, kỹ năng thực hành, kỹ năng mềm, có khả năng thích ứng và hội nhập trong bối cảnh kinh tế-xã hội hiện đại.

Fig. 12 shows the visual results on an instance of the VQA-TQU1 dataset, with an input image, then a text recognition model is used to recognize the text in the image, and objects in the image, resulting in a descriptive response of the image (image captioning) after using the Vintern-1B model.



Figure 12: Illustrating the image captioning result of the VQA-TQU1 dataset after using the Vintern-1B model.

Furthermore, the model demonstrates superior structural understanding through an effective noise filtering mechanism: automatically removing unnecessary interface elements (logos, stickers) to focus on core content, while also separating data logic.

Finally, strict format compliance is ensured with standard Markdown output, allowing direct integration of results into Web/App systems without complex post-processing steps, significantly optimizing the software development process. Vintern-1B marks a significant step forward in the democratization of multimodal AI technology in Vietnam. The strategic combination of the powerful VE (InternViT) and the lightweight language model (Qwen2-0.5B) has established an optimal balance between performance and cost. This is considered an ideal solution for document digitization, automated information extraction, and the development of intelligent Vietnamese virtual assistants, feasible even on limited hardware platforms.

b) Wav2vec-2.0 ASR Results

This performance level is incredibly impressive for a large architecture like Wav2Vec 2.0 (Large/XLS-R-53 variant), confirming the optimal efficiency of GPU-based parallel computing. Recognizing the inherent semantic limitations of pure ASR models, the research team proposed a solution that integrates the API of the Big LLM—specifically, Gemini—acting as a post-processor. This mechanism leverages the superior contextual understanding and vast knowledge base of the LLM to review and correct the raw text output from Wav2Vec 2.0. Table 9 shows the results of the Wav2Vec 2.0 ASR model for speech-to-text when combined with the API of the LLM.

Experimental results show a groundbreaking improvement. All phonetic spelling errors (such as “vững chắc” → “vững chắc”, “y dựng” → “y dựng”) were automatically detected and corrected accurately based on the sentence context. In particular, previously misidentified local proper names (such as “kiên quang”) have now been perfectly standardized as “Tuyên Quang” and “Tân Trào”. Furthermore, LLM also performs advanced stylistic standardization operations such as adding punctuation, capitalizing first letters of sentences, and creating appropriate paragraph breaks, transforming raw text into a complete, coherent, and grammatically correct Vietnamese paragraph. The combination of Wav2Vec 2.0’s audio processing speed and Gemini’s linguistic intelligence has created a comprehensive speech recognition pipeline, significantly improving real-world accuracy of WER while keeping execution times close to the original model and ready for high-quality commercial applications.

Table 9: The results of the Wav2Vec 2.0 ASR model for speech-to-text when combined with the API of the LLM.

Original Text from ASR (Unprocessed)	Post-Processing Text (LLM Integration)
trải qua hành trình hơn sáu mươi năm xây dựng và phát triển đến nay trường đại học tân trào đã tạo dựng được vị trí vững chắc trong hệ thống giáo dục việt nam từng bước khẳng định được uy tín thương hiệu trong cộng đồng và xã hội trên hai mươi chương trình đào tạo đón đầu xu thế của thị trường lao động đa dạng ở các lĩnh vực sư phạm y dược kinh tế văn hóa du lịch nông lâm nghiệp thu hút sinh viên từ ba mươi tỉnh thành trong cả nước và nước ngoài theo học mạng lưới hợp tác quốc tế với hơn ba mươi trường đại học viện nghiên cứu và các tổ chức giáo dục trên thế giới tạo nên môi trường năng động hội nhập cho giảng viên và sinh viên một trăm phần trăm giảng viên của trường có trình độ từ thạc sĩ trở lên tỷ lệ tăng viên có trình độ tiến sĩ đáp ứng mức bình quân trung của cả nước gần ba nghìn sản phẩm khoa học ở trong nước và quốc tế với mục tiêu chất lượng đào tạo là yếu tố cấp lõi để tồn tại và phát triển trường chú trọng đầu tư nâng cấp cơ sở vật chất công nghệ nâng cao chất lượng đội ngũ để tổ chức đào tạo đa ngành chủ động đổi mới phương thức đào tạo theo hướng giảm giờ lý thuyết tăng giờ thực hành tăng cường sự tham gia của đơn vị sử dụng lao động vào quá trình đào tạo hàng trăm hoạt động sự kiện dành cho sinh viên với nhiều hình thức đa dạng đã tạo nên môi trường năng động toàn diện để sinh viên thực hành rèn luyện và phát triển vị thế của trường đại học tân trào đã và đang để lại dấu ấn trong hệ thống giáo dục đại học với tầm nhìn trở thành một trung tâm phát triển bền vững góp phần thay đổi cuộc sống và phục vụ cộng đồng trường đại học tân trào là địa chỉ tin cậy cung cấp nguồn nhân lực chất lượng cao cho công cuộc xây dựng và phát triển kinh tế xã hội của tỉnh kiênh quang và trên cả nước.	Trải qua hành trình hơn sáu mươi năm xây dựng và phát triển, đến nay Trường Đại học Tân Trào đã tạo dựng được vị trí vững chắc trong hệ thống giáo dục Việt Nam, từng bước khẳng định được uy tín thương hiệu trong cộng đồng và xã hội. Trên hai mươi chương trình đào tạo đón đầu xu thế của thị trường lao động đa dạng ở các lĩnh vực sư phạm, y dược, kinh tế, văn hóa, du lịch, nông lâm nghiệp thu hút sinh viên từ ba mươi tỉnh thành trong cả nước và nước ngoài theo học. Mạng lưới hợp tác quốc tế với hơn ba mươi trường đại học, viện nghiên cứu và các tổ chức giáo dục trên thế giới tạo nên môi trường năng động, hội nhập cho giảng viên và sinh viên. Một trăm phần trăm giảng viên của trường có trình độ từ thạc sĩ trở lên, tỷ lệ giảng viên có trình độ tiến sĩ đáp ứng mức bình quân chung của cả nước. Gần ba nghìn sản phẩm khoa học ở trong nước và quốc tế. Với mục tiêu chất lượng đào tạo là yếu tố cốt lõi để tồn tại và phát triển, trường chú trọng đầu tư nâng cấp cơ sở vật chất, không ngừng nâng cao chất lượng đội ngũ để tổ chức đào tạo đa ngành, chủ động đổi mới phương thức đào tạo theo hướng giảm giờ lý thuyết, tăng giờ thực hành, tăng cường sự tham gia của đơn vị sử dụng lao động vào quá trình đào tạo. Hàng trăm hoạt động, sự kiện dành cho sinh viên với nhiều hình thức đa dạng đã tạo nên môi trường năng động, toàn diện để sinh viên thực hành, rèn luyện và phát triển. Vị thế của Trường Đại học Tân Trào đã và đang để lại dấu ấn trong hệ thống giáo dục đại học với tầm nhìn trở thành một trung tâm phát triển bền vững, góp phần thay đổi cuộc sống và phục vụ cộng đồng. Trường Đại học Tân Trào là địa chỉ tin cậy cung cấp nguồn nhân lực chất lượng cao cho công cuộc xây dựng và phát triển kinh tế xã hội của tỉnh Tuyên Quang và trên cả nước.

Besides accuracy, processing speed plays a crucial role in determining the feasibility of the model in real-world applications. Based on system logs with a total execution time of 40 s for an audio segment of 112 s (1 min 52 s), the RTF is determined as follows:

RTF is a metric that answers the question, “*Is the computer faster or slower than the speaker?*”. RTF is calculated by comparing [the time the computer takes to process] with [the actual length of the recording]. In an ideal state ($RTF < 1$), the system runs “*faster*” than the speaker. For example, if a 10-s recording takes the computer only 2 s to process, the result is $RTF = 0.2$. This application is often used in live streaming/conversation applications. In a slow state ($RTF > 1$), the system runs “*slower*” than the speaker. For example, a 10-s recording takes the computer 15 s to process. This is only acceptable when processing pre-recorded files online. RTF is calculated using the following [Formula \(8\)](#).

$$RTF = \frac{T_{\text{process}}}{T_{\text{audio}}} = \frac{40}{112} \approx 0.36 \quad (8)$$

With an *RTF* of approximately 0.36 (less than 1), the system achieves processing speeds about 2.8 times faster than real-time. This means decoding 1 h of audio data takes only about 21–22 min.

c) *SmolVLM2 Results*

The model was tested on a video introducing “*Trường Đại học Tân Trào*”⁵ with a duration of 1 min and 52 s. Quantitative results clearly demonstrate the advantages of the “*Smol*” architecture (500M parameters) when operating in a resource-constrained environment. Specifically, the model only took 27 s to process a 112-s (1 min and 52 s) video. The results of the SmolVLM2 model are described in English in the video “*Trường Đại học Tân Trào*” and are presented in [Table 10](#).

This performance is shown by the Real-Time Factor (RTF) calculated as in [Formula \(9\)](#).

$$RTF = \frac{\text{Processing time}}{\text{Video length}} \approx \frac{27}{112} \approx 0.24 \quad (9)$$

With an *RTF* of 1, the model achieves a performance speed four times faster than real-time, demonstrating ultra-fast responsiveness. In terms of load handling, for a stream of 64 regularly sampled frames as input, the model simultaneously performs vision encoding and language decoding at an average speed of approximately 0.42 s per frame.

This impressive performance for a video LMM model demonstrates the feasibility of the SmolVLM2 [\[114\]](#) for deploying real-time video analytics applications on mainstream GPUs (such as NVIDIA T4, RTX 3060/4060) without relying on expensive server infrastructure like the A100. Despite its modest parameter size, the SmolVLM2 demonstrates unexpectedly deep video understanding capabilities. The power of the SigLIP encoder enables accurate object recognition and OCR in natural environments, demonstrated by the successful decoding of proper names (“*DAI HOC*”, “*TAN TRAO*”) and detailed timestamps against an event backdrop. Its ability to capture subtle visual attributes (such as clothing color) provides a solid foundation for location identification and automated metadata extraction applications. In terms of semantic thinking, the model goes beyond surface recognition to understand diverse contexts (from drone shots to graduation ceremonies) and character actions. Importantly, the output ensures temporal consistency: instead of generating disjointed annotations, the model structures the content into a coherent story. The flexible use of transitional phrases (“*The video opens with...*”, “*Next...*”) demonstrates the ability to string together fragmented visual events into a complete and logical narrative. Besides their advantages,

⁵<https://www.youtube.com/watch?v=U-WK3ISjKA>

experimental results also reveal inherent limitations of small language models (SLMs), typically the problem of hallucination when processing unstructured text. Failed OCR attempts resulting in the appearance of meaningless character strings (e.g., “*Lien Voi Nha Khoai...*”) often stem from the model attempting to decode Vietnamese text with stylized, blurry, or abbreviated fonts without a clear basis for recognition. The core issue lies in the bias of the foundational training data (primarily English), causing the model to tend to generate content based on phonetic similarity rather than actual semantics. This highlights a significant barrier to local language sensitivity and highlights the urgent need for additional fine-tuning on high-quality Vietnamese annotated video datasets to completely overcome this problem.

Table 10: The results of the SmolVLM2 model are described in English in the video “*Trường Đại học Tân Trào*”.

Input	English Description
The video in details	Assistant: The video opens with an aerial view of a large building complex, prominently featuring the ‘DAI HOC’ and ‘TAN TRAO’ sign on its facade, surrounded by lush greenery under clear blue skies. It transitions to scenes showcasing various groups of people gathered at tables or desks at different locations within the same building complex. A group of women holding bouquets walk along garden paths, interacting with each other. A camera, followed by another group celebrating inside what appears to be a conference room filled with laptops. The scene shifts again to a classroom where students are working diligently on their computers. Next, individuals dressed in matching red and navy blue uniforms stand together outside a glass structure labeled ‘UNIVERSITY’, gesturing and giving instructions. Inside a lecture hall, several individuals work on laptops while others prepare for exams. In an outdoor setting near blooming flowers, students gather around a table covered with floral arrangements, engaged in discussions and possibly preparing presentations. Two groups walk through a green corridor lined with trees and buildings, passing banner reading ‘Lien Voi Nha Khoa Ha Lut Trang Cau Phuc’. Students present papers during a seminar event set up indoors against a backdrop that includes banners displaying Thai script and logos representing various organizations. In another context within the same building, a group of flags waves in the background, and some members perform a dance routine, masks hand-held along a path adorned with colorful flower petals. Another group stands tall in front of a yellow building with red columns, waving flags and appearing excited. Inside an indoor cafeteria, students engage in lively discussion, accompanied by food offerings like bowls and plates arranged neatly on trays. A graduation ceremony unfolds behind a podium with text indicating it’s Tuesday August 9th, marked by a ‘Happy Graduation!’ message displayed above the stage. Students celebrate individually before parting ways into small lines standing side by side, one line carrying a flag bearing national symbols such as stars and hearts, symbolizing unity and pride.

In summary, the SmolVLM2 strikes an optimal balance between speed and comprehension. The model demonstrates superior performance in content summarization and entity recognition tasks, making it perfectly suited for deployment on edge devices. However, for tasks requiring absolute accuracy in OCR in resource-limited languages, like Vietnamese, the system still requires the support of human-in-the-loop monitoring or specialized post-processing steps to ensure the reliability of the output data.

The convergence of audio, visual, and linguistic signals marks a significant shift, pushing VLM systems beyond the limitations of “*silent vision*” toward a unified cognitive architecture. Instead of discrete processing, advanced models like ImageBind [115], AudioGPT [115] synthesize information from all three data streams. This allows the system to analyze acoustic features to understand context, recognize emotions through tone, and detect events beyond visual observation.

However, the differences in the data between the models pose significant technical challenges. To address this, the research community is focusing on developing specialized audio encoders and multimodal attention mechanisms. The core objective is to precisely align the embedded space of audio data (which is a time series) in perfect synchronization with the semantic space of images and text.

In terms of hardware, the biggest challenge comes from running the wav2vec2-large-xlsr-53 model [108] (300 million parameters). The minimum VRAM required to start is 12–16 GB (equivalent to Tesla T4, RTX 3060); however, users must reduce the batch size and apply gradient accumulation techniques to ensure convergence. To optimize the training process, the recommended configuration is GPUs with 24 GB VRAM or more (such as RTX 3090, A10G). In addition, system memory (RAM) plays a crucial role when processing large float32 data arrays, with a minimum requirement of 16 GB (ideally 32 GB+). Under limited resources, a technical solution is to enable `streaming = True` mode for the datasets. Finally, equipping an SSD is mandatory to ensure fast read/write (I/O) speeds, avoiding bottlenecks that cause the GPU to idle.

To verify the model’s performance in a real-world environment, the study conducted experiments using audio data extracted from the video “*Trường Đại học Tân Trào*”⁶. This sample exhibited characteristics of a natural environment with background noise, moderate speaking speed, and standard Northern Vietnamese accent (news/document format). The main challenge for the model was its ability to accurately identify specialized educational terminology and local proper names (such as “*Tuyên Quang*”, “*Tân Trào*”). The decoding process using Greedy Search and Beam Search algorithms successfully reproduced approximately 85%–90% of the core content, particularly excelling in capturing long-term dependencies in complex sentences. However, the lack of a supporting Language Model revealed significant limitations in its operation. Common errors include spelling mistakes based on phonetic similarities (such as “*vững chắc*”, “*y dựng*”) and misidentification of proper nouns outside the training dataset’s coverage area (such as “*Tuyên Quang*” as “*kiên quang*”). Notably, expanding the search space using Beam Search did not yield a significant improvement over Greedy Search, confirming that relying solely on phonetic probability is insufficient to correct semantic errors without incorporating external linguistic knowledge.

5.3 VLM’s Development Direction

- 1) **Extended Model to Video:** Extending VLM to the video domain is a crucial step, requiring solutions to the spatio-temporal representation problem to capture dynamic content and causal relationships. Advanced architectures such as Video-LLaMA [114] and VideoChat [116] integrate temporal attention mechanisms or adapters to connect visual encoders to LLMs (the architecture is illustrated in Fig. 13), serving tasks such as video captioning, VideoQA, and action recognition. In terms of evaluation, the Video-MME [73] benchmark plays a core role in testing the model’s ability to perform complex reasoning and synthesize information in a dynamic environment.

⁶<https://www.youtube.com/watch?v=U-WK3ISjjKA>

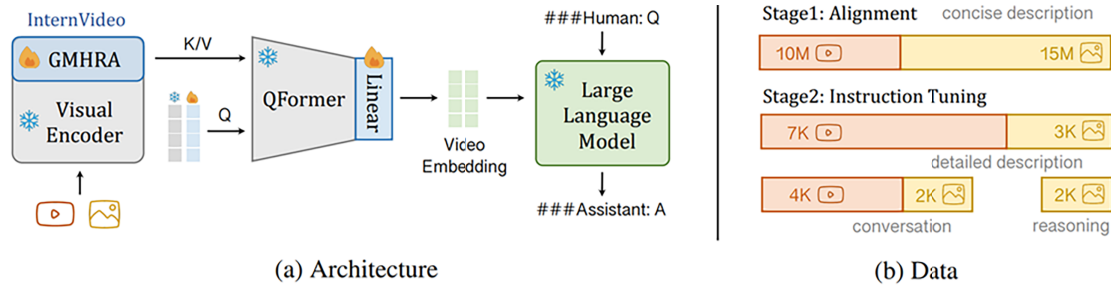


Figure 13: Architecture and training process of VideoChat [116].

- 2) **The SmolVLM2 Model and Video Comprehension Capabilities on Edge Devices:** In the context of the Generative AI boom, alongside the development of LMMs requiring massive infrastructure, the trend of optimizing compression architectures for edge devices is becoming a crucial counterweight. The SmolVLM2 [109] model on Hugging Face is a prime example of the strategy of moving AI out of the cloud to operate directly on user devices. With its ultra-compact size of just 500M, this model challenges the traditional notion that complex video comprehension tasks require massive models (7B+). Its impressive performance is the result of knowledge distillation and neural network optimization techniques, maintaining high accuracy in video tutorials and question-answering tasks while ensuring low latency. Technically, the SmolVLM2 architecture is built on the SmolLM2 backbone (135M/360M ultra-lightweight variants) rather than the latest SmolLM3 generation. This is a strategic engineering decision: prioritizing processing speed and performance on low-VRAM devices (such as phones and laptops), accepting a trade-off in language knowledge depth to achieve real-time responsiveness.
- 3) **The Role of NPL/LLM in VLM and the Output of NPL:** In modern multimodal architecture, the roles of NLP and LLM have shifted to a central (“LLM-*first*”) position, acting as the primary decoder and reasoning tool. Typical models such as MiniGPT-4 [10] and BLIP-2 [23] utilize linear projection layers or Q-Former [28] networks to align image embedding spaces with LLM, thereby transforming all visual tasks into auto-regressive text generation problems. This integration enables multimodal few-shot learning and allows models like Flamingo [83] or GPT-4V [5] to leverage outside knowledge to solve complex inferences. However, to ensure reliability, instruction-tuning and RLHF techniques play a crucial role in aligning the output with user intent and minimizing hallucination. Based on experimental results from sample videos, it can be affirmed that the current model has reached maturity at the Acoustic Model level with excellent phoneme recognition capabilities, but still reveals significant limitations at the Language Model level. The lack of semantic knowledge leads to contextual spelling errors (e.g., confusing “y được” with “y dựng”) and significant inaccuracies in handling proper names or local vocabulary. To overcome these challenges and move towards commercial application, the primary solution is to integrate an External LM. Combining decoding algorithms with n-gram LMs (such as KenLM) or Transformer LMs will act as post-processing tools, using statistical probabilities to automatically correct semantic errors (e.g., restoring “Tuyen Quang” from “kienh quang”), promising to raise overall accuracy to over 95%. Simultaneously, a data augmentation strategy with diverse dialects and background noise is also crucial for increasing the robustness of the system.

6 Conclusions and Future Work

The development of VLM models in recent years has marked a pivotal architectural shift, from contrastive learning methods to the era of generative AI architectures centered on LLMs. We have presented a survey of VLM, classifying its development directions according to model architecture and learning methods,

databases, evaluation metrics, and overall results. At the same time, we have also presented the challenges posed by VLM. Within this technological trend, this report demonstrates the feasibility and effectiveness of small language models specifically optimized for Vietnamese. We collected and published the VQA-TQU1 dataset with 180 color images and 5 question-answer pairs, each with a detailed description of approximately 100 words and 1080 interactions. Simultaneously, we fine-tuned Viterbin-1B on the VQA-TQU1 dataset, with the BLEU measure being 0.5677 and the ROUGE-L measure being 0.7397. A model that has proven superior in OCR and document information extraction tasks, establishing an optimal balance between accuracy and operating costs on edge devices. Simultaneously, the successful integration of an LLM as a post-processor for the ASR Wav2Vec 2.0 system has thoroughly addressed the semantic and local spelling issues, creating a highly reliable speech-to-text conversion process. We performed ASR assessments on the SmolVLM2 and Wav2vec2 models, with results showing $RTF = 0.24$ and $RTF = 0.36$, respectively.

Despite impressive achievements, the practical testing process still reveals core technical hurdles that need to be overcome. The biggest challenge to system reliability is “*hallucination*,” manifested by the model sometimes generating inaccurate information about dates or language mixing when processing complex unstructured texts. Furthermore, the lack of semantic knowledge in purely acoustic models is a significant limitation, making the system prone to errors in identifying homonyms or rare proper names without the intervention of auxiliary language models. In addition, the ability to ground and count objects in small-scale models still lags behind that of large-scale commercial models.

Looking ahead, the development roadmap for this field will focus on creating comprehensive Perceptual AI systems. The primary approach is Holistic Integration, directly integrating speech recognition modules into the VLM architecture to build a unified Video-LLM system capable of simultaneously processing audio and visual signals, rather than operating as separate pipelines. Simultaneously, expanding and standardizing specialized datasets (such as TQU-VQA) is crucial to minimizing noise and enhancing model robustness. Further research will move towards Embodied AI, integrating streamlined VLM models into robotic agents or 3D virtual assistants, enabling them not only to observe but also to plan and interact physically in the real world. In the near future, we will be adding data to the VQA-TQU1 dataset and fine-tuning the models for better accuracy.

Acknowledgement: This research is supported by Hung Vuong University.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Van-Thuan Nguyen, Van-Nui Nguyen, Van-Hung Le; data collection: Van-Thuan Nguyen, Van-Hung Le; analysis and interpretation of results: Van-Nui Nguyen; draft manuscript preparation: Van-Thuan Nguyen, Van-Hung Le; model optimization: Van-Hung Le; visual validation and perceptual metric analysis: Van-Thuan Nguyen; academic norm checking: Van-Hung Le. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Van-Hung Le, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(8):5625–44. doi:10.1109/TPAMI.2024.3369699.

2. Sun S, Chen W, Wang L, Liu X, Liu TY. On the depth of deep neural networks: a theoretical view. *Proc AAAI Conf Artif Intell.* 2016;30(1):2066–72. doi:10.1609/aaai.v30i1.10243.
3. Du Y, Liu Z, Li J, Zhao WX. A survey of vision-language pre-trained models. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*; 2022 Jul 23–29; Vienna, Austria. p. 5436–43. doi:10.24963/ijcai.2022/762.
4. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. *arXiv:2103.00020.* 2021.
5. Wu T, Yang G, Li Z, Zhang K, Liu Z, Guibas L, et al. GPT-4V(ision) is a human-aligned evaluator for text-to-3D generation. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2024 Jun 16–22; Seattle, WA, USA. p. 22227–38. doi:10.1109/CVPR52733.2024.02098.
6. Li Z, Wu X, Du H, Liu F, Nghiem H, Shi G. A survey of state of the art large vision language models: alignment, benchmark, evaluations and challenges. In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; 2025 Jun 11–12; Nashville, TN, USA. p. 1578–97. doi:10.1109/CVPRW67362.2025.00147.
7. Jiang Z, Meng R, Yang X, Yavuz S, Zhou Y, Chen W. VLM2Vec: training vision-language models for massive multimodal embedding tasks. *arXiv:2410.05160.* 2024.
8. Kim S, Jo D, Lee D, Kim J. MAGVLT: masked generative vision-and-language transformer. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023 Jun 17–24; Vancouver, BC, Canada. p. 23338–48. doi:10.1109/CVPR52729.2023.02235.
9. Jiang H, Zhang J, Huang R, Ge C, Ni Z, Song S, et al. Cross-modal adapter for vision-language retrieval. *Pattern Recognit.* 2025;159:111144. doi:10.1016/j.patcog.2024.111144.
10. Zhu D, Chen J, Shen X, Li X, Elhoseiny M. Minigpt-4: enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592.* 2024.
11. Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choromanski K, et al. RT-2: vision-language-action models transfer web knowledge to robotic control. In: *Proceedings of the 7th Machine Learning Research*; 2023 Nov 6–9; Atlanta, GA, USA. London, UK: PMLR; 2023. p. 2165–83.
12. Zhou X, Liu M, Yurtsever E, Zagar BL, Zimmer W, Cao H, et al. Vision language models in autonomous driving: a survey and outlook. *IEEE Trans Intell Veh.* 2025. doi:10.1109/tiv.2024.3402136.
13. Chen FL, Zhang DZ, Han ML, Chen XY, Shi J, Xu S, et al. VLP: a survey on vision-language pre-training. *Mach Intell Res.* 2023;20(1):38–56. doi:10.1007/s11633-022-1369-5.
14. Danish S, Sadeghi-Niaraki A, Khan SU, Dang LM, Tightiz L, Moon H. A comprehensive survey of vision-language models: pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Inf Fusion.* 2026;126(Pt B):103623. doi:10.1016/j.inffus.2025.103623.
15. Pham MT, Pham QH, Tran QD, Ho HT, Nguyen LV, Huy DNM, et al. A review on vision-language-based approaches: challenges and applications. *Comput Mater Continua.* 2025;82(2):1733–56. doi:10.32604/cmc.2025.060363.
16. Masry A, Rodriguez JA, Zhang T, Wang S, Wang C, Feizi A, et al. A LIGN VLM: bridging vision and language latent spaces for multimodal understanding. *arXiv:2502.01341.* 2025.
17. Tschannen M, Gritsenko A, Wang X, Naeem MF, Alabdulmohsin I. SigLIP 2: multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786.* 2025.
18. Gustafsson F, Danelljan M, Timofte R, Schön T. How to train your energy-based model for regression. *arXiv:2005.01698.* 2020. doi:10.5244/c.34.43.
19. Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, et al. FLAVA: a foundational language and vision alignment model. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022 Jun 18–24; New Orleans, LA, USA. p. 15617–29. doi:10.1109/CVPR52688.2022.01519.
20. Kwon G, Cai Z, Ravichandran A, Bas E, Bhotika R, Soatto S. Masked vision and language modeling for multimodal representation learning. *arXiv:2208.02131.* 2023.

21. Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, et al. Reproducible scaling laws for contrastive language-image learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 2818–29. doi:10.1109/CVPR52729.2023.00276.
22. Tsimpoukelli M, Menick J, Cabi S, Eslami SMA, Vinyals O, Hill F. Multimodal few-shot learning with frozen language models. In: NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2021. p. 200–12.
23. Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML'23: Proceedings of the 40th International Conference on Machine Learning. London, UK: PMLR; 2023. p. 19730–42.
24. Zheng K, He X, Wang XE. Interleaved vision-and-language generation via generative voken. In: 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2026 Mar 6–10; Tucson, AZ, USA. p. 472–82. doi:10.1109/WACV61042.2026.00054.
25. Chen J, Zhu D, Shen X, Li X, Liu Z, Zhang P, et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. In: The Twelfth International Conference on Learning Representations, ICLR 2024; 2024 May 7–11; Vienna, Austria.
26. Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, et al. Qwen-vl: a versatile vision-language model for understanding, localization, text reading, and beyond. In: The Twelfth International Conference on Learning Representations, ICLR 2024; 2024 May 7–11; Vienna, Austria.
27. Kim S, Lee A, Park J, Chung A, Oh J, Lee J-Y. Towards efficient visual-language alignment of the q-former for visual reasoning tasks. In: Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA, USA: ACL; 2024. p. 15155–65.
28. Avogaro N, Frick T, Bartezzaghi A, Malossi C, Schindler K, Assaf R, et al. Show or Tell? Effectively prompting vision language models for semantic segmentation. In: Transactions on Machine Learning Research. The Thirteenth International Conference on Learning Representations; 2025 Apr 24; Singapore.
29. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y. CoCa: contrastive captioners are image-text foundation models. arXiv:2205.01917. 2022.
30. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer Vision—ECCV 2014. Cham, Switzerland: Springer International Publishing; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.
31. Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W. Visualbert: a simple and performant baseline for vision and language. arXiv:1908.03557. 2019.
32. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, et al. VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 2425–243. doi:10.1109/iccv.2015.279.
33. Zhang T, Wang S, Li L, Zhang G, Taslakian P, Rajeswar S, et al. VCR: a task for pixel-level complex reasoning in vision language models via restoring occluded text. arXiv:2406.06462. 2025.
34. Suhr A, Artzi Y. NLVR2 visual bias analysis. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Red Hook, NY, USA: Curran Associates, Inc.; 2019. p. 6418–28.
35. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 2641–9. doi:10.1109/ICCV.2015.303.
36. Lu J, Batra D, Parikh D, Lee S. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and language tasks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2019. p. 13–23.
37. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86.
38. Changpinyo S, Sharma P, Ding N, Soricut R. Conceptual 12M: pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 3557–67. doi:10.1109/CVPR46437.2021.00356.

39. Karamcheti S, Nair S, Balakrishna A, Liang P, Kollar T, Sadigh D. Prismatic VLMS: investigating the design space of visually-conditioned language models. In: ICML'24: Proceedings of the 41st International Conference on Machine Learning. London, UK: PMLR; 2024. p. 23123–2314.
40. Cho J, Zala A, Bansal M. Visual programming for text-to-image generation and evaluation. In: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2023. p. 6048–69.
41. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 26286–96. doi:10.1109/CVPR52733.2024.02484.
42. Panagopoulou A, Xue L, Yu N, Li J, Li D, Joty S, et al. X-InstructBLIP: a framework for aligning image, 3D, audio, video to LLMs and its emergent cross-modal reasoning. In: Computer Vision—ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 177–97. doi:10.1007/978-3-031-72995-9_11.
43. Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model. In: ICML'23: Proceedings of the 40th International Conference on Machine Learning. Cambridge, MA, USA: PMLR; 2022. p. 8469–88.
44. Wiecek TJ, Daun N, Khan ME, Rohrbach M. Variational visual question answering for uncertainty-aware selective prediction. arXiv:2505.09591. 2026.
45. Marino K, Rastegari M, Farhadi A, Mottaghi R. OK-VQA: a visual question answering benchmark requiring external knowledge. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 3190–9. doi:10.1109/CVPR.2019.00331.
46. Zhang Y, Wu JH, Chen Z, Huang H, Zhu G. Text-based visual question answering based on text-aware pre-training. In: 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML); 2024 Nov 22–24; Shenzhen, China. p. 1017–24. doi:10.1109/ICICML63543.2024.10957919.
47. Li Y, Du Y, Zhou K, Wang J, Zhao X, Wen JR. Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore Association for Computational Linguistics. Kerrville, TX, USA: ACL; 2023. p. 292–305.
48. Lu P, Mishra S, Xia T, Qiu L, Chang K-W, Zhu S-C, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. arXiv:2209.09513. 2022.
49. Liu Y, Duan H, Zhang Y, Li B, Zhang S, Zhao W. MMBench: is your multi-modal model an all-around player? In: Computer Vision—ECCV 2024: 18th European Conference. Cham, Switzerland: Springer; 2024. p. 216–33.
50. Agrawal H, Desai K, Wang Y, Chen X, Jain R, Johnson M, et al. Nocaps: novel object captioning at scale. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 8947–56. doi:10.1109/iccv.2019.00904.
51. Mathew M, Karatzas D, Jawahar CV. DocVQA: a dataset for VQA on document images. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 3–8; Waikoloa, HI, USA. p. 2199–208. doi:10.1109/wacv48630.2021.00225.
52. Masry A, Long D, Tan JQ, Joty S, Hoque E. ChartQA: a benchmark for question answering about charts with visual and logical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022. Kerrville, TX, USA: ACL; 2022. p. 2263–79.
53. Mathew M, Bagal V, Tito R, Karatzas D, Valveny E, Jawahar CV. InfographicVQA. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3–8; Waikoloa, HI, USA. p. 2582–91. doi:10.1109/WACV51458.2022.00264.
54. Thrush T, Jiang R, Bartolo M, Singh A, Williams A, Kiela D, et al. Winoground: probing vision and language models for visio-linguistic compositionality. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5228–38. doi:10.1109/CVPR52688.2022.00517.
55. Wu F, Pang Y, Zhang J, Pang L, Yin J. CoRe: context-regularized text embedding learning for text-to-image personalization. Proc AAAI Conf Artif Intell. 2025;3(8):8377–85.

56. Cheng K, Song W, Fan J, Ma Z, Sun Q, Xu F, et al. CapArena: benchmarking and analyzing detailed image captioning in the LLM era. In: Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg, PA, USA: ACL; 2025. p. 14077–94. doi:10.18653/v1/2025.findings-acl.724.
57. Wang U, Duan H, Wang J, Jia Z, Zhai G, Min X. TIT-score: evaluating long-prompt based text to-image alignment via text-to-image-to-text consistency. arXiv:2510.02987. 2025.
58. Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y. CLIPScore: a reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2021. p. 7514–28. doi:10.18653/v1/2021.emnlp-main.595.
59. Hu Y, Liu B, Kasai J, Wang Y, Ostendorf M, Krishna R, et al. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. 2024. p. 20349–60. doi:10.1109/ICCV51070.2023.01866.
60. Cho J, Hu Y, Garg R, Anderson P, Krishna R, Baldrige J, et al. Davidsonian scene graph: improving reliability in fine-grained evaluation for text-to-image generation. In: The Twelfth International Conference on Learning Representations, ICLR 2024; 2024 May 11; Vienna, Austria.
61. Jiu S, Weng Y, Zhu M, Feng C, Liu Z. TVQACML: benchmarking text-centric visual question answering in multilingual Chinese minority languages. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics; 2025. p. 13957–67.
62. Vu YL, Duong DT, Duong TB, Nguyen AK, Nguyen TH, Phuc Nguyen LT, et al. Describe anything model for visual question answering on text-rich images. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2025 Oct 19–20; Honolulu, HI, USA. p. 7534–44. doi:10.1109/ICCVW69036.2025.00776.
63. Tang J, Liu Q, Ye Y, Lu J, Wei S, Wang AL, et al. MTVQA: benchmarking multilingual text-centric visual question answering. In: Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg, PA, USA: ACL; 2025. p. 7748–63. doi:10.18653/v1/2025.findings-acl.404.
64. Yang Y, Pan L, Li D, Liu L. EZSR: event-based zero-shot recognition. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 4628–38. doi:10.1109/CVPR52734.2025.00436.
65. Xia H, Dong Q. ImageNetVC: zero- and few-shot visual commonsense evaluation on 1000 ImageNet categories. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Kerrville, TX, USA: ACL. p. 2009–26.
66. Mishra S, Saenko K, Saligrama V. SCRAMBLE: enhancing multimodal LLM compositionality with synthetic preference data. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2025 Oct 19–20; Honolulu, HI, USA. p. 6292–302. doi:10.1109/ICCVW69036.2025.00654.
67. Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JMF, et al. Visual dialog. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1080–9. doi:10.1109/CVPR.2017.121.
68. Goto K, Hirose T, Ukai M, Kurita S, Inoue N. Referring expression comprehension for small objects. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV); 2025 Oct 19–25; Honolulu, HI, USA. p. 21231–42. doi:10.1109/ICCV51701.2025.01972.
69. Kazemzadeh S, Ordonez V, Matten M, Berg T. ReferItGame: referring to objects in photographs of natural scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: ACL; 2014. p. 787–98. doi:10.3115/v1/d14-1086.
70. Fujitake M. LayoutLLM: large language model instruction tuning for visually rich document understanding. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 10219–24. doi:10.63317/428wjtmc3xvh.
71. Biten AF, Tito R, Mafla A, Gomez L, Rusinol M, Jawahar CV, et al. Scene text visual question answering. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 4290–300. doi:10.1109/iccv.2019.00439.
72. Yue X, Ni Y, Zheng T, Zhang K, Liu R, Zhang G, et al. MMMU: a massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 9556–67. doi:10.1109/cvpr52733.2024.00913.

73. Shu Y, Liu Z, Zhang P, Qin M, Zhou J, Liang Z, et al. Video-XL: extra-long vision language model for hour-scale video understanding. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 26160–9. doi:10.1109/CVPR52734.2025.02436.
74. Fu C, Dai Y, Luo Y, Li L, Ren S, Zhang R, et al. Video-MME: the first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2025. p. 24108–18.
75. Raj C, Wei B, Caliskan A, Anastasopoulos A, Zhu Z. VIGNETTE: socially grounded bias evaluation for vision language models. arXiv:2505.22897. 2026.
76. Wang S, Cao X, Zhang J, Yuan Z, Shan S, Chen X, et al. VLBiasBench: a comprehensive benchmark for evaluating bias in large vision-language model. IEEE Trans Pattern Anal Mach Intell. 2026;48(8):10143–56 doi:10.1109/tpami.2026.3683747.
77. Sahili ZA, Patras I, Purver M. Data matters most: auditing social bias in contrastive vision language models. Trans Mach Learn Res. 2025. [cited 2026 May 13]. Available from: <https://openreview.net/forum?id=3vF2fn9owm>.
78. Hao Y, Tan Y, Wang S, Zhang H, Zhu B, Zhu X. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. In: Advances in Neural Information Processing Systems; 2024 Dec 10–15; Vancouver, BC, Canada. p. 2001–25. doi:10.52202/079017-0064.
79. Todescato MV, Carbonera JL. No labels needed: zero-shot image classification with collaborative self-learning. In: 2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI); 2025 Nov 3–5; Athens, Greece. p. 972–6. doi:10.1109/ICTAI66417.2025.00142.
80. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv:2201.12086. 2022.
81. Alayrac J-B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. In: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. San Diego, CA, USA: NIPS; 2022. p. 23716–36.
82. Maller S, Aster F, Tronger S, Beyer L, Padlewski P, Salz D, et al. PALI-3 vision language models: smaller, faster, stronger. In: The Twelfth International Conference on Learning Representations, ICLR 2024; 2024 May 7–11; Vienna, Austria. p. 1–16.
83. Hiippala T, Alikhani M, Haverinen J, Kalliokoski T, Logacheva E, Orekhova S, et al. AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. Lang Resour Eval. 2021;55(3):661–88. doi:10.1007/s10579-020-09517-1.
84. Ma F, Zhou Y, Rao F, Zhang Y, Sun X. Image captioning with multi-context synthetic data. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2024. p. 4089–97. doi:10.1609/aaai.v38i5.28203.
85. Popp E, Hlavacs H, Winiwarter W. AI-driven web game development with Gemini 2.5 Pro. In: Information Integration and Web Intelligence: 27th International Conference, iiWAS 2025; 2025 Dec 8–10; Matsue, Japan. p. 474–9. doi:10.1007/978-3-032-11976-6_36.
86. Masry A, Islam MS, Ahmed M, Bajaj A, Kabir F, Kartha A, et al. ChartQAPro: a more diverse and challenging benchmark for chart question answering. In: Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg, PA, USA: ACL; 2025. p. 19123–51. doi:10.18653/v1/2025.findings-acl.978.
87. Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of The 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual.
88. Lavoie S, Kirichenko P, Ibrahim M, Assran M, Wilson AG, Courville A, et al. Modeling caption diversity in contrastive vision-language pretraining. In: ICML'24: Proceedings of the 41st International Conference on Machine Learning. Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 26070–84.
89. Beaumont R, Cherti M, Coombes T, Crowson K, Gordon C, Jitsev J, et al. LAION-5B: an open large-scale dataset for training next generation image-text models. In: Advances in Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 25278–94. doi:10.52202/06431-1833.

90. Bezemer CP, Habchi S, Nguyen A, Nguyen G, Taesiri MR. ImageNet-hard: the hardest images remaining from a study of the power of zoom and spatial biases in image classification. In: *Advances in Neural Information Processing Systems*; 2023 Dec 10–16; New Orleans, LA, USA. p. 35878–953. doi:10.52202/075280-1558.
91. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis*. 2017;123(1):32–73. doi:10.1007/s11263-016-0981-7.
92. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high performance deep learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. San Diego, CA, USA: NIPS; 2019. p. 8026–37.
93. Zhang J, Zhang Y, Gu J, Dong J, Kong L, Yang X. Xformer: hybrid x-shaped transformer for image denoising. arXiv:2303.06440. 2024.
94. Leclerc G, Ilyas A, Engstrom L, Park SM, Salman H, Mądry A. FFCV: accelerating training by removing data bottlenecks. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023 Jun 17–24; Vancouver, BC, Canada. p. 12011–20. doi:10.1109/CVPR52729.2023.01156.
95. Zeng Y, Zhang X, Li H. Multi-grained vision language pre-training: aligning texts with visual concepts. arXiv:2111.08276. 2022.
96. Vicente TFY, Hou L, Yu CP, Hoai M, Samaras D. Large-scale training of shadow detectors with noisily-annotated shadow examples. In: *Computer Vision—ECCV 2016*. Cham, Switzerland: Springer International Publishing; 2016. p. 816–32. doi:10.1007/978-3-319-46466-4_49.
97. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl*. 2023;82(3):3713–44. doi:10.1007/s11042-022-13428-4.
98. Dai W, Fung PN, Hoi S, Li B, Li J, Li D, et al. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In: *Advances in Neural Information Processing Systems*; 2023 Dec 10–16; New Orleans, LA, USA. p. 49250–67. doi:10.52202/075280-2142.
99. Sun Z, Shen S, Cao S, Liu H, Li C, Shen Y, et al. Aligning large multimodal models with factually augmented RLHF. In: *Findings of the Association for Computational Linguistics ACL 2024*. Bangkok, Thailand and Virtual Meeting. Stroudsburg, PA, USA: ACL; 2024. p. 13088–110. doi:10.18653/v1/2024.findings-acl.775.
100. Zhang Y, Zhang R, Gu J, Zhou Y, Lipka N, Yang D, et al. Lllavar: enhanced visual instruction tuning for text-rich image understanding. In: *The Twelfth International Conference on Learning Representations, ICLR 2024*; 2024 May 7–11; Vienna, Austria.
101. Li Z, Yang B, Liu Q, Ma Z, Zhang S, Yang J, et al. Monkey: image resolution and text label are important things for large multi-modal models. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2024 Jun 16–22; Seattle, WA, USA. p. 26753–63. doi:10.1109/CVPR52733.2024.02527.
102. Nematollahi I, DeMoss B, Chandra AL, Hawes N, Burgard W, Posner I. Lumos: language-conditioned imitation learning with world models. arXiv:2503.10370. 2025.
103. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: low-rank adaptation of large language models. arXiv:2106.09685. 2021.
104. Dettmers T, Holtzman A, Pagnoni A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. In: *Advances in Neural Information Processing Systems*; 2023 Dec 10–16; New Orleans, LA, USA. p. 10088–115. doi:10.52202/075280-0441.
105. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. *Int J Comput Vis*. 2022;130(9):2337–48. doi:10.1007/s11263-022-01653-1.
106. Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, et al. CLIP-adapter: better vision-language models with feature adapters. *Int J Comput Vis*. 2024;132(2):581–95. doi:10.1007/s11263-023-01891-x.
107. Lupaşcu M, Rogoz AC, Sorin Stupariu M, Tudor Ionescu R. Large multimodal models for low-resource languages: a survey. *Inf Fusion*. 2026;131(13):104189. doi:10.1016/j.inffus.2026.104189.
108. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. In: *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: NIPS; 2020. p. 12449–60.

109. Marafoti A, Zohar O, Farré M, Noyan M, Bakouch E, Cuenca P, et al. Smolvlm: redefining small and efficient multimodal models. arXiv:2504.05299. 2025.
110. Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, et al. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 24185–98. doi:10.1109/CVPR52733.2024.02283.
111. Chen Z, Wang W, Tian H, Ye S, Gao Z, Cui E, et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Sci China Inf Sci.* 2024;67(12):220101. doi:10.1007/s11432-024-4231-5.
112. Nagayi M, Khan A, Frank T, Swart R, Nyirenda C. Evaluating OCR performance on food packaging labels in South Africa. In: *Artificial Intelligence Research (SACAIR 2025)*. Cham, Switzerland: Springer; 2025. p. 127–43. doi:10.1007/978-3-032-11733-5_8.
113. Rahimi N. POVQA: preference-optimized video question answering with rationales for data efficiency. arXiv:2510.01009. 2026.
114. Zhang H, Li X, Bing L. Video-LLaMA: an instruction-tuned audio-visual language model for video understanding. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Singapore. Stroudsburg, PA, USA: ACL; 2023. p. 543–53. doi:10.18653/v1/2023.emnlp-demo.49.
115. Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, et al. ImageBind one embedding space to bind them all. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 15180–90. doi:10.1109/CVPR52729.2023.01457.
116. Li K, He Y, Wang Y, Li Y, Wang W, Luo P, et al. VideoChat: chat-centric video understanding. *Sci China Inf Sci.* 2025;68(10):200102. doi:10.1007/s11432-024-4321-9.