Qianfan-VL: Domain-Enhanced Universal Vision-Language Models

Qianfan Team, Baidu AI Cloud

https://github.com/baidubce/Qianfan-VL

Abstract

We present Qianfan-VL, a series of multimodal large language models ranging from 3B to 70B parameters, achieving state-of-the-art performance through innovative domain enhancement techniques. Our approach employs multi-stage progressive training and high-precision data synthesis pipelines, which prove to be critical technologies for enhancing domain-specific capabilities while maintaining strong general performance. Qianfan-VL achieves comparable results to leading open-source models on general benchmarks, with state-of-the-art performance on benchmarks such as CCBench, SEEDBench IMG, ScienceQA, and MMStar. The domain enhancement strategy delivers significant advantages in OCR and document understanding, validated on both public benchmarks (OCRBench 873, DocVQA 94.75%) and in-house evaluations. Notably, Qianfan-VL-8B and 70B variants incorporate long chain-of-thought capabilities, demonstrating superior performance on mathematical reasoning (MathVista 78.6%) and logical inference tasks. All models are trained entirely on Baidu's Kunlun P800 chips, validating the capability of large-scale AI infrastructure to train SOTA-level multimodal models with over 90% scaling efficiency on 5000 chips for a single task. This work establishes an effective methodology for developing domain-enhanced multimodal models suitable for diverse enterprise deployment scenarios.

1 Introduction

The rapid advancement of vision-language models (VLMs) has enabled remarkable progress in multimodal understanding (Radford et al., 2021; Liu et al., 2024a; Alayrac et al., 2022). However, enterprise applications often require not only general multimodal capabilities but also domain-specific expertise in critical areas such as document processing, OCR recognition, and mathematical reasoning (Mathew et al., 2021; Masry et al., 2022). Existing VLMs typically face a trade-off between maintaining broad general capabilities and achieving deep domain expertise (Chen et al., 2024d; Bai et al., 2023), despite recent advances in multimodal instruction tuning (Liu et al., 2024a) and document understanding (Hu et al., 2024).

We introduce Qianfan-VL, a family of domain-enhanced vision-language models that addresses this challenge through innovative training strategies and architectural designs. Our approach centers on three key contributions:

First, we propose a four-stage progressive training pipeline that systematically enhances domain capabilities while preserving general performance. This pipeline progresses from cross-modal alignment (100B tokens) through general knowledge injection (2.66T tokens) and domain enhancement (0.32T tokens) to final instruction tuning (1B tokens). The careful staging and data mixture design enables the model to acquire specialized skills without catastrophic forgetting of general knowledge.

Second, we develop comprehensive data synthesis pipelines for critical enterprise scenarios. By combining traditional computer vision models with programmatic generation techniques, we create high-quality training data at scale. Our synthesis covers six major task categories: document OCR, mathematical problem-solving, chart understanding, table recognition, formula recognition, and natural scene OCR. Each pipeline incorporates domain-specific augmentation strategies and quality verification mechanisms to ensure data reliability.

Third, we demonstrate the feasibility of training large-scale VLMs entirely on proprietary hardware infrastructure. All Qianfan-VL models are trained on Baidu's Kunlun P800 chips, utilizing innovative parallel strategies and communication-computation fusion techniques to achieve over 90% scaling efficiency on 5000+ chip clusters. This represents a significant milestone in developing independent AI capabilities.

Qianfan-VL offers three model variants to address different deployment scenarios:

Model	Context	CoT	Target Deployment
Qianfan-VL-3B	32K	✗	Edge devices, real-time OCR
Qianfan-VL-8B	32K	✓	Servers, general applications
Qianfan-VL-70B	32K	✓	Cloud, complex reasoning

Table 1: Qianfan-VL model variants and their capabilities.

Extensive evaluations demonstrate that Qianfan-VL achieves competitive performance on general multimodal benchmarks while excelling in domain-specific tasks. On document understanding benchmarks, Qianfan-VL-70B achieves 94.75% on DocVQA, demonstrating strong document processing capabilities. For mathematical reasoning, Qianfan-VL-70B reaches 78.60% on Mathvista, demonstrating strong problem-solving capabilities. The models also show significant improvements in OCR tasks, with scores of 873 on OCRBench (Chen et al., 2024c) for the 70B variant.

2 Model Architecture

Qianfan-VL adopts a modular architecture that combines proven components with targeted innovations for domain enhancement. The architecture consists of three main components: a language model backbone, a visual encoder, and a cross-modal adapter.

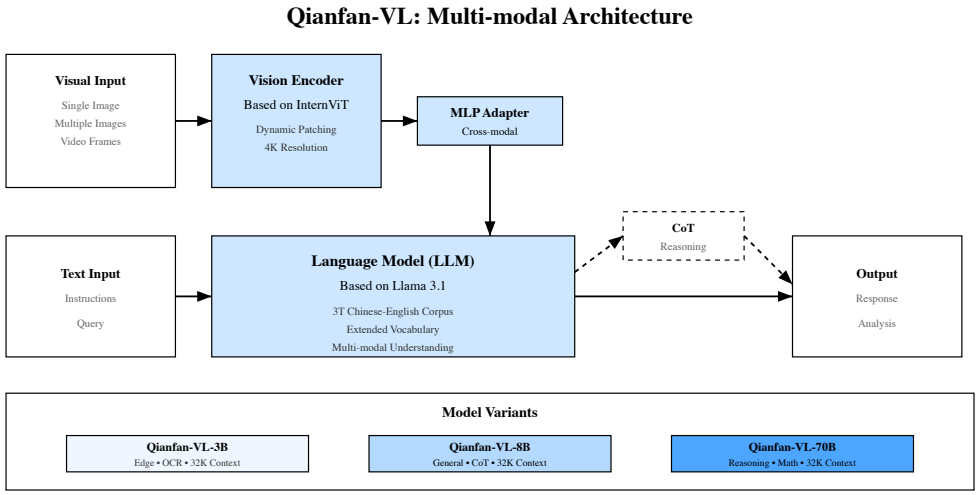


Figure 1: Overall architecture of Qianfan-VL. The model integrates the InternViT architecture for visual encoding and Llama 3.1 (8B/70B) or Qwen2.5-3B (3B) architectures for language modeling, initialized with their original pretrained weights. The cross-modal MLP adapter is randomly initialized. Based on this assembled multimodal architecture, we conduct multi-stage progressive pretraining to build the final Qianfan-VL models.

2.1 Language Model Backbone

The language model backbone varies across our model variants: the 8B and 70B models are based on the Llama 3.1 architecture (Touvron et al., 2023), while the 3B model is based on Qwen2.5-3B (Bai et al., 2025), both enhanced with vocabulary expansion and localization improvements. We extend the original vocabulary with additional tokens and train on 3T tokens of multilingual mixed data to improve cross-lingual understanding. The model employs Grouped-Query Attention (GQA) (Ainslie et al., 2023) to optimize memory efficiency and inference speed, while RMSNorm (Zhang and Sennrich, 2019) is used to improve training stability.

We provide three model variants with different parameter scales to serve diverse deployment scenarios. Table 2 presents the detailed architectural parameters for each variant.

Configuration	Qianfan-VL-3B	Qianfan-VL-8B	Qianfan-VL-70B
Vision Encoder			
Hidden Size	1024	1024	1024
# Layers	24	24	24
# Num Heads	16	16	16
Intermediate Size	4096	4096	4096
Patch Size	14	14	14
Cross-Modal Adapter			
In Channel	4096	4096	4096
Out Channel	2048	4096	8192
Language Model Backbone			
Hidden Size	2048	4096	8192
# Layers	36	32	80
# KV Heads	2	8	8
Head Size	128	128	128
Intermediate Size	11008	14336	28672
Embedding Tying	✓	×	×
Vocabulary Size	151673	182025	182025

Table 2: Architectural parameters for Qianfan-VL model variants. Note: The 3B variant uses Qwen2.5-3B as the language model backbone, while the 8B and 70B variants use Llama 3.1.

2.2 Vision Encoder

The vision encoder is initialized from InternViT (Chen et al., 2024d) and supports dynamic image tiling for variable resolution inputs, building upon the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2020). The encoder processes images at multiple scales, with maximum support for 4K resolution inputs. We employ a tile-based approach where the image is divided into dynamic number of 448×448 pixel tiles, plus an additional global snapshot where the entire image is resized to 448×448 to help the model capture holistic image information. The vision transformer configurations are as follows:

Model	Vision Params	Image Tokens/Tile	Max Tiles
Qianfan-VL-3B	300M	256	12
Qianfan-VL-8B	300M	256	12
Qianfan-VL-70B	300M	256	12

Table 3: Visual encoder configurations for Qianfan-VL models.

The dynamic tiling strategy allows the model to process high-resolution images by splitting them into multiple tiles, each processed independently and then aggregated. This approach maintains detail preservation while managing computational costs.

2.3 Cross-Modal Adapter

The cross-modal adapter employs a two-layer MLP with GELU activation to project visual features into the language model's embedding space. It starts with layer normalization on the input visual features, followed by dimensional reduction through the first linear layer, GELU activation for non-linearity, and a final linear transformation. This design ensures stable training dynamics and efficient cross-modal alignment between vision and language representations.

3 Training Methodology

3.1 Four-Stage Progressive Training

Our training methodology employs a carefully designed four-stage pipeline that progressively builds model capabilities:

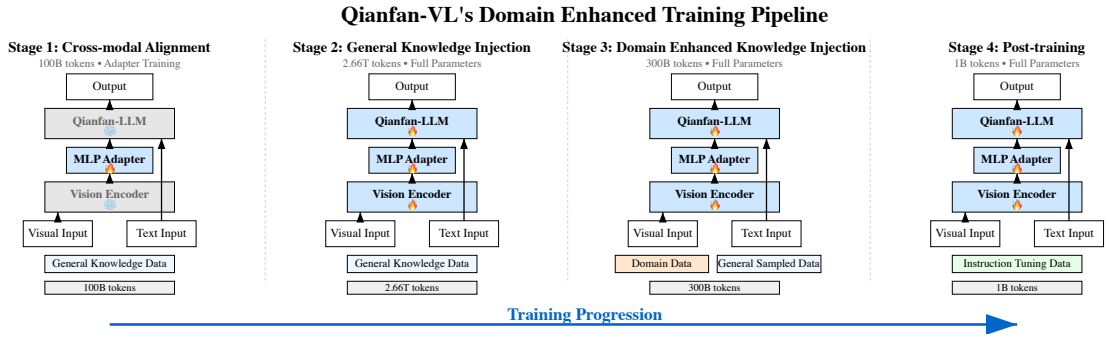


Figure 2: Four-stage progressive training pipeline of Qianfan-VL. The pipeline systematically builds capabilities from cross-modal alignment (Stage 1) through general knowledge injection (Stage 2), domain enhancement (Stage 3), to final instruction tuning (Stage 4). Each stage carefully balances different data types and training objectives to achieve domain enhancement while maintaining general capabilities.

Stage 1: Cross-Modal Alignment (100B tokens) In this initial stage, we establish the fundamental connection between visual and linguistic modalities to build a vision-language mapping foundation. We adopt a conservative training strategy where only the MLP adapter parameters are updated while keeping both the vision encoder and language model completely frozen. The training data consists of 100B tokens of high-quality image-caption pairs and basic visual question-answering tasks. Our experiments demonstrate that this stage is necessary for stable training - without it, we observe unstable loss curves during the early phase of Stage 2, which negatively impacts final model performance. The frozen encoder strategy ensures that pre-trained representations remain intact while the adapter learns to bridge the modality gap.

Stage 2: General Knowledge Injection (2.66T tokens) This stage focuses on injecting massive amounts of general knowledge while establishing robust multimodal understanding through our comprehensive dataset collection. We perform full parameter updates across all model components - vision encoder, language model, and adapter.

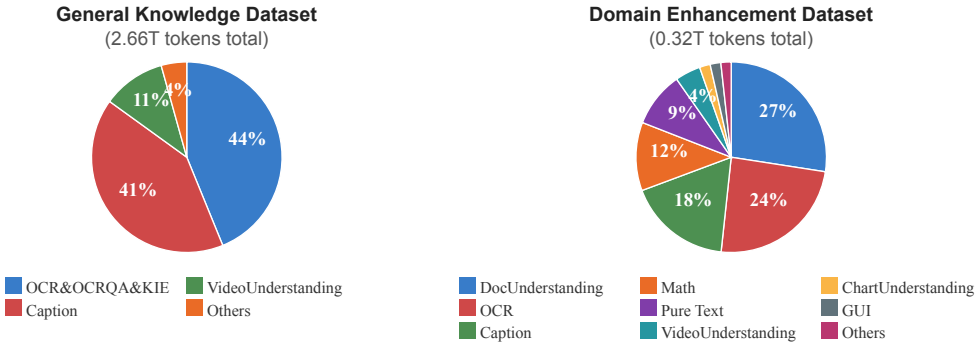


Figure 3: Distribution of training data across different task categories. Left: General Knowledge Dataset (2.66T tokens total) with OCR&OCRQA&KIE (43.8%) and Caption (41.1%) tasks dominating, VideoUnderstanding (10.7%), and Others (4.3%) including Grounding, ChartUnderstanding, DocUnderstanding, GUI, Knowledge, Math, and VQA. Right: Domain Enhancement Dataset (0.32T tokens total) with more balanced distribution across specialized domains.

The General Knowledge Dataset comprises 2.66T tokens distributed across major task categories, as shown in Figure 3. The dataset emphasizes document understanding and visual description capabilities, with OCR&OCRQA&KIE tasks (43.8%) combining open-source datasets including EATEN (Guo et al., 2019), VCR (Zhang et al., 2024b), CASIA (Liu et al., 2020), WKVVQA (Shah et al., 2019), LLaVAR (Zhang et al., 2023), and A-OKVQA (Schwenk et al., 2022) with proprietary OCR synthesis datasets for robust text recognition and key information extraction. Caption generation (41.1%) leverages ShareGPT4Video (Chen et al., 2024b), GRIT (You et al., 2023), MedTrinity (Xie et al., 2024), WebSight (Laurençon et al., 2024), and COYO-700M along with in-house video caption datasets. VideoUnderstanding (10.7%) integrates LLaVA-Video-178K and ChinaOpen with proprietary video datasets for temporal reasoning. The remaining

tasks (4.3%) encompass Grounding (All-Seeing-V2, V3Det, RefCOCO/+/g (Kazemzadeh et al., 2014)), ChartUnderstanding (Masry et al., 2022), DocUnderstanding (Mathew et al., 2021), GUI understanding, general Knowledge tasks, Math reasoning datasets, and VQA benchmarks, ensuring comprehensive coverage of multimodal understanding scenarios.

Stage 3: Domain Enhancement (0.32T tokens) This critical stage implements targeted capability enhancement for enterprise-critical domains. We maintain full parameter updates while carefully balancing the training mixture with 70% meticulously curated domain-specific data and 30% general data sampling to maintain broad capabilities. As shown in the right panel of Figure 3, the Domain Enhancement Dataset (0.32T tokens) has a more balanced distribution across specialized domains. DocUnderstanding (27.4%) focuses on comprehensive document analysis including contracts, invoices, reports, and academic papers with complex layout understanding. OCR (24.3%) addresses advanced text recognition tasks including handwriting, scene text, formula recognition, and multi-language scripts. Caption generation (17.6%) targets domain-specific visual descriptions for specialized content and technical imagery. Mathematical reasoning (11.6%) covers K-12 to university-level problems with detailed solutions and proof verification. Pure Text Pretraining Data (9.4%) maintains language modeling capabilities, while VideoUnderstanding (4.3%) provides specialized video analysis, ChartUnderstanding (1.8%) enables business intelligence and data visualization interpretation, and GUI tasks (1.8%) support user interface understanding. We employ curriculum learning with adaptive difficulty scheduling, starting with simple OCR tasks and progressively introducing complex multi-step reasoning problems, ensuring stable learning and preventing overfitting to specific task patterns.

Stage 4: Post-training with Instruction Tuning (1B tokens) The final stage focuses on post-training through comprehensive instruction tuning to enhance the model's instruction following capabilities. We continue full parameter updates with a carefully curated dataset of 1B tokens encompassing complex instruction following (multi-step tasks, conditional logic, edge cases), writing and generation (reports, summaries, creative content), question answering (factual, analytical, and reasoning-based queries), programming assistance (code understanding, debugging, documentation), and domain-specific instructions (OCR formatting, mathematical notation, chart interpretation). For tasks requiring logical reasoning and mathematical computation—such as chart question answering, visual problem solving, and visual reasoning—we employ long chain-of-thought (Long CoT) techniques to significantly enhance the model's reasoning capabilities. These Long CoT traces provide detailed step-by-step reasoning paths and intermediate computational steps, enabling the model to tackle complex multi-step problems with improved accuracy. Critically, we include substantial pure-text instruction data to maintain the language model's capabilities while systematically improving its ability to handle reasoning-intensive visual tasks. Additionally, we perform model merging on the best-performing checkpoints from different training runs to combine their complementary strengths, resulting in enhanced overall performance across all evaluation metrics.

3.2 Data Synthesis Pipeline

We develop comprehensive data synthesis pipelines for six major task categories, combining traditional computer vision models with programmatic generation to create high-quality training data at unprecedented scale. Our synthesis approach emphasizes diversity, accuracy, and real-world applicability:

Pipeline	Key Features
Document OCR	Multi-format, noise simulation
Mathematics	K-12 to university, step-by-step
Charts	15+ types, Q&A pairs
Tables	Complex structures, 50+ themes
Formulas	Multi-engine, handwriting
Scene OCR	Natural embedding, multilingual

Table 4: Data synthesis pipeline summary showing key characteristics of each category.

Document OCR Pipeline: Our document OCR pipeline implements three core functionalities: full document parsing through multi-dimensional analysis combining layout detection, content extraction, and structural understanding (supporting multilingual documents including handwritten and scanned materials); image-to-Markdown conversion for efficient transformation of single/multi-page documents into structured format preserving formatting and hierarchy; and document Q&A capabilities supporting summarization, reasoning, and multi-turn dialogue about content. Data sources include DocVQA (Mathew et al., 2021), DocReason25K, and proprietary synthesized datasets, with robustness enhancements applied through bitmap rendering, morphological operations (erosion/dilation), Gaussian blur, and other noise simulation techniques inspired by document augmentation approaches (Groleau et al., 2022). Quality

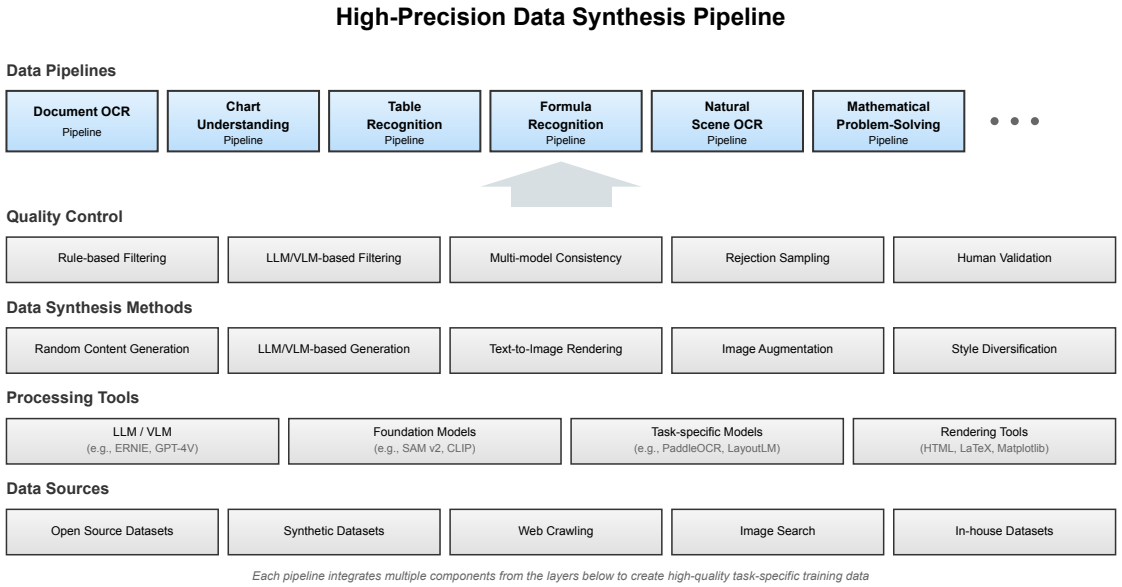


Figure 4: Overview of the data synthesis pipeline architecture, showing the integrated workflow from data sources through various processing stages including synthesis tools, quality control, and specialized data pipelines for different modalities.

assurance employs multi-VLM cross-validation with agreement thresholds, and the pipeline generates documents with varying complexity levels, from simple forms to complex multi-column layouts with embedded tables and figures.

Mathematical Problem-Solving Pipeline: Our mathematical pipeline addresses the unique challenges of educational scenarios through comprehensive educational data preprocessing that collects multilingual high-quality problem-solving data with standardized terminology and notation, performing structured decomposition into problem statements, conditions, solution steps, and formulas. The solution synthesis follows a knowledge-guided generation pipeline from structured representation through LaTeX formatting and HTML rendering to image generation, creating realistic photo-solving scenarios with specialized handling of charts, formulas, and geometric figures using Markdown, LaTeX, and Asymptote formal description languages. We enhance diversity through multiple handwriting styles, paper backgrounds, and lighting conditions to simulate K-12 through university-level scenarios, while quality verification employs rule-based filtering, rejection sampling, multi-model voting, and OCR character-level validation to ensure mathematical correctness. The pipeline covers algebra, geometry, calculus, statistics, and linear algebra with adaptive difficulty scaling.

Chart Understanding Pipeline: Our chart understanding pipeline automates the generation of high-quality chart Q&A pairs through a comprehensive process beginning with data expansion via open-source dataset sampling combined with web crawling through image search APIs, followed by deduplication. Pre-trained VLMs generate structured summaries containing both visual and numerical information, which feed into a two-stage generation process where questions are first generated from summaries, then answers are produced based on questions and summaries. We incorporate LaTeX rendering through ArXiv paper crawling with regex extraction and TexLive re-rendering for precise mathematical chart descriptions inspired by Nougat (Blecher et al., 2023), while quality control employs chain-of-thought model verification combined with human review to ensure accuracy. The pipeline produces three question types: data retrieval (exact value extraction), visual attributes (color, style, layout), and computational Q&A (aggregation, comparison, trend analysis), covering 15+ chart types including bar, line, pie, scatter, heatmap, box plot, and complex composite visualizations.

Table Recognition Pipeline: Our table pipeline addresses two core capabilities: table structure recovery for precise conversion of image tables to HTML/LaTeX (supporting borderless tables, contract forms, and complex layouts with merged cells), and table Q&A for numerical computation, comparative analysis, and information retrieval based on table images. The synthesis process employs content generation through random table structures (3-20 rows/columns) populated via the Faker library and LLM-based realistic data filling with random cell merging, combined with visual rendering using 50+ professional CSS themes (statistical reports, technical documents, financial statements) rendered through Jinja2+KaTeX engines. Data augmentation applies geometric transformations, color perturbations, and blur effects for diversity, drawing from sources including TabMWP (Lu et al., 2022b), MMC-Inst (Liu et al., 2023),

BigDocs (Rodriguez et al., 2024), and proprietary synthetic datasets. The pipeline handles complex scenarios like multi-level headers, footnotes, and cross-references.

Formula Recognition Pipeline: Our formula pipeline achieves symbol recognition, syntax parsing, and semantic understanding through comprehensive symbol coverage (mathematical symbols, Greek letters, special notations, and domain-specific markers) and structure parsing of complex elements including fractions, radicals, subscripts/superscripts, matrices, and tensor notations. Semantic mapping links formula semantics to mathematical concepts for deeper understanding, while multi-engine rendering with MathJax and KaTeX ensures cross-platform consistency, complemented by handwriting simulation featuring diverse writing styles, paper textures, ink variations, and noise patterns. The dataset spans from elementary topics (arithmetic operations, basic algebra) through intermediate concepts (trigonometry, logarithms, sequences) to advanced mathematics (calculus, differential equations, linear algebra, abstract algebra), with each formula including LaTeX source, rendered image, and semantic annotations.

Natural Scene OCR Pipeline: Our pipeline, inspired by SynthText (Gupta et al., 2016), implements systematic text-in-image synthesis through background screening using lightweight OCR models (Cui et al., 2025) and image type detection to eliminate text-containing and non-static samples, combined with scene understanding via semantic segmentation (Ravi et al., 2024) and monocular depth estimation (Yang et al., 2024) for region division and 3D structure. Realistic projection employs plane detection with perspective transformation and random text styling for natural embedding, while fusion enhancement through Poisson blending ensures consistent occlusion, shadows, and texture integration. The pipeline generates diverse scenes including street environments (signs, storefronts, billboards), indoor settings (product labels, menus, notices), and documents in context (whiteboards, presentations, posters), with annotations providing character-level and word-level bounding boxes along with reading order information. Text languages cover over 12 languages with appropriate fonts and styles.

3.3 Complex Instruction Enhancement

Enterprise applications require sophisticated instruction-following beyond simple queries. We address this gap by evolving simple prompts into multi-constraint instructions through a systematic pipeline:

(1) Seed Construction and Mining: We select domain-relevant images paired with initial prompts, then expand coverage using SFTMiner, our retrieval engine supporting text-to-image and image-to-image search across massive repositories.

(2) Prompt Evolution: Starting from 5 manually designed seed prompts per scenario, we generate 5-10 simple single-constraint variants, then systematically introduce diverse constraint types (conditional reasoning, quantity limitations, sequential constraints) to create complex multi-step instructions.

(3) Response Generation: We generate responses using advanced VLMs, forming `<image, prompt, response>` triplets with multi-model voting and consistency checking for quality assurance.

Through this pipeline, we synthesize approximately 200K complex instruction samples covering conditional extraction, multi-constraint analysis, reasoning with rejection handling, and hierarchical decomposition, significantly enhancing the model's ability to handle sophisticated enterprise requirements.

3.4 Chain-of-Thought Training

For the 8B and 70B variants, we implement sophisticated chain-of-thought (CoT) reasoning capabilities (Wei et al., 2022) through a multi-faceted approach:

Token-Activated Reasoning: We introduce special tokens (`<think>` and `</think>`) to delineate reasoning processes, allowing users to explicitly request reasoning by including these tokens. The model learns to generate intermediate reasoning steps within these boundaries while providing final answers outside the thinking tokens for clarity. Figure 5 illustrates the difference between standard mode and token-activated reasoning mode, showing how the thinking tokens enable explicit chain-of-thought generation.

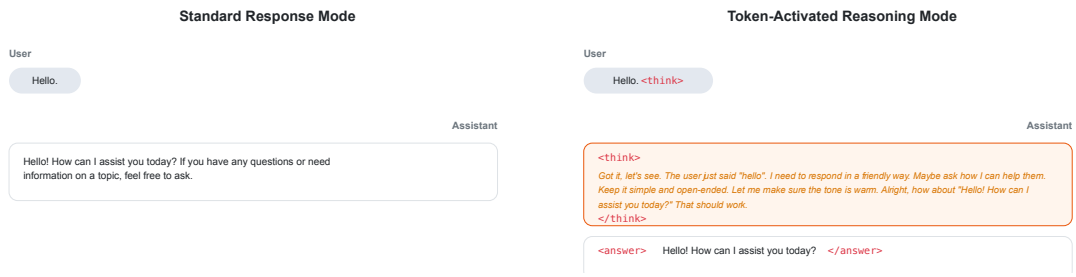


Figure 5: Comparison between standard response mode and token-activated reasoning mode. The left panel shows direct response generation, while the right panel demonstrates how <think> tokens trigger internal reasoning processes (shown in orange dashed box) that are hidden from the user, with only the final answer visible.

Training Data Construction: Our CoT training corpus primarily consists of approximately 200K mathematical reasoning problems, many of which are synthesized into multimodal data by combining mathematical problems with visual elements. We leverage advanced models with thinking capabilities such as DeepSeek-R1 to generate long chain-of-thought data with detailed step-by-step reasoning processes, ensuring comprehensive coverage of mathematical concepts from basic arithmetic to advanced calculus.

Quality Assurance: We employ rejection sampling to generate 5-10 solutions per problem and select the best via reward models, combined with process supervision to verify correctness of intermediate steps rather than just final answers. Consistency checking ensures reasoning chains are logically coherent and complete, while human validation through expert review of 10% of samples maintains quality standards.

4 Infrastructure and Implementation

4.1 Kunlun Chip Training

All Qianfan-VL models are trained on Baidu's Kunlun P800 chips, demonstrating the viability of proprietary AI infrastructure for large-scale model development.

Model Adaptation and Selection: Our infrastructure successfully adapted multiple foundation model families to the Kunlun P800 architecture, including the Llama 3.1 series across various scales and the Qwen2.5 series in different parameter configurations. We also successfully adapted the InternViT vision encoder for efficient visual processing on Kunlun chips. Through extensive ablation experiments evaluating different model combinations and scales, we determined the optimal configurations: 70B and 8B models based on Llama 3.1 architecture for their superior performance on general multimodal tasks, and the 3B model based on Qwen2.5 for its excellent efficiency in edge deployment scenarios. These experiments fully validate the capability of both Llama and Qwen series models to undergo parallel training on large-scale Kunlun chip clusters with massive datasets, achieving excellent performance outcomes. This demonstrates the maturity of proprietary AI infrastructure in supporting diverse model architectures and training paradigms at scale.

Cluster Configuration: Our training infrastructure comprises over 5000 Kunlun P800 chips operating in parallel, processing a massive 3T token training corpus while maintaining over 90% scaling efficiency at large scale.

3D Parallelism Strategy: Our parallelism strategy combines three dimensions for optimal scaling: Data Parallelism (DP) distributes batch samples across nodes with gradient accumulation, Tensor Parallelism (TP) splits model layers across chips within nodes for memory efficiency, and Pipeline Parallelism (PP) divides model depth across node groups to maximize throughput.

Key optimizations include dynamic load balancing with adaptive distribution based on layer computation patterns, optimized AllReduce gradient synchronization achieving 60% communication reduction, and 1F1B (one forward, one backward) pipeline scheduling with bubble rate below 5%. Additionally, sequence parallelism for splitting long sequences reduces memory by 50% for 32K contexts, while dynamic batching adapts batch sizes based on sequence length distribution, and selective recomputation strategically places checkpoints to balance memory and computation.

Communication-Computation Fusion: *Architectural Advantage:* The Kunlun P800 chip features a unique architecture where communication and matrix multiplication units are physically separated, unlike traditional GPUs where they compete for resources. This hardware design enables resource isolation where communication operations never block computation resources, true parallelism with simultaneous

data transfer and matrix operations, and overlap optimization that hides communication latency behind computation.

GEMM Fusion Implementation: We establish bypass streams (BypassStream) for seamless integration, enabling independent scheduling where bypass streams run parallel to main computation streams, data prefetching that initiates communication before computation needs data, and result pipelining for immediate transfer of computation results.

Multi-Stream Optimization: Taking AllGather+GEMM fusion as an example, traditional approaches complete AllGather, wait, then start GEMM sequentially. Our optimized approach transfers and computes data blocks in a pipeline fashion, achieving a 40% reduction in end-to-end latency for large operations.

5 Evaluation Results

We conduct comprehensive evaluations across general multimodal benchmarks and domain-specific tasks using the VLMEvalKit framework (Duan et al., 2024), an open-source toolkit designed for evaluating large multi-modality models. Most benchmarks employ the framework's built-in rule-based evaluation methods for consistent and reproducible assessment. For benchmarks requiring more nuanced evaluation—including mathematical reasoning tasks, subjective assessments, and those with strict output format requirements—we implement LLM-as-a-judge evaluation using Ernie-4.5-Turbo-VL (Baidu-ERNIE-Team, 2025), achieving over 95% accuracy as validated through manual assessment.

5.1 General Multimodal Benchmarks

We conduct extensive evaluations across 14 standard multimodal benchmarks to assess general visual understanding capabilities. Table 5 presents comprehensive comparisons with state-of-the-art open-source models including InternVL 3 (Zhu et al., 2025) and Qwen2.5-VL (Bai et al., 2025).

Benchmark	Qianfan-VL			InternVL-3		Qwen2.5-VL	
	3B	8B	70B	8B	78B	7B	72B
A-Bench_VAL (Zhang et al., 2024c)	75.65	75.72	**78.10**	75.86	75.86	76.49	**79.22**
CCBench (Liu et al., 2024b)	66.86	70.39	**80.98**	77.84	70.78	57.65	**73.73**
SEEDBench_IMG (Li et al., 2023a)	76.55	78.02	**79.13**	77.0	77.52	76.98	**78.34**
SEEDBench2_Plus (Li et al., 2024)	67.59	70.97	**73.17**	69.52	68.47	70.93	**73.25**
MMVet (Yu et al., 2023)	48.17	53.21	57.34	**80.28**	**78.9**	70.64	75.69
MMMU_VAL (Yue et al., 2024)	46.44	47.11	58.33	56.11	**60.78**	51.0	**65.78**
ScienceQA_TEST (Lu et al., 2022a)	95.19	97.62	**98.76**	97.97	97.17	85.47	92.51
ScienceQA_VAL (Lu et al., 2022a)	93.85	97.62	**98.81**	**97.81**	95.14	83.59	91.32
MMT-Bench_VAL (Ying et al., 2024)	62.23	63.22	**71.06**	65.17	63.67	61.4	**69.49**
MTVQA_TEST (Tang et al., 2024)	26.50	30.14	**32.18**	30.30	27.62	29.08	**31.48**
BLINK (Fu et al., 2024)	49.97	56.81	**59.44**	55.87	51.87	54.55	**63.02**
MMStar (Chen et al., 2024a)	57.93	64.07	**69.47**	**68.40**	66.07	61.53	66.00
RealWorldQA (xAI, 2024)	65.75	70.59	71.63	71.11	**74.25**	69.28	**73.86**
Q-Bench1_VAL (Wu et al., 2023)	73.51	75.25	77.46	75.99	77.99	**78.10**	**79.93**
POPE (Li et al., 2023b)	85.08	86.06	88.97	**90.59**	88.87	85.97	83.35
RefCOCO (Avg) (Kazemzadeh et al., 2014)	85.94	89.37	**91.01**	89.65	**91.40**	86.56	90.25

Table 5: Performance on general multimodal benchmarks. Top-2 results for each benchmark are in bold.

Qianfan-VL demonstrates competitive performance across general benchmarks:

- **Scientific reasoning:** Achieves 98.17% on ScienceQA_TEST (8B/70B), surpassing most comparable models
- **Hallucination resistance:** Strong performance on POPE (88.79% for 70B) indicates robust grounding
- **Visual perception:** Competitive scores on SEEDBench_IMG (78.85% for 70B) and Q-Bench (78.33% for 70B)
- **Real-world understanding:** Solid performance on RealWorldQA (71.76% for 70B) demonstrates practical applicability

However, we observe relatively lower performance on MMMU (46.44%-58.33%) and MMVet (48.17%-57.34%) compared to leading models. Analysis of failure cases reveals that these gaps primarily stem from insufficient coverage of general knowledge questions, particularly those requiring broad understanding of diverse topics and complex reasoning across multiple domains. This limitation can be addressed in future iterations by incorporating more interleaved image-text data covering general knowledge domains,

which would enhance the model's ability to handle open-ended questions requiring extensive world knowledge while maintaining our strong domain-specific capabilities.

Despite these areas for improvement, Qianfan-VL achieves competitive results across most benchmarks while being optimized for domain-specific tasks, validating our approach of enhancing specialized capabilities without sacrificing overall general performance.

5.2 OCR and Document Understanding

Table 6 shows performance on OCR and document understanding tasks, where Qianfan-VL exhibits significant advantages.

Benchmark	Qianfan-VL 3B	Qianfan-VL 8B	Qianfan-VL 70B	InternVL-3 8B	InternVL-3 78B	Qwen2.5-VL 3B	Qwen2.5-VL 7B	Qwen2.5-VL 72B
OCRBench (Chen et al., 2024c)	831	854	873	**881**	847	810	**883**	874
AI2D_TEST (**?**)	81.38	**85.07**	**87.73**	85.07	83.55	77.07	80.47	83.84
OCRVQA_TEST (**?**)	66.15	68.98	**74.06**	39.03	35.58	69.24	**71.02**	66.8
TextVQA_VAL (**?**)	80.11	82.13	**84.48**	82.15	83.52	79.09	**84.96**	83.26
DocVQA_VAL (Mathew et al., 2021)	90.85	93.54	**94.75**	92.04	83.82	92.71	**94.91**	**95.75**
ChartQA_TEST (Masry et al., 2022)	81.79	**87.72**	**89.60**	85.76	82.04	83.40	86.68	87.16

Table 6: Performance on OCR and document understanding benchmarks. Top-2 results for each benchmark are in bold.

The domain enhancement strategy yields exceptional results:

- **Document understanding:** Qianfan-VL-70B achieves 94.75% on DocVQA, with the series maintaining competitive performance across all model sizes
- **Chart analysis:** 89.60% on ChartQA_TEST (70B), leading among comparable model sizes
- **OCR accuracy:** Strong showing on OCRBench (Chen et al., 2024c) (873 for 70B), with competitive performance across the model series
- **Visual question answering:** Competitive on TextVQA (84.48% for 70B) and OCRVQA (74.06% for 70B)

These results validate our targeted data synthesis and domain-focused training approach, showing particular strength in document-centric tasks critical for enterprise applications.

5.3 Mathematical Reasoning

Mathematical reasoning capabilities are evaluated on specialized benchmarks as shown in Table 7.

Benchmark	Qianfan-VL 8B	Qianfan-VL 70B	InternVL-3 8B	InternVL-3 78B	Qwen2.5-VL 7B	Qwen2.5-VL 72B
Mathvista-mini (Lu et al., 2023)	69.19	**78.60**	69.50	70.10	67.20	73.90
Mathvision (Wang et al., 2024)	32.82	**50.29**	29.61	34.8	25.95	39.34
Mathverse (Zhang et al., 2024a)	48.4	**61.04**	43.68	49.26	44.21	55.18
ChartQA Pro (Masry et al., 2025)	50.41	**52.00**	37.32	44.43	43.73	45.30
HallusionBench (Guan et al., 2024)	51.72	**54.52**	49.20	40.20	47.90	49.90
InHouse Dataset A	59.87	**71.78**	40.64	41.47	45.58	57.20
InHouse Dataset B	61.33	**75.60**	36.25	42.65	30.62	59.68

Table 7: Mathematical reasoning performance with CoT enabled. Best results for each benchmark are in bold.

The chain-of-thought training delivers remarkable improvements in mathematical and visual reasoning:

- **MathVista performance:** Qianfan-VL-70B achieves 78.60% on Mathvista-mini, demonstrating state-of-the-art performance among open-source models
- **Complex visual math:** 50.29% on Mathvision (70B), significantly outperforming comparable open models
- **Multi-step reasoning:** 61.04% on Mathverse (70B), demonstrating strong capability in problems requiring multiple reasoning steps
- **Chart-based reasoning:** 52.00% on ChartQA Pro (70B), leading among open-source models
- **Hallucination mitigation:** 54.52% on HallusionBench (70B), showing improved factual grounding with CoT

- **Proprietary benchmarks:** Strong performance on internal datasets (71.78% and 75.60% for 70B) validates real-world applicability

The consistent improvements across diverse reasoning tasks demonstrate that our CoT training methodology effectively enhances complex problem-solving capabilities beyond simple pattern matching.

5.4 Ablation Studies On The Validity of Domain Enhancement

We conduct comprehensive ablation studies to validate the effectiveness of our domain enhancement strategy (Stage 3). To isolate the impact of domain-specific training, we fix the data in Stages 1, 2, and 4 while comparing models trained with and without Stage 3.

Config	Document Understanding Tasks				
	DocVQA	AI2D	ChartQA	Doc Text*	Simple HTML*
w/o Stage 3	93.67	84.59	86.68	73.70	68.70
w/ Stage 3	94.13	85.33	88.00	75.00	72.10
Gain	+0.46	+0.74	+1.32	+1.30	+3.40

Config	OCR Tasks					
	OCRBench	Complex HTML*	Chart Dict*	LaTeX*	Struct Extract*	Handwriting*
w/o Stage 3	837	78.33	75.73	85.04	33.17	75.69
w/ Stage 3	852	82.00	77.21	86.16	35.85	83.89
Gain	+15	+3.67	+1.48	+1.12	+2.68	+8.20

Table 8: Impact of Stage 3 (Domain Enhancement) on OCR and document understanding tasks. * indicates in-house datasets.

Configuration	Mathematical Benchmarks				
	Mathvista-mini	Mathvision	Mathverse	In-House A	In-House B
w/o Stage 3	75.50	45.52	57.34	59.28	65.83
w/ Stage 3	77.10	51.54	59.95	77.28	73.01
Gain	+1.60	+6.02	+2.61	+18.00	+7.18

Table 9: Impact of Stage 3 (Domain Enhancement) on mathematical reasoning tasks.

Key Findings: The ablation results clearly demonstrate the effectiveness of our domain enhancement strategy:

- **OCR Performance:** Stage 3 consistently improves OCR-related tasks, with particularly strong gains in handwritten text recognition (+8.20%) and complex HTML table structure (+3.67%). Even challenging tasks like structured information extraction show meaningful improvements (+2.68%).
- **Mathematical Reasoning:** Domain enhancement yields substantial improvements across all mathematical benchmarks. The most significant gains appear in our in-house datasets (+18.00% and +7.18%), while public benchmarks also benefit (Mathvision +6.02%, Mathverse +2.61%).
- **Document Understanding:** Complex document tasks benefit significantly from domain-specific training. HTML table structure recognition improves by 3.40% for simple tables and 3.67% for complex tables, demonstrating enhanced structural understanding.
- **Consistent Improvements:** All 16 evaluated tasks show positive gains with Stage 3, with no performance regressions observed. This validates that domain enhancement complements rather than conflicts with general training.

These results confirm that targeted domain enhancement in Stage 3 is crucial for achieving state-of-the-art performance in specialized tasks while maintaining general multimodal capabilities. The consistent improvements across diverse benchmarks validate our four-stage training strategy.

Importantly, Stage 3 training incurs relatively low computational costs while delivering significant performance improvements for domain-specific tasks. This cost-effectiveness enables efficient customization for different application domains—organizations can start from our Stage 2 checkpoint and apply only Stage 3 and Stage 4 training with domain-specific data to achieve targeted enhancements. This modular approach significantly reduces the barrier for developing specialized vision-language models tailored to specific industry needs.

6 Limitations and Future Work

While the Qianfan-VL series demonstrates strong performance across various benchmarks, several functional limitations remain that we plan to address in future iterations. The current models support a maximum context length of 32K tokens, which limits their ability to process lengthy documents, multi-page PDFs, or engage in extended multi-turn conversations. This constraint particularly affects applications requiring comprehensive document analysis or complex reasoning chains that span extensive textual and visual information. Additionally, despite various optimizations implemented throughout the training process, the models still require substantial computational resources for inference, especially when processing high-resolution images or multiple visual inputs simultaneously. This computational burden limits deployment scenarios in resource-constrained environments such as mobile devices or edge computing platforms. Furthermore, while the models excel at OCR and document understanding tasks, they currently lack certain advanced capabilities such as video understanding, 3D spatial reasoning, and fine-grained temporal analysis that are becoming increasingly important in real-world multimodal applications.

To address these limitations, we are pursuing several technical solutions and future research directions. We are actively working on extending the context window to 128K tokens and beyond through techniques like sparse attention mechanisms and hierarchical encoding strategies, which will enable the processing of entire books, lengthy technical documents, and maintenance of context across extended dialogues. For computational efficiency, we plan to integrate NaViT (Native Resolution ViT) techniques (Dehghani et al., 2024) to process images at their native resolutions without resizing, thereby reducing computational overhead while maintaining accuracy. We are also exploring quantization methods and model distillation approaches to create lighter variants suitable for edge deployment without significant performance degradation. In terms of capability expansion, future versions will incorporate specialized training for video understanding, 3D scene comprehension, and temporal reasoning. We also plan to enhance multilingual capabilities across more languages, and develop domain-specific variants for medical imaging, scientific diagrams, and technical blueprints through targeted fine-tuning strategies. These improvements will position the Qianfan-VL series as a more versatile and efficient solution for diverse multimodal understanding tasks across various domains and deployment scenarios.

7 Conclusion

Qianfan-VL represents a significant advancement in domain-enhanced vision-language models, successfully balancing general multimodal capabilities with specialized expertise in critical enterprise domains. Through innovative training strategies, large-scale data synthesis, and efficient infrastructure utilization, we demonstrate that targeted domain enhancement can be achieved without sacrificing broad applicability.

The four-stage progressive training pipeline provides a principled approach to capability development, while our comprehensive data synthesis techniques ensure high-quality training data for specialized tasks. The successful training on Kunlun chips validates the maturity of proprietary AI infrastructure for large-scale model development.

Qianfan-VL's strong performance on both general and domain-specific benchmarks, combined with flexible deployment options across different model sizes, makes it a practical solution for diverse enterprise multimodal applications. Detailed application showcases demonstrating the model's capabilities in document processing, educational scenarios, and business intelligence are provided in Appendix A. As we continue to expand capabilities and optimize performance, Qianfan-VL aims to bridge the gap between research advances and real-world deployment needs.

Acknowledgments

We thank the Baidu AI Cloud team for infrastructure support, the Baige and Kunlun teams for AI infrastructure assistance, and all contributors to the QianFan platform. We are deeply grateful to the operations, storage, and network teams for maintaining the stability of the P800 clusters, enabling Qianfan-VL to train successfully at massive scale on 5000+ P800 chips. Special thanks to our annotation teams and quality assurance engineers for their meticulous work in data validation. There are many more colleagues who contributed to this project's success whom we cannot acknowledge individually.

Contributors

Core Contributors

- Daxiang Dong*
- Mingming Zheng
- Dong Xu
- Bairong Zhuang
- Wenyu Zhang
- Chunhua Luo
- Haoran Wang
- Zijian Zhao
- Jie Li
- Yuxuan Li
- Hanjun Zhong
- Mengyue Liu
- Jieting Chen
- Shupeng Li
- Jianmin Wu

Contributors

- Lun Tian
- Yaping Feng
- Xin Li
- Donggang Jiang
- Yong Chen
- Yehua Xu
- Duohao Qin
- Chen Feng
- Dan Wang
- Henghua Zhang
- Jingjing Ha
- Jinhui He
- Yanfeng Zhai
- Chengxin Zheng
- Jiayi Mao
- Jiacheng Chen
- Ruchang Yao
- Ziye Yuan
- Guangjun Xie**
- Dou Shen**

* Project Lead
** Project Sponsor

References

J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Baidu-ERNIE-Team. Ernie 4.5 technical report. https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf, 2025.

L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, Z. Tang, L. Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024b.

Y. Chen, L. Zhang, H. Liu, et al. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2024c.

Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2024d.

C. Cui, T. Sun, M. Lin, T. Gao, Y. Zhang, J. Liu, X. Wang, Z. Zhang, C. Zhou, H. Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

A. Groleau, K. W. Chee, S. Larson, S. Maini, and J. Boarman. Augraphy: A data augmentation library for document images. *arXiv preprint arXiv:2208.14558*, 2022.

T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

H. Guo, X. Qin, J. Liu, J. Han, J. Liu, and E. Ding. Eaten: Entity-aware attention for single shot visual text extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 254–259. IEEE, 2019.

A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016.

A. Hu, H. Xu, J. Ye, M. Yan, L. Zhang, B. Zhang, C. Li, J. Zhang, Q. Jin, F. Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

H. Laurençon, L. Tronchon, and V. Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024. URL https://arxiv.org/abs/2403.09029.

B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

B. Li, Y. Ge, Y. Chen, Y. Ge, R. Zhang, and Y. Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

B. Liu, X. Xu, and Y. Zhang. Offline handwritten chinese text recognition with convolutional neural networks. *arXiv preprint arXiv:2006.15619*, 2020.

F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoob, and D. Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024b.

P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022a.

P. Lu, L. Qiu, K.-W. Chang, Y. N. Wu, S.-C. Zhu, T. Rajpurohit, P. Clark, and A. Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022b.

P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

A. Masry, M. S. Islam, M. Ahmed, A. Bajaj, F. Kabir, A. Kartha, M. T. R. Laskar, M. Rahman, S. Rahman, M. Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

M. Mathew, D. Karatzas, and C. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, pages 8748–8763, 2021.

N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

J. Rodriguez, X. Jian, S. S. Panigrahi, T. Zhang, A. Feizi, A. Puri, A. Kalkunte, F. Savard, A. Masry, S. Nayak, et al. Bigdocs: An open dataset for training multimodal models on document and code tasks. *arXiv preprint arXiv:2412.04626*, 2024.

D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019.

J. Tang, Q. Liu, Y. Ye, J. Lu, S. Wei, C. Lin, W. Li, M. F. F. B. Mahmood, H. Feng, Z. Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, and H. Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.

xAI. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model., 2024. URL https://x.ai/blog/grok-1.5v.

Y. Xie, C. Zhou, L. Gao, J. Wu, X. Li, H.-Y. Zhou, S. Liu, L. Xing, J. Zou, C. Xie, and Y. Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine, 2024. URL https://arxiv.org/abs/2408.02900.

L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, et al. Ferret: Refer and ground anything at any granularity, 2023. URL https://arxiv.org/abs/2310.07704.

W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

B. Zhang and R. Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, Y. Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024a.

T. Zhang, S. Wang, L. Li, G. Zhang, P. Taslakian, S. Rajeswar, J. Fu, B. Liu, and Y. Bengio. Vcr: A task for pixel-level complex reasoning in vision language models via restoring occluded text. *arXiv preprint arXiv:2406.06462*, 2024b.

Y. Zhang, R. Koh, J. Zhang, J. Aneja, Y. Yang, H. Ding, H. Li, and H. Pfister. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

Z. Zhang, H. Wu, C. Li, Y. Zhou, W. Sun, X. Min, Z. Chen, X. Liu, W. Lin, and G. Zhai. A-bench: Are lmms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*, 2024c.

J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Application Showcases

This appendix presents detailed examples showcasing Qianfan-VL's capabilities across diverse enterprise scenarios and real-world applications. All images used in these showcases are sourced either from open-source datasets or collected from publicly available internet resources. The copyright of these images belongs to their respective owners, and they are used here solely for demonstration purposes.

A.1 Intelligent Document Processing

Scene Text Recognition:

This example demonstrates Qianfan-VL's advanced OCR capabilities in challenging real-world scenarios:

- **Model Capabilities:** Complex scene text detection, multi-directional text recognition, and handling of various fonts and sizes in natural images
- **Technical Challenges:** The image contains text at different angles, varying lighting conditions, partial occlusions, and multilingual content with special characters
- **Performance Highlights:** Accurate extraction of all visible text including phone numbers, distances, and location descriptions despite visual complexity

Image:	
	**Prompt:** 请识别图中文字 **Response:** 实景现房中赫出品 颐和园公园南门800米/北侧距万柳高尔夫80米 心30米/距地铁巴沟站C出口10米 T·6271

Table 10: Scene Text Recognition in Real Estate Advertisement

Invoice Information Extraction:

This case showcases Qianfan-VL's structured information extraction and JSON generation capabilities:

- **Model Capabilities:** Document layout understanding, table structure recognition, multi-language processing, and structured output generation
- **Technical Challenges:** Complex invoice layouts with nested tables, multilingual text, numerical data extraction, and maintaining relationships between different invoice fields
- **Performance Highlights:** Precise extraction of all invoice fields into well-formatted JSON, including item-level details with quantities and prices, while handling missing information gracefully

Image:

Prompt:
你是发票提取专家。分析提供的图像（这是一个中文增值税发票），提取：
- 发票号码
- 开票日期 (YYYY-MM-DD)
- 供应商名称
- 购买方名称
- 总金额 (含货币)
- 项目列表：每个含描述、数量、单价、小计
缺失信息用"N/A"。仅输出JSON。

Response:


```
{
 "发票号码": "0110020001",
 "开票日期": "2021年10月",
 "供应商名称": "北京成药安宫牛黄丸有限公司",
 "购买方名称": "华康医药集团有限公司",
 "含税总额": "1580.70",
 "项目列表": [
  {
   "项目名称": "成药-安宫牛黄丸",
   "数量": "1",
   "金额": "749.11"
  },
  {
   "项目名称": "成药*益安宁丸",
   "数量": "1",
   "金额": "758.71"
  },
  {
   "项目名称": "学资品制剂*薯香正气软胶囊",
   "数量": "2",
   "金额": "72.88"
  }
 ]
}
```


Table 11: Invoice Information Extraction with Structured JSON Output

A.2 Educational Applications

Advanced Mathematical Reasoning:

This example demonstrates Qianfan-VL's sophisticated mathematical understanding and problem-solving abilities:

- **Model Capabilities:** Mathematical notation recognition, multi-step reasoning, geometric understanding, and step-by-step solution generation
- **Technical Challenges:** Complex quadratic function analysis, coordinate system interpretation, multiple solution paths, and clear mathematical explanation in multiple languages
- **Performance Highlights:** Complete solution with vertex formula application, discriminant analysis, and systematic verification of all answer choices with detailed reasoning

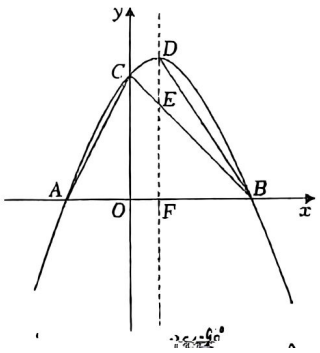
Technical Diagram Analysis:

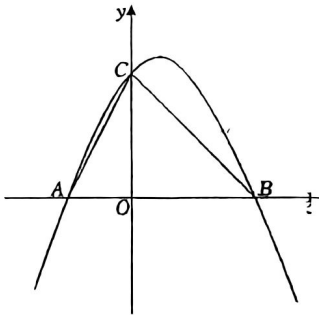
Image:

如图，二次函数 $y=-x^2+2mx+2m+1$（m 是常数，且 $m>0$）的图象与 x 轴交于 A，B 两点（点 A 在点 B 的左侧），与 y 轴交于点 C，顶点为 D．其对称轴与线段 BC 交于点 E，与 x 轴交于点 F．连接 AC，BD．

（1）求 A，B，C 三点的坐标（用数字或含 m 的式子表示），并求 $\angle OBC$ 的度数；

（2）若 $\angle ACO=\angle CBD$，求 m 的值；





备用图

Prompt: 帮我求解这道题，给出详细步骤。

Response:

<think>
用户现在需要解决这道二次函数的题目，分两问。首先看第一问，要求求出A、B、C三点的坐标，并且求∠OBC的度数。首先，二次函数的解析式是y = -x² + 2mx + 2m + 1，m > 0。
...
[详细推导过程省略]
...
所以第二问的答案是m = 1。
</think>

<answer>
（1）求A，B，C三点的坐标及∠OBC的度数

求A、B两点坐标
二次函数与x轴交于A、B两点，即令\(y=0\)，解方程：
\[-x^2 + 2mx + 2m + 1 = 0 \]
两边同乘\(-1\)得：

[篇幅原因，省略大量内容]

答案
（1）\(A(-1, 0)\)，\(B(2m + 1, 0)\)，\(C(0, 2m + 1)\)，\(\angle OBC = 45^\circ\)；
（2）\(m = 1\)。
</answer>

Table 12: Advanced Mathematical Reasoning

Image:

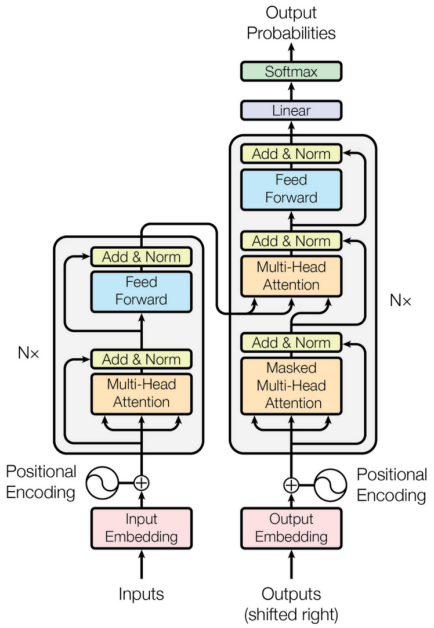


Figure 1: The Transformer - model architecture.

Prompt:
非常详细解释这张图表的含义，解释清楚每一个元素。

Response:
这张图展示了Transformer模型的架构，这是一种广泛应用于自然语言处理和其他序列任务的深度学习模型。我将详细解释图中的每个元素：
整体结构
图1展示了Transformer模型的基本架构，包括编码器（左侧）和解码器（右侧）的组件。Transformer模型由多个这样的编码器-解码器模块堆叠而成（图中标注为"N×"，表示重复多次）。
左侧：编码器（**Encoder**）
1. 输入嵌入（**Input Embedding**）：接收输入序列（如词序列），将每个词转换为固定维度的向量表示。
2. 位置编码（**Positional Encoding**）：添加表示序列中元素位置的信息。
3. 多头注意力（**Multi-Head Attention**）：允许模型关注输入序列的不同位置。
4. 前馈神经网络（**Feed Forward**）：包含ReLU激活函数的简单神经网络。
5. 残差连接与归一化（**Add & Norm**）：残差连接和层归一化处理。
右侧：解码器（**Decoder**）
包括额外的掩码多头注意力和编码器-解码器注意力组件，用于序列生成任务。
输出层
线性层和Softmax层将解码器输出转换为概率分布。
核心创新在于自注意力机制，使模型能够高效处理长序列并捕捉不同位置间的关系。

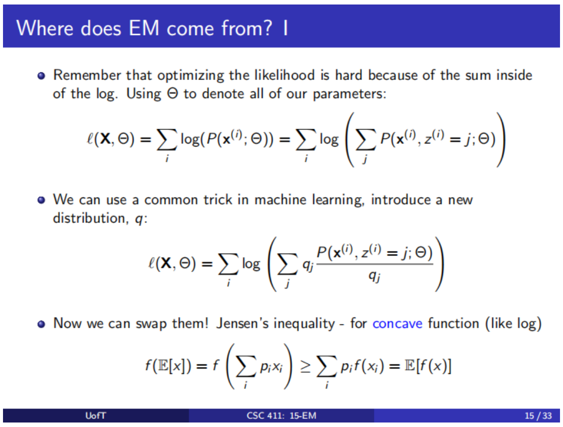
Table 13: Technical Architecture Diagram Analysis and Explanation

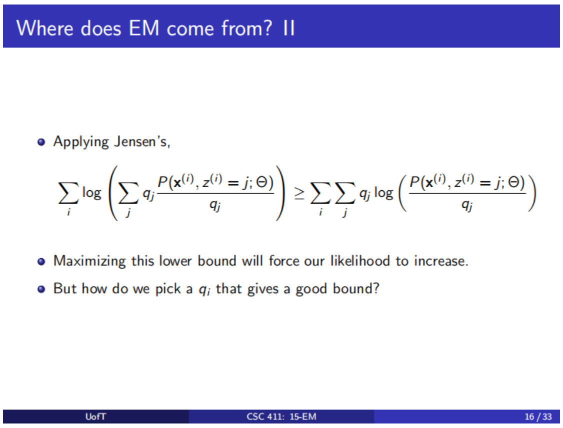
This case highlights Qianfan-VL's ability to understand and explain complex technical architectures:

- **Model Capabilities:** Technical diagram interpretation, component relationship understanding, data flow analysis, and technical documentation generation
- **Technical Challenges:** Understanding transformer architecture details, identifying component connections, explaining attention mechanisms, and providing implementation-ready descriptions
- **Performance Highlights:** Accurate identification of all architectural components, clear explanation of information flow, and practical code structure suggestions

Algorithm Theory Explanation:

Images:





Prompt:
帮我解释下面这些slides的详细内容。<think>

Response:

用户现在需要我解释这两张关于EM算法来源的幻灯片内容。首先，我得回忆一下EM算法的基本原理，特别是它如何通过Jensen不等式来处理含有和的log问题。首先看第一张幻灯片，标题是"Where does EM come from? I"。

第一张幻灯片开头提到，优化似然函数很困难，因为log里面有和。这里用Θ表示所有参数，似然函数ℓ(X, Θ)是log(P(X; Θ))，而P(X; Θ)可以写成对每个观测x^(i)和潜在变量z^(i)=j的概率之和，也就是Σ_j P(x^(i), z^(i)=j; Θ)。所以原式是ℓ(X, Θ) = Σ_i log(Σ_j P(x^(i), z^(i)=j; Θ))。

然后，幻灯片说引入一个新的分布q，这里q是一个关于潜在变量z的分布，可能对于每个i，q有不同的分布？不过这里写的是q_j，可能是指对于每个z^(i)=j的概率q_j^(i)？不过幻灯片里写的是"introduce a new distribution, q:"，然后表达式是Σ_i log(Σ_j q_j P(x^(i), z^(i)=j; Θ)/q_j)。这里可能是笔误或者简写，通常EM算法中，对于每个i，会有一个条件分布q(z^(i))，所以可能这里的q_j应该是q(z^(i)=j)，即对于每个观测i，潜在变量z^(i)的概率分布q。所以原式变成了ℓ(X, Θ) = Σ_i log(Σ_j q_j P(x^(i), z^(i)=j; Θ)/q_j)，这里用了q_j作为中间变量，目的是应用Jensen不等式。

接下来，Jensen不等式部分，幻灯片提到对于凹函数（concave function），比如log，有f(E[X]) ⩾ E[f(X)]。这里log是凹函数，所以可以应用Jensen不等式。然后，幻灯片说现在可以交换求和顺序（swap them），也就是把log和Σ_i的位置交换？不过这里可能需要更仔细推导。正确的步骤应该是：对于凹函数f，Σ q_j f(a_j/q_j) ⩾ f(Σ q_j a_j)，其中a_j = P(x^(i), z^(i)=j; Θ)/q_j？或者这里的Σ_i log(Σ_j (q_j P(x^(i), z^(i)=j)/q_j))可以看作是对每个i，里面是log(Σ_j (q_j * [P(x^(i), z^(i)=j)/q_j]))，也就是log(Σ_j (q_j * r_j))，其中r_j = P(x^(i), z^(i)=j)/q_j。因为log是凹函数，所以Σ_j q_j log(r_j) ⩾ log(Σ_j q_j r_j)，也就是Σ_j q_j log(r_j) ⩾ Σ_j q_j r_j = log(Σ_j q_j r_j)。所以Σ_i log(Σ_j q_j r_j) ⩾ Σ_i Σ_j q_j log(r_j)。这里的r_j就是P(x^(i), z^(i)=j)/q_j，所以代入的话，就是ℓ(X, Θ) ⩾ Σ_i Σ_j q_j log(P(x^(i), z^(i)=j)/q_j)，这就是Jensen不等式的应用结果，得到一个下界。

第一张幻灯片最后提到了Jensen不等式的公式，针对凹函数f，有f(E[X]) ⩾ E[f(X)]，这里E是期望，对于离散变量就是求和，所以E[f(X)]就是Σ q_j f(x_j)，f(E[X])就是f(Σ q_j x_j)，所以Σ q_j f(x_j) ⩽ f(Σ q_j x_j)，也就是Σ q_j f(x_j) ⩾ f(Σ q_j x_j)当f是凹函数时，这里x_j = log(r_j)，所以Σ q_j log(r_j) ⩾ log(Σ q_j r_j)，即log(Σ q_j r_j) ⩽ Σ q_j log(r_j)，所以原式的log(Σ q_j r_j)的下界就是Σ q_j log(r_j)，因此ℓ(X, Θ) ⩾ Σ_i Σ_j q_j log(P(x^(i), z^(i)=j)/q_j)。这就是EM算法中的下界函数。

然后第二张幻灯片标题是"Where does EM come from? II"，开头说应用Jensen不等式得到不等式：Σ_i log(Σ_j q_j P(x^(i), z^(i)=j)/q_j) ⩾ Σ_i Σ_j q_j log(P(x^(i), z^(i)=j)/q_j)。这里应该是笔误，左边其实是ℓ(X, Θ)的表达式，右边是Jensen不等式得到的下界，所以正确的应该是ℓ(X, Θ) ⩾ Σ_i Σ_j q_j log(P(x^(i), z^(i)=j)/q_j)。

然后幻灯片说，最大化这个下界将迫使我们的似然函数增加。因为下界是ℓ(X, Θ)的一个下界，如果我们最大化下界，那么下界增大，由于ℓ(X, Θ)⩾下界，所以ℓ(X, Θ)也会增大。不过这里需要注意，EM算法是交替最大化，先固定参数Θ，求最优的q（E步），再固定q，求最优的Θ（M步）。

接下来，幻灯片问如何选择q_j使得这个下界很好？也就是如何选择q使得下界尽可能接近上界ℓ(X, Θ)，即下界是紧的（tight）。在EM算法中，通常选择q使得q_j = P(z^(i)=j; Θ)，也就是在E步，计算后验概率作为q。这时候，因为Σ_j q_j = 1，且q_j = P(z^(i)=j|X, Θ)，代入下界公式，Σ_j q_j log(P(x^(i), z^(i)=j)/q_j) = Σ_j P(z^(i)=j|X, Θ) log(P(x^(i), z^(i)=j)/P(z^(i)=j|X, Θ)) = Σ_j P(z^(i)=j|X, Θ) log(P(x^(i)|z^(i)=j; Θ))，这时候下界就变成了Σ_i Σ_j P(z^(i)=j|X, Θ) log(P(x^(i), z^(i)=j; Θ)) - Σ_i Σ_j P(z^(i)=j|X, Θ) log(P(z^(i)=j|X, Θ))，而后一项在固定Θ时是常数，所以优化下界等价于优化Σ_i Σ_j P(z^(i)=j|X, Θ) log(P(x^(i)|z^(i)=j; Θ))，这就是E步的目标。然后在M步，固定q（即固定后验概率），最大化这个下界，就可以得到新的参数Θ，这样交替进行，似然函数单调递增。

现在需要把这些内容组织起来，解释清楚每一步。首先，第一张幻灯片解释了为什么直接优化log里面有和的似然函数困难，然后引入分布q，应用Jensen不等式得到下界。第二张幻灯片说明最大化下界可以提升似然，并且引出如何选择q的问题，为E步和M步做铺垫。

可能用户需要更详细的数学推导步骤，比如Jensen不等式的具体应用，以及q的选择为什么是关键。另外，要区分EM算法中的E步（求q，即后验概率）和M步（求Θ，最大化下界）。

总结一下，第一张幻灯片的核心是：直接优化log(Σ...)困难，引入q，应用Jensen不等式得到下界。第二张幻灯片的重点是：最大化下界使似然增加，如何选择q使得下界有效（即E步求后验概率）。</think>

<answer> 这两张幻灯片详细阐述了期望最大化（EM）算法的起源，核心在于如何通过**Jensen不等式**解决含潜在变量模型中似然函数优化的难题。以下从数学推导和算法思想两方面详细解释：

第一张幻灯片：EM的动机与Jensen不等式的引入 标题"Where does EM come from? I"（EM从何而来？I）聚焦于**为什么直接优化似然函数困难**，以及如何通过引入辅助分布和Jensen不等式建立下界。

1. 优化似然函数的困难：log内的求和 在含潜在变量（如混合高斯模型中的聚类标签）的概率模型中，似然函数通常形式为：ℓ(X, Θ) = Σ_i log P(x^(i); Θ) 其中，X = {x^(1), ..., x^(n)}是观测数据，Θ是模型参数（如均值、方差、混合系数），P(x^(i); Θ)是观测数据x^(i)的似然，通常依赖于潜在变量z^(i)（如聚类标签），可展开为：P(x^(i); Θ) = Σ_j P(x^(i), z^(i)=j; Θ)（对潜在变量z^(i)求和） 因此，似然函数变为：ℓ(X, Θ) = Σ_i log(Σ_j P(x^(i), z^(i)=j; Θ)) **核心困难**：log函数内部有求和项Σ_j P(x^(i), z^(i)=j; Θ)，直接对Θ求导优化会涉及复杂的复合函数求导，难以处理。

2. 引入辅助分布q，构造下界 为解决上述困难，EM算法引入一个**辅助分布q**（关于潜在变量z的分布），将log与求和的顺序"交换"。具体步骤：

假设对每个观测x^(i)，存在潜在变量z^(i)的辅助分布q(z^(i)=j)（满足Σ_j q_j^(i)=1），则可将原式改写为：ℓ(X, Θ) = Σ_i log(Σ_j q_j^(i) · P(x^(i), z^(i)=j; Θ)/q_j^(i)) 这里，q_j^(i)是辅助分布中z^(i)=j的概率，目的是将"log内求和"转化为"求和后log"的形式，以便应用Jensen不等式。

3. 应用Jensen不等式：建立下界 Jensen不等式指出：**对凹函数f，有f(E[X]) ⩾ E[f(X)]**（log函数是凹函数，满足此性质）。

将上式中的求和视为"期望"：令X_j = log(P(x^(i), z^(i)=j; Θ)/q_j^(i))，权重为q_j^(i)（即E[X] = Σ_j q_j^(i) X_j），则：log(Σ_j q_j^(i) · P(x^(i), z^(i)=j; Θ)/q_j^(i)) ⩽ Σ_j q_j^(i) log(P(x^(i), z^(i)=j; Θ)/q_j^(i)) 两边对同一个观测i求和，得到似然函数的下界：ℓ(X, Θ) ⩾ Σ_i Σ_j q_j^(i) log(P(x^(i), z^(i)=j; Θ)/q_j^(i))（下界函数L(Θ, q)）

第二张幻灯片：最大化下界与q的选择 标题"Where does EM come from? II"（EM从何而来？II）聚焦于**如何利用下界提升似然函数**，以及**如何选择辅助分布q**。

1. 最大化下界：迫使似然函数增加 下界L(Θ, q)是ℓ(X, Θ)的**紧下界**（即L(Θ, q) ⩽ ℓ(X, Θ)）。若最大化L(Θ, q)，则下界增大，由于L(Θ, q) ⩽ ℓ(X, Θ)，必然有ℓ(X, Θ)增大。

但L(Θ, q)同时依赖于参数Θ和辅助分布q，因此需交替优化：- **E步（Expectation）**：固定Θ，选择q使下界最大；- **M步（Maximization）**：固定q，选择Θ使下界最大。

2. 如何选择q：最优辅助分布 为使下界L(Θ, q)尽可能接近上界ℓ(X, Θ)（即下界"紧"），EM算法选择q为**潜在变量的后验概率**：q_j^(i) = P(z^(i)=j | x^(i), Θ) = P(x^(i), z^(i)=j; Θ)/Σ_k P(x^(i), z^(i)=k; Θ)

此时，代入下界公式，L(Θ, q)可简化为：L(Θ, q) = Σ_i Σ_j P(z^(i)=j | x^(i), Θ) log P(x^(i), z^(i)=j; Θ) 这一项仅依赖于Θ，在E步中固定Θ时，q_j^(i)可直接计算（即E步的核心是"计算后验概率"）。

总结：EM算法的核心思想 EM算法通过以下两步交替迭代，逐步提升似然函数：1. **E步**：固定参数Θ^(t)，计算后验概率q_j^(i) = P(z^(i)=j | x^(i), Θ^(t))（即选择最优辅助分布q）；2. **M步**：固定q，最大化下界L(Θ, q)，得到新参数Θ^(t+1)。

通过Jensen不等式建立的下界保证了每次迭代后似然函数ℓ(X, Θ)单调递增，最终收敛到局部最优解。

Table 14: Algorithm Theory Analysis and Mathematical Derivation Explanation

This example showcases Qianfan-VL's deep understanding of machine learning algorithms and mathematical concepts:

- **Model Capabilities:** Algorithm comprehension, mathematical formula interpretation, multi-slide correlation, and theoretical explanation with Chain-of-Thought reasoning
- **Technical Challenges:** Understanding EM algorithm mathematics, connecting concepts across multiple slides, explaining complex iterative processes, and providing intuitive interpretations
- **Performance Highlights:** Comprehensive explanation of E-step and M-step, clear mathematical derivations, and practical insights into convergence properties and applications

A.3 Business Intelligence

Political Polling Analysis:

This case demonstrates Qianfan-VL's advanced data visualization interpretation and trend analysis capabilities:

- **Model Capabilities:** Time-series data interpretation, multi-variable trend analysis, contextual understanding of political events, and comprehensive data storytelling
- **Technical Challenges:** Identifying multiple data series, correlating trends with historical events, extracting precise percentage values, and providing meaningful interpretations
- **Performance Highlights:** Accurate identification of all political parties, precise trend description with timeline markers, and insightful analysis of electoral implications

This example demonstrates Qianfan-VL's sophisticated capability to analyze complex multi-party political polling charts:

Image:	Prompt: 帮我详细解释这张图表的含义。 Response: 这张图表展示了英国主要政党的选举投票意向变化趋势，数据来源于经济学人杂志的英国民意调查追踪器。图表标题为"困境中的保守党"（Tories in trouble），副标题说明这是英国大选投票意向百分比。 图表主要内容包括： 1. 两条主要曲线：- 红色曲线代表工党（Labour）支持率 - 蓝色曲线代表保守党（Conservatives）支持率 2. 下方还有四条较小的曲线，分别代表：- 黄色：自由民主党（Lib Dems）- 青色：绿党（Green）- 灰色：改革英国党（Reform UK）- 浅灰色：其他政党（Other） 3. 横轴表示时间，从2020年到2023年 4. 纵轴表示支持率百分比，从0%到60% 从图表可以观察到：- 保守党（蓝色）在2020年初支持率较高（约45-50%），但呈现明显下降趋势，到2023年降至约25%左右 - 工党（红色）在2020年时支持率较低（约30-35%），但整体呈上升趋势，到2023年上升至约45%左右 - 两党支持率大约在2021-2022年间出现交叉，之后工党保持领先 - 其他小党的支持率相对稳定，基本维持在较低水平 总体趋势表明，保守党的支持率在下降，而工党的支持率在上升，标题"困境中的保守党"也反映了这一政治动态。

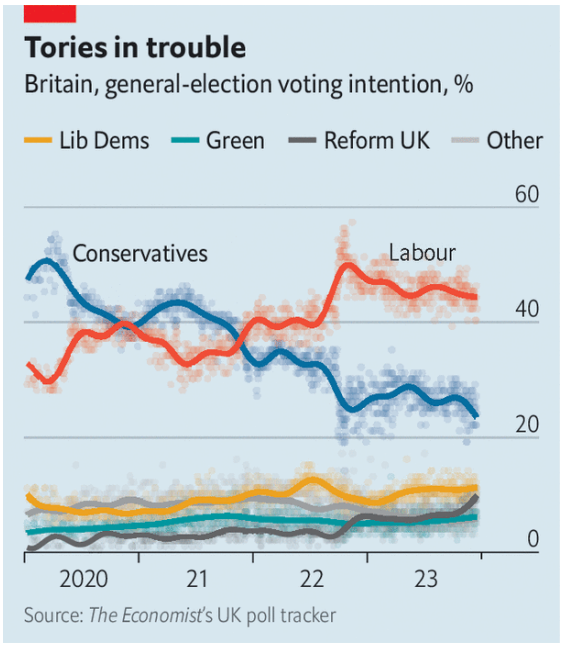


Table 15: Political Polling Data Analysis and Trend Interpretation

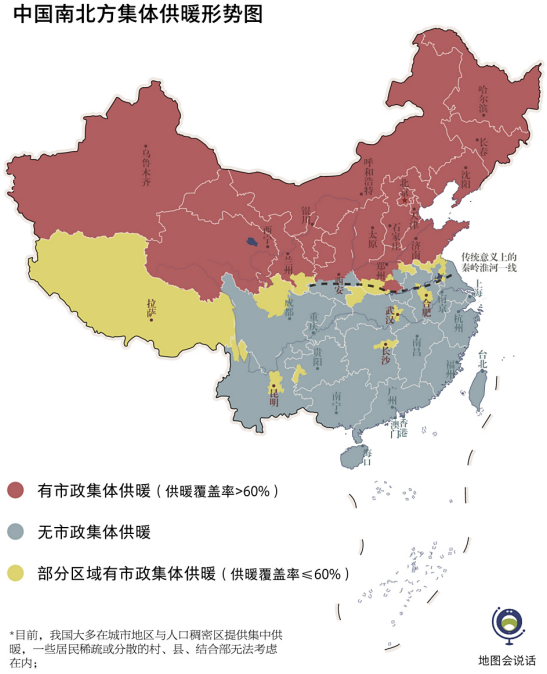
- **Multi-trend analysis:** Simultaneous tracking and interpretation of multiple political parties' polling trends over time
- **Cross-referencing data sources:** Recognition and attribution of data sources (The Economist's UK poll tracker)
- **Temporal pattern recognition:** Identification of trend reversals, crossover points, and long-term trajectories across different time periods
- **Contextual interpretation:** Understanding of political implications and providing meaningful analysis beyond raw data extraction

Geographical Data Analysis:

This example demonstrates Qianfan-VL's capability to interpret complex geographical and meteorological visualizations:

- **Model Capabilities:** Heat map interpretation, geographical feature recognition, temperature scale understanding, and regional climate analysis
- **Technical Challenges:** Color gradient mapping to temperature values, identifying geographical boundaries, understanding seasonal patterns, and providing location-specific insights
- **Performance Highlights:** Accurate temperature range identification, detailed regional analysis, and comprehensive explanation of climate distribution patterns

Image:



Prompt:
云南有任何区域有集体供暖吗？<think>

Response:
用户问的是"云南有任何区域有集体供暖吗？"首先，我需要查看提供的地图信息。地图的标题是"中国南北方集体供暖形势图"，图例分为三种：红色是有市政集体供暖（覆盖率¿60%），灰色是无市政集体供暖，黄色是部分区域有市政集体供暖（覆盖率≤60%）。
首先，找到云南省的位置。云南省位于中国西南，地图上昆明是云南省的省会。在地图上，昆明所在的区域颜色需要确认。根据图例，黄色代表部分区域有市政集体供暖（覆盖率≤60%）。查看地图，昆明被标记为黄色，而云南省的其他区域可能大部分是灰色（无市政集体供暖），但问题问的是"任何区域"，只要有一个区域就算。
另外，地图下方的说明提到"目前，我国大多在城市地区与人口稠密区提供集中供暖"，昆明作为云南省的省会，属于较大城市，可能属于部分区域有集体供暖的情况，因为昆明的颜色是黄色。所以结论应该是云南的部分区域（如昆明）有市政集体供暖，属于黄色区域，即覆盖率≤60%。需要确认地图上昆明是否确实是黄色，从描述中"昆明"被标记为黄色，所以答案应该是肯定的，云南有部分区域（如昆明）有市政集体供暖。</think>
<answer>根据"中国南北方集体供暖形势图"及图例信息：
云南省的省会昆明所在区域在地图上被标记为黄色，对应图例中"部分区域有市政集体供暖（供暖覆盖率≤60%）"。
因此，云南的部分区域（如昆明）存在市政集体供暖，属于"部分区域有集体供暖"的范畴。
答案：云南的部分区域（如昆明）有市政集体供暖（供暖覆盖率≤**60%**）。</answer>

Table 16: Geographical Map Analysis and Regional Information Extraction

This example showcases Qianfan-VL's capability to analyze geographical maps and extract regional information:

- **Map legend interpretation:** Accurate understanding of color-coded legends and their corresponding meanings in geographical context
- **Spatial reasoning:** Identification of specific regions (Yunnan province, Kunming city) and their geographical relationships within the map
- **Data cross-referencing:** Matching visual map elements with legend information to provide accurate regional analysis
- **Localized information extraction:** Providing specific, targeted answers about regional characteristics based on map data

Advertisement Analysis:

This case showcases Qianfan-VL's understanding of marketing materials and brand communication strategies:

- **Model Capabilities:** Advertisement layout analysis, brand messaging interpretation, visual hierarchy understanding, and marketing strategy identification
- **Technical Challenges:** Analyzing multi-element compositions, understanding cultural context, identifying target audience messaging, and interpreting visual metaphors
- **Performance Highlights:** Complete identification of all visual elements, understanding of brand positioning, and insightful analysis of marketing effectiveness

Image:	Prompt: 请描述这张广告牌上的主要内容，并且分析一下这句广告词的含义。 Response: 广告牌上是可口可乐的广告，主色调为红色，上面有一个简约风格的女性头像剪影，广告词是'THEY DON'T MAKE 'EM LIKE THEY USED TO. WE DO.'（他们不再制造像过去那样的产品了。但我们仍然制造。） 这句广告词的含义是强调可口可乐始终保持着其原始配方和经典口味，暗示其他饮料品牌可能已经改变了原有配方或品质，但可口可乐依然坚持传统，保持着最初的味道和品质。广告通过怀旧情感唤起消费者对经典产品的认同和信任。

Table 17: Advertisement Content Analysis and Marketing Message Interpretation

This example demonstrates Qianfan-VL's sophisticated capability to analyze advertising content and interpret marketing strategies:

- **Visual brand recognition:** Accurate identification of brand elements, color schemes, and design aesthetics in complex urban advertising displays
- **Text extraction and translation:** Precise reading of advertising copy and providing contextual translation for cross-linguistic understanding
- **Marketing strategy analysis:** Deep interpretation of advertising messages, understanding the psychological appeal and competitive positioning
- **Cultural and emotional context:** Recognition of nostalgic marketing approaches and their intended impact on consumer sentiment and brand loyalty

Media Information Extraction:

This example illustrates Qianfan-VL's ability to extract and analyze information from media promotional materials:

- **Model Capabilities:** Media content understanding, celebrity recognition, program information extraction, and entertainment industry knowledge
- **Technical Challenges:** Processing stylized text layouts, understanding program formats, extracting scheduling information, and identifying cast members
- **Performance Highlights:** Accurate extraction of all program details, understanding of show format and genre, and comprehensive cast identification

Image:

Prompt:
请抽取这张电影海报中所有的关键信息。

Response:
这张电影海报是王家卫导演的《花样年华》(In the Mood for Love)的宣传海报。以下是我从海报中抽取的关键信息：
1. 电影名称： - 英文：In the Mood for Love - 中文：花样年华
2. 导演与主演： - 导演：王家卫 - 主演：张曼玉、梁朝伟
3. 上映信息： - 上映日期：2月14日 - 标语："好久不见"
4. 版本信息： - "25周年导演特别版4K"
5. 技术规格： - IMAX - CINITY - CINITY LED - 中国巨幕 - Dolby Atmos
6. 电影节荣誉： - "CANNES 2000"（戛纳2000） - "最佳男主角及最佳艺术成就奖"（The best actor award & Grand Prix technique） - "世界影史百大最佳电影"（《时代》杂志评选）
7. 其他文字信息： - "何止98分钟" - "史上最长版本全球首映"
海报整体以红色为主色调，展现了一个背对观众的人物形象，站在走廊中，符合王家卫电影特有的美学风格和氛围。

Table 18: Movie Poster Information Extraction and Content Analysis

This example showcases Qianfan-VL's comprehensive capability to extract and analyze multimedia content information:

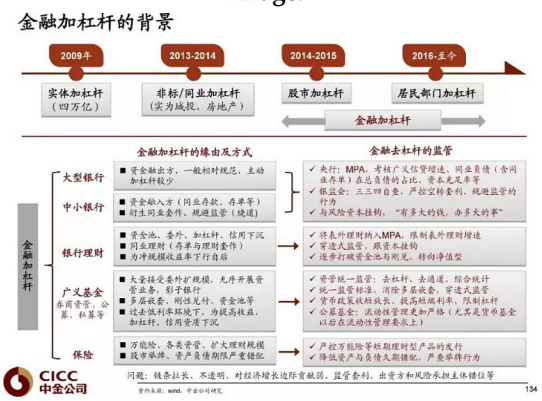
- **Multilingual text recognition:** Accurate extraction of multilingual text from complex poster layouts with varying fonts and sizes
- **Structured information organization:** Systematic categorization of extracted information including titles, credits, release details, and technical specifications
- **Cultural and artistic analysis:** Recognition of directorial style, aesthetic elements, and cultural significance of the film poster
- **Complete metadata extraction:** Comprehensive identification of all relevant information including awards, technical formats, and promotional elements

Financial Policy Analysis:

This case demonstrates Qianfan-VL's ability to understand and analyze complex financial policy presentations:

- **Model Capabilities:** Financial terminology understanding, policy interpretation, structured information extraction, and professional document analysis
- **Technical Challenges:** Processing dense financial content, understanding regulatory language, extracting key policy points, and maintaining logical structure
- **Performance Highlights:** Accurate extraction of all policy details, clear categorization of financial measures, and comprehensive understanding of implementation guidelines

Image:



Prompt:
请分条回答：金融去杠杆的监管是如何应对广义基金的金融加杠杆的?

Response:
1. 资管统一监管：去杠杆、去通道、综合统计
2. 统一监管标准：消除多层嵌套，穿透式监管
3. 货币政策收紧：提高市场利率，限制杠杆
4. 公募基金：流动性管理更加严格（尤其是货币基金以后在流动性管理要求上）

Table 19: Financial Policy and Regulatory Analysis

This example demonstrates Qianfan-VL's expertise in analyzing complex financial and regulatory policy documents:

- **Policy comprehension:** Deep understanding of financial regulatory frameworks including deleveraging mechanisms and asset management oversight
- **Structured policy analysis:** Systematic breakdown of regulatory approaches with clear categorization of different policy instruments
- **Financial terminology mastery:** Accurate usage and explanation of specialized financial terms and asset management regulations including penetrating supervision mechanisms
- **Regulatory impact assessment:** Understanding of how different policy measures address specific financial risks and leverage issues in the fund industry

A.4 Structured Information Extraction

Table Structure Analysis and Data Extraction:

This sophisticated example demonstrates Qianfan-VL's advanced capability to handle complex multi-table extraction tasks:

- **Model Capabilities:** Complex table structure recognition, multi-table relationship understanding, structured JSON generation, and data validation
- **Technical Challenges:** Processing multiple interconnected tables, maintaining data relationships across different measurement systems, handling dense numerical data, and ensuring output accuracy
- **Performance Highlights:** Perfect extraction of all size conversions, accurate mapping between different sizing systems, and well-formatted JSON output with complete data integrity

Image:

类型		贵人鸟鞋子尺码参考表											
男鞋&女鞋	美码	2.5	3.5	4.5	5	5.5	6.5	7	8	8.5	9.5	10	11
	英码	2	3	4	4.5	5	6	6.5	7.5	8	9	9.5	10.5
	法码	35	36	37	38	39	40	41	42	43	44	45	46
	中国码	220	230	235	240	245	250	255	260	265	270	280	285

温馨提示：皮鞋比运动小一码，如39的皮鞋，运动鞋选40

类别		贵人鸟服装尺码参考表									
男	上衣	尺码	XS	S	M	L	XL	2XL	3XL	4XL	5XL
		身高 胸围	160 /80A	165 /84A	170 /88A	175 /92A	180 /96A	185 /100A	190 /104A	195 /108A	200 /112A
	裤装	尺码	XS	S	M	L	XL	2XL	3XL	4XL	5XL
		身高 腰围	160 /68A	165 /72A	170 /76A	175 /80A	180 /84A	185 /88A	190 /92A	195 /96A	200 /100A
女	上衣	尺码	XS	S	M	L	XL	2XL	3XL	4XL	5XL
		身高 胸围	155 /80A	160 /84A	165 /88A	170 /92A	175 /96A	180 /100A	185 /104A	190 /108A	195 /112A
	裤装	尺码	XS	S	M	L	XL	2XL	3XL	4XL	5XL
		身高 腰围	115 /64A	160 /68A	165 /72A	170 /76A	175 /80A	180 /84A	185 /88A	190 /92A	195 /96A

温馨提示：据贵人鸟公司战略规划，自2018年春季产品开始，所有贵人鸟服装产品尺码调整（服装与配件未调整），原贵人鸟的S码是现在的XS码，原M码是现在的S码，原L码是现在的M码，原XL码是现在的L码，原2XL码是现在的XL码，原3XL码是现在的2XL码，原4XL码是现在的3XL码，原5XL码是现在的4XL码，原6XL码是现在的5XL码，取消原6XL码，具体尺码可参考上方尺码表

品牌形象 BRAND

THE BRAND IMAGE

关于仓库

Prompt:
请从图片中的"品牌形象"以上的图片部分，抽取以下关键信息：鞋子尺码对照表（包括美码、英码、法码、中国码的对应关系），服装尺码对照表（包括男女上衣和裤装的尺码、身高、胸围、腰围数据）。以JSON格式输出并且只输出JSON信息，不要输出其他内容，确保信息的准确性和完整性。

Response:


```
{
  "鞋子尺码对照表": {
    "美码": ["2.5", "3.5", "4.5", "5", "5.5", "6.5", "7", "8",
             "8.5", "9.5", "10", "11"],
    "英码": ["2", "3", "4", "4.5", "5", "6", "6.5", "7.5", "8",
             "9", "9.5", "10.5"],
    "法码": ["35", "36", "37", "38", "39", "40", "41", "42",
             "43", "44", "45", "46"],
    "中国码": ["220", "230", "235", "240", "245", "250", "255",
              "260", "265", "270", "280", "285"]
  },
  "服装尺码对照表": {
    "男上衣": {
      "尺码": ["XS", "S", "M", "L", "XL", "2XL", "3XL", "4XL", "5XL"],
      "身高/胸围": ["160/80A", "165/84A", "170/88A", "175/92A",
                  "180/96A", "185/100A", "190/104A", "195/108A",
                  "200/112A"]
    },
    "男裤装": {
      "尺码": ["XS", "S", "M", "L", "XL", "2XL", "3XL", "4XL", "5XL"],
      "身高/腰围": ["160/68A", "165/72A", "170/76A", "175/80A",
                  "180/84A", "185/88A", "190/92A", "195/96A",
                  "200/100A"]
    },
    "女上衣": {
      "尺码": ["XS", "S", "M", "L", "XL", "2XL", "3XL", "4XL", "5XL"],
      "身高/胸围": ["155/80A", "160/84A", "165/88A", "170/92A",
                  "175/96A", "180/100A", "185/104A", "190/108A",
                  "195/112A"]
    },
    "女裤装": {
      "尺码": ["XS", "S", "M", "L", "XL", "2XL", "3XL", "4XL", "5XL"],
      "身高/腰围": ["115/64A", "160/68A", "165/72A", "170/76A",
                  "175/80A", "180/84A", "185/88A", "190/92A",
                  "195/96A"]
    }
  }
}
```


Table 20: Structured Information Extraction from Complex Tables

This example demonstrates Qianfan-VL's capability to extract structured information from complex multi-table layouts:

- **Multi-table processing:** Accurately identifies and processes multiple interconnected tables within a single image
- **Cross-reference understanding:** Maintains relationships between different sizing systems and measurements
- **Structured output generation:** Produces well-formatted JSON output with proper hierarchical organization
- **Data integrity preservation:** Ensures accuracy of numerical data and dimensional correspondences across different measurement systems

A.5 Sports Schedule Analysis

Event Calendar Information Extraction:

This final example showcases Qianfan-VL's ability to extract and analyze complex scheduling information from sports calendars:

- **Model Capabilities:** Calendar layout understanding, event scheduling interpretation, multi-competition tracking, and temporal relationship analysis
- **Technical Challenges:** Processing dense calendar layouts, distinguishing between different competition types, extracting dates and venues, and understanding tournament structures
- **Performance Highlights:** Complete extraction of all match schedules, accurate identification of

competitions and venues, and clear temporal organization of events

Image:

Prompt:
巴塞罗那会在什么时候和德甲球队进行比赛?

Response:
10月23日（对阵拜仁慕尼黑）和12月11日（对阵多特蒙德）。

Table 21: Sports Schedule Analysis and Information Extraction

This example showcases Qianfan-VL's ability to analyze sports calendars and extract specific scheduling information:

- **Calendar interpretation:** Accurately reads and interprets complex sports scheduling calendars with multiple teams and dates
- **Selective information extraction:** Identifies specific matches based on league criteria (German teams in this case)
- **Date and opponent recognition:** Precisely extracts match dates and opponent team names from visual calendar data
- **Context-aware filtering:** Understands the relationship between team leagues and provides relevant match information