Qwen2.5-VL Technical Report

Qwen Team, Alibaba Group

Feb 2025

Jaehyeon Son, Junhyun Kim

Georgia Tech

Presenter

- Jaehyeon Son
- ML PhD
- Interests: Embodied AI and Robotics

- Junhyun Kim
- MS Robotics
- Interests : Robot Manipulation and Embodied AI

Outline

- Timeline and Overview
- Model Architecture
- Pre-training and Post-training
- Experiments and Results
- Strengths
- Weaknesses and Limitations

Timeline

Qwen-VL

- Aug 2023
- Qwen-VL / Qwen-VL-Chat
- First attempt to integrate vision encoder on Qwen

Qwen2-VL

- Sep 2024
- 2B / 8B / 72B
- MRoPE

Qwen2.5-VL

- Feb 2025
- 3B / 7B / 32B / 72B
- Dynamic FPS sampling
- SwiGLU / RMSNorm
- Window Attention

Qwen2.5-Omni

- Mar 2025
- 7B
- Added audio modality
- TMRoPE
- Thinker-Talker

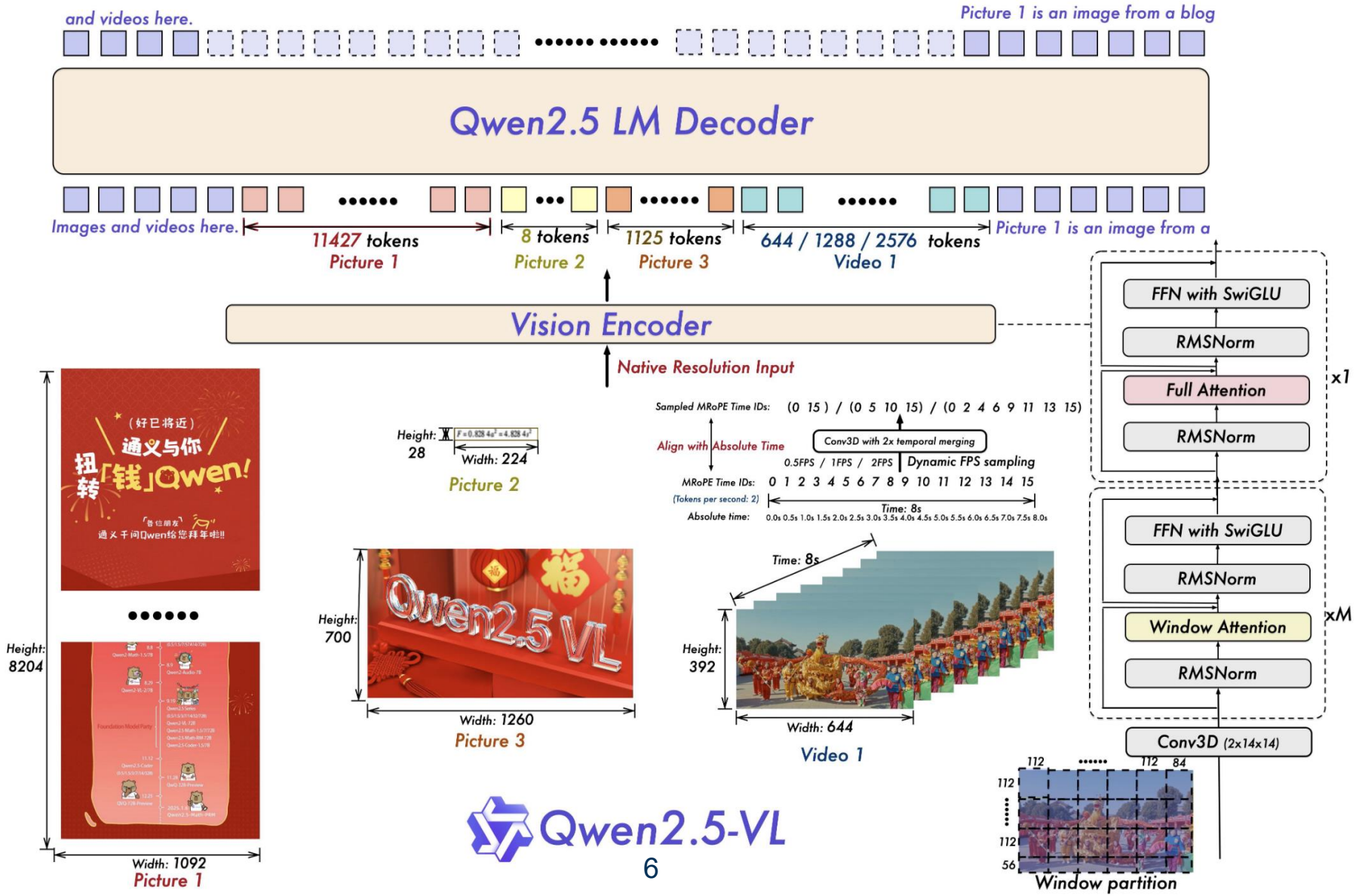
Qwen3-Omni

- **Sep 22**, 2025
- 30B
- Mixture-of-Experts(MoE)
- Audio Transformer(AuT)

Overview Qwen2.5-VL

- Developer: Qwen Team, Alibaba
- Release: Feb 2025
- Model variants: 3B / 7B / 32B / 72B
- License: Apache 2.0 (✓ commercial use) / Qwen (✗ commercial use)
- Key features
  - Vision-language model for text, image, *video*
  - Long-video understanding (up to hours)
  - Multilingual support (English, Chinese, and many other languages)

Model Architecture

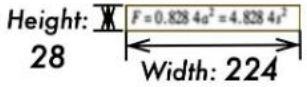


Qwen2.5-VL

Georgia Tech

Image Encoding

- Image resizing: multiples of 28
- Support for dynamic resolution
  - Vanilla ViT[1]: fixed resolution only



Picture 2

Picture 3

[1] Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2020

Georgia Tech

Image Encoding

- Patch size: 14 x 14
  - Even number of patches in both width and height
- Each patch is passed to Conv3D

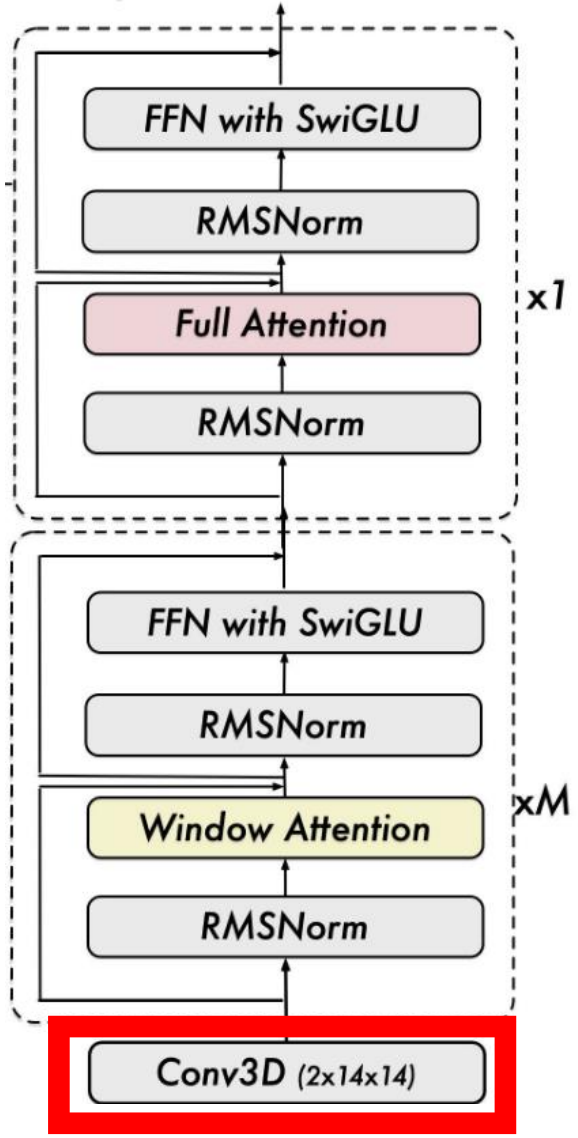
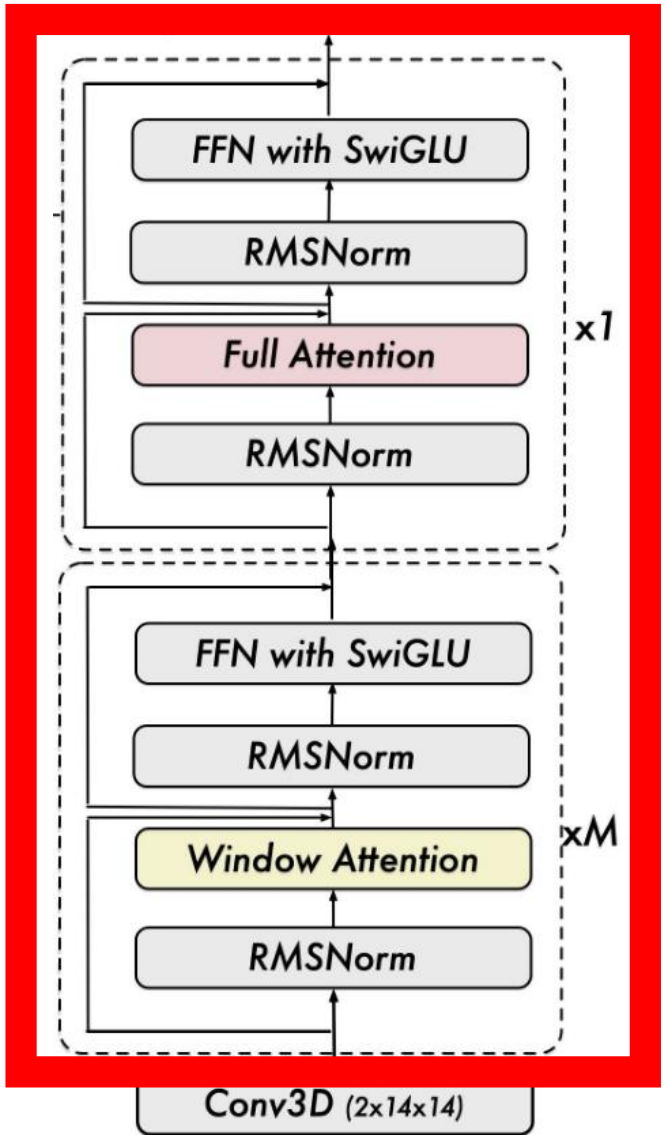


Image Encoding

- Vision Encoder: ViT[1]
- Several modifications are applied
- After ViT: Vision-Language Merger
  - Merge 4 adjacent patch features into one token
  - Two-layer MLP



[1] Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2020

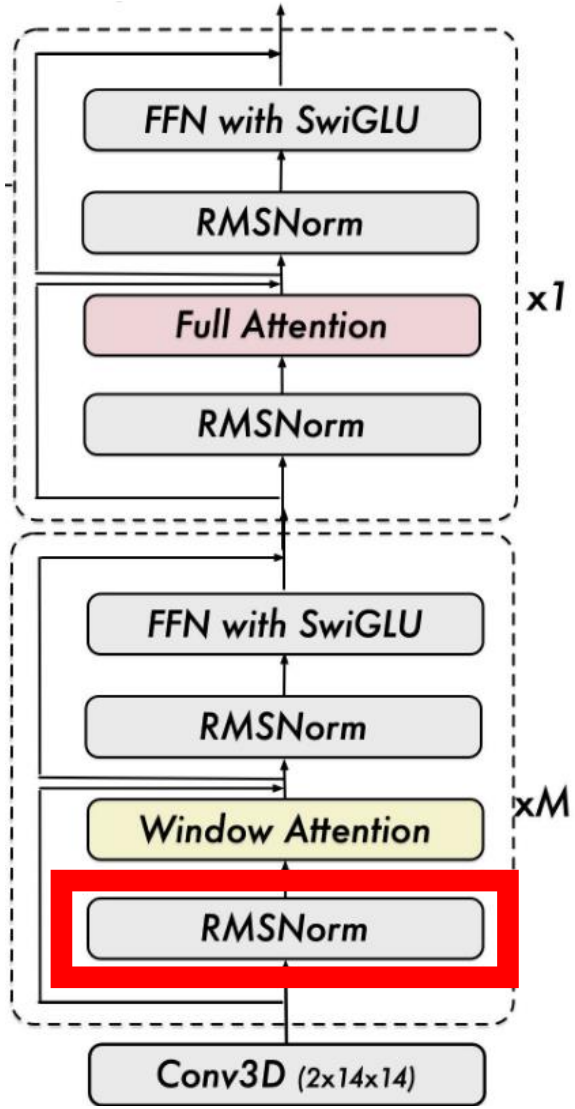
Image Encoding

- RMSNorm[1]
  - Cheap & competitive alternative to vanilla LayerNorm
  - Employed in modern Transformers
    - T5, LLaMA, Mistral, etc.

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sigma} \cdot \gamma + \beta$$

$$\text{RMSNorm}(x) = \frac{x}{\text{RMS}(x)} \cdot \gamma$$

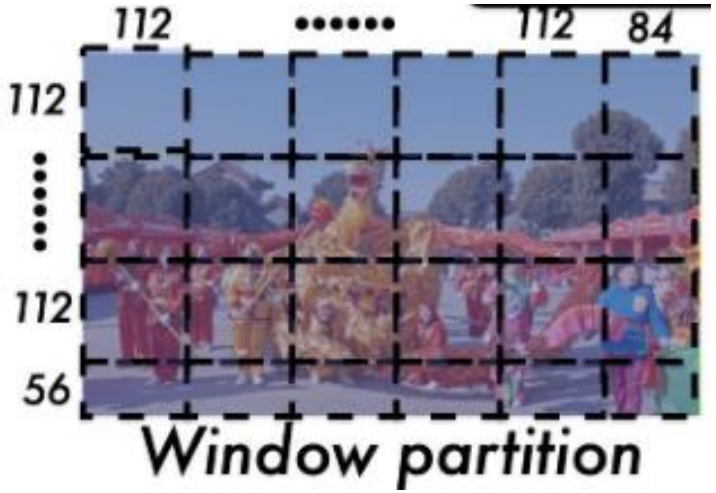
$$\text{RMS}(x) = \sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2 + \epsilon}$$



[1] Zhang & Sennrich, *Root Mean Square Layer Normalization*, 2019

Image Encoding

- Window Attention
  - Patch token attends only to adjacent 8×8 patches
    - 8x8 patches = 112x112 pixels
  - Linear complexity in the number of input tokens
- Full Attention is only used in 4 blocks



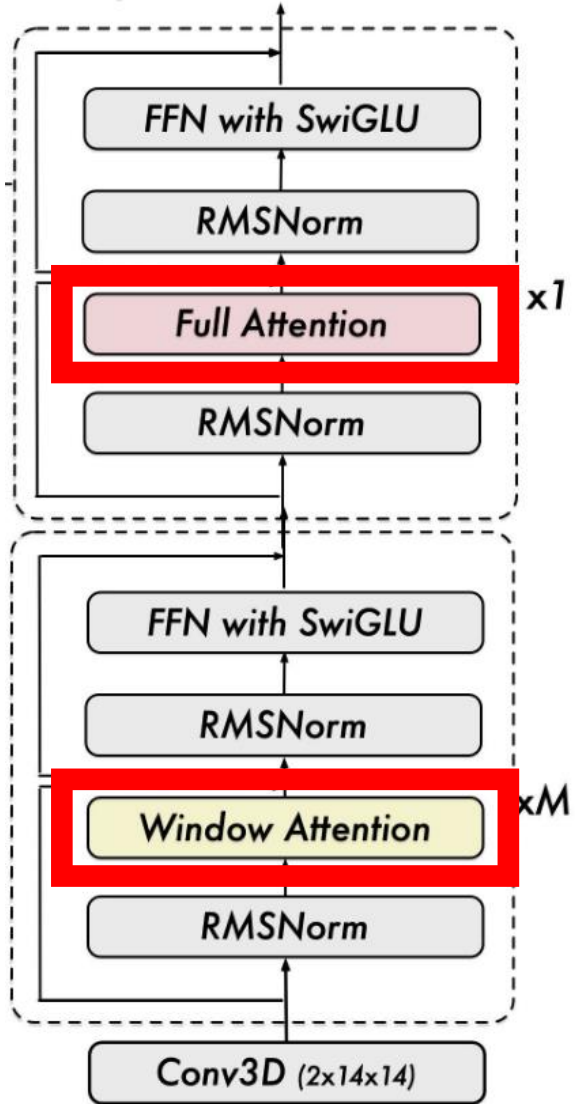


Image Encoding

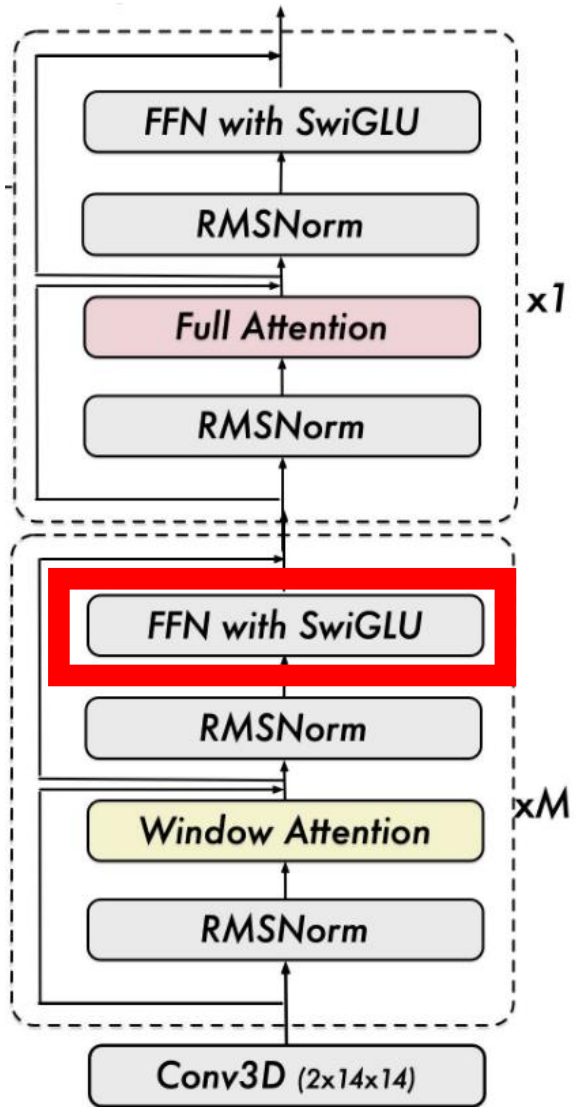
- FFN with SwiGLU[1]
  - Better expressivity than ReLU/GELU
  - Employed in modern Transformers
    - PaLM, LLaMA, Mistral, etc.

$$\mathrm{FFN}(x) = W_2\, f(W_1 x + b_1) + b_2$$

$$\mathrm{GLU}(x) = (W_1 x) \odot \sigma(W_2 x)$$

$$\mathrm{SwiGLU}(x) = (W_1 x) \odot \mathrm{Swish}(W_2 x)$$

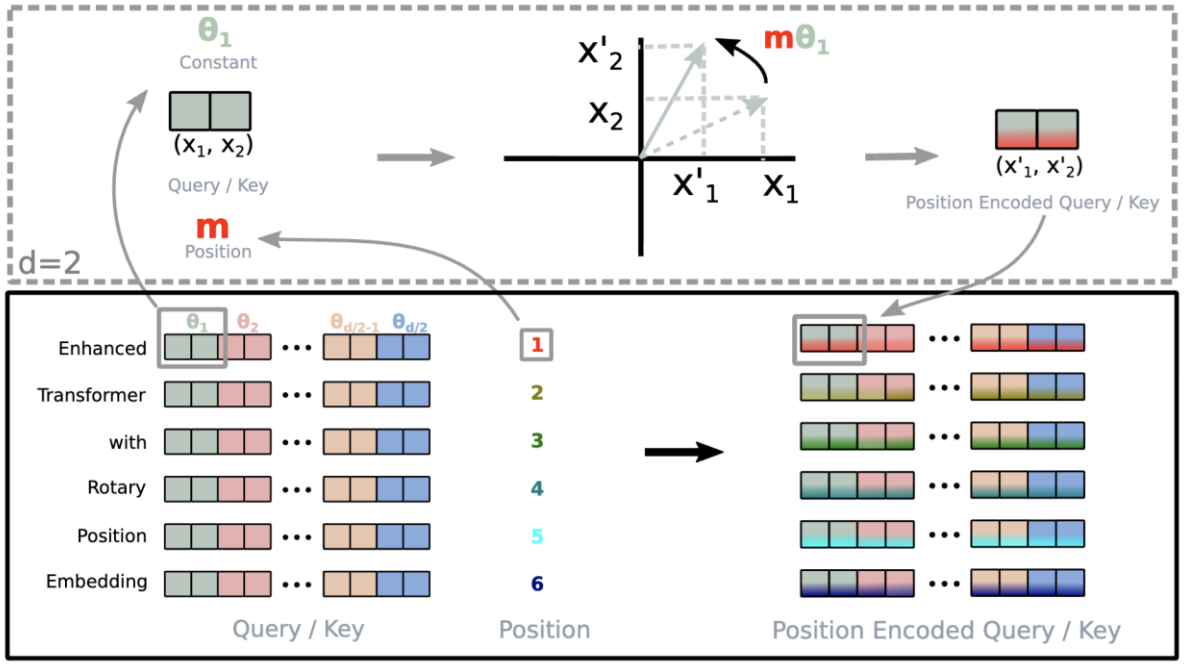
$$\mathrm{Swish}(z) = z \cdot \sigma(z)$$



[1] Shazeer et al., *GLU Variants Improve Transformer*, 2020

Rotary Positional Embedding (RoPE[1])

- Applies rotary transformations to Q/K
  - in *each* Attention layer
  - not to input embeddings
- Encodes relative distance directly in the dot product
- Preserves original token meaning through V



$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m}\boldsymbol{W}_{\{q,k\}}\boldsymbol{x}_m$$

where

$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

$$\Theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, ..., d/2]\}.$$

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = (\boldsymbol{R}^d_{\Theta,m}\boldsymbol{W}_q\boldsymbol{x}_m)^\intercal(\boldsymbol{R}^d_{\Theta,n}\boldsymbol{W}_k\boldsymbol{x}_n) = \boldsymbol{x}^\intercal\boldsymbol{W}_q R^d_{\Theta,n-m}\boldsymbol{W}_k\boldsymbol{x}_n$$

[1] Su et al., *RoFormer: Enhanced Transformer with Rotary Position Embedding, 2021*

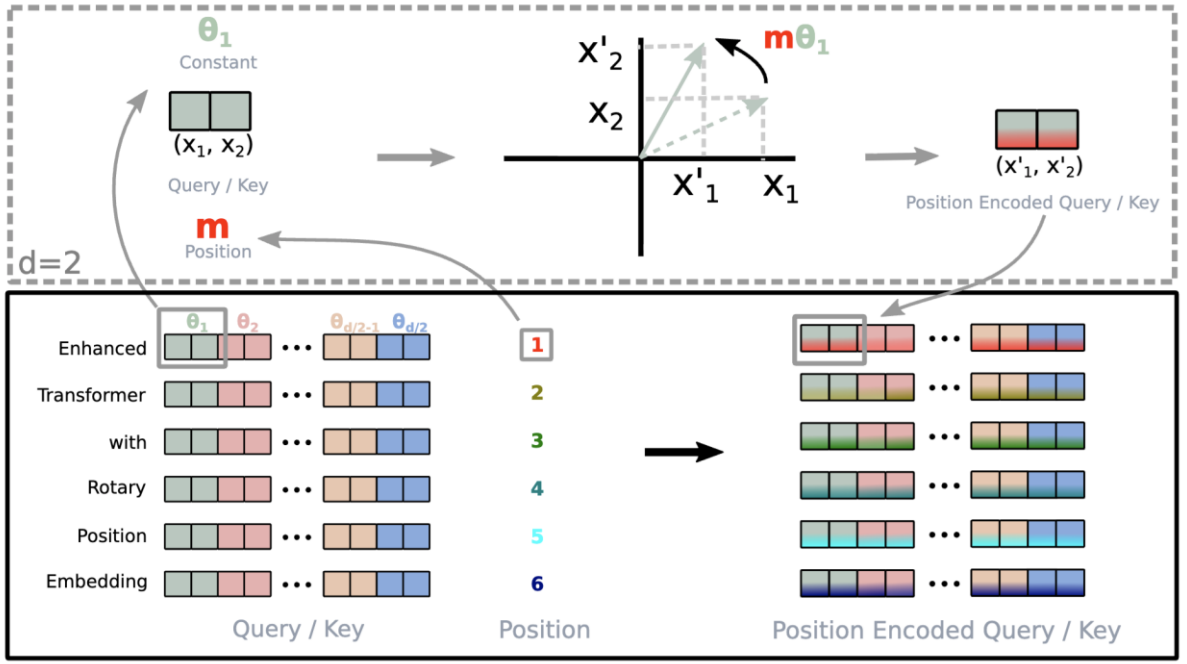
Georgia Tech

Rotary Positional Embedding (RoPE[1])

- Image patch: 2D-RoPE
  - Halve the feature dimension
  - Encode row and column indices separately

Patch K/V

Height position	Width position



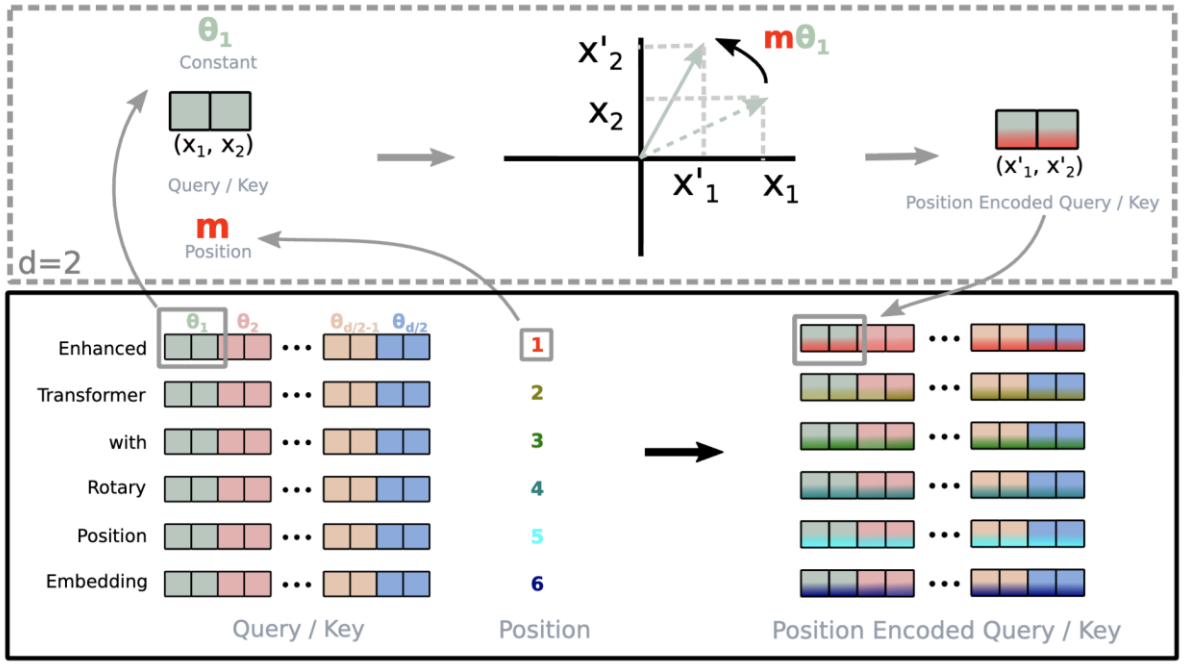
[1] Su et al., *RoFormer: Enhanced Transformer with Rotary Position Embedding, 2021*

Georgia Tech

Rotary Positional Embedding (RoPE[1])

- Image patch: 2D-RoPE
  - Halve the feature dimension
  - Encode row and column indices separately

Patch K/V	Height position	Width position



- Video patch: Multimodal RoPE (MRoPE)

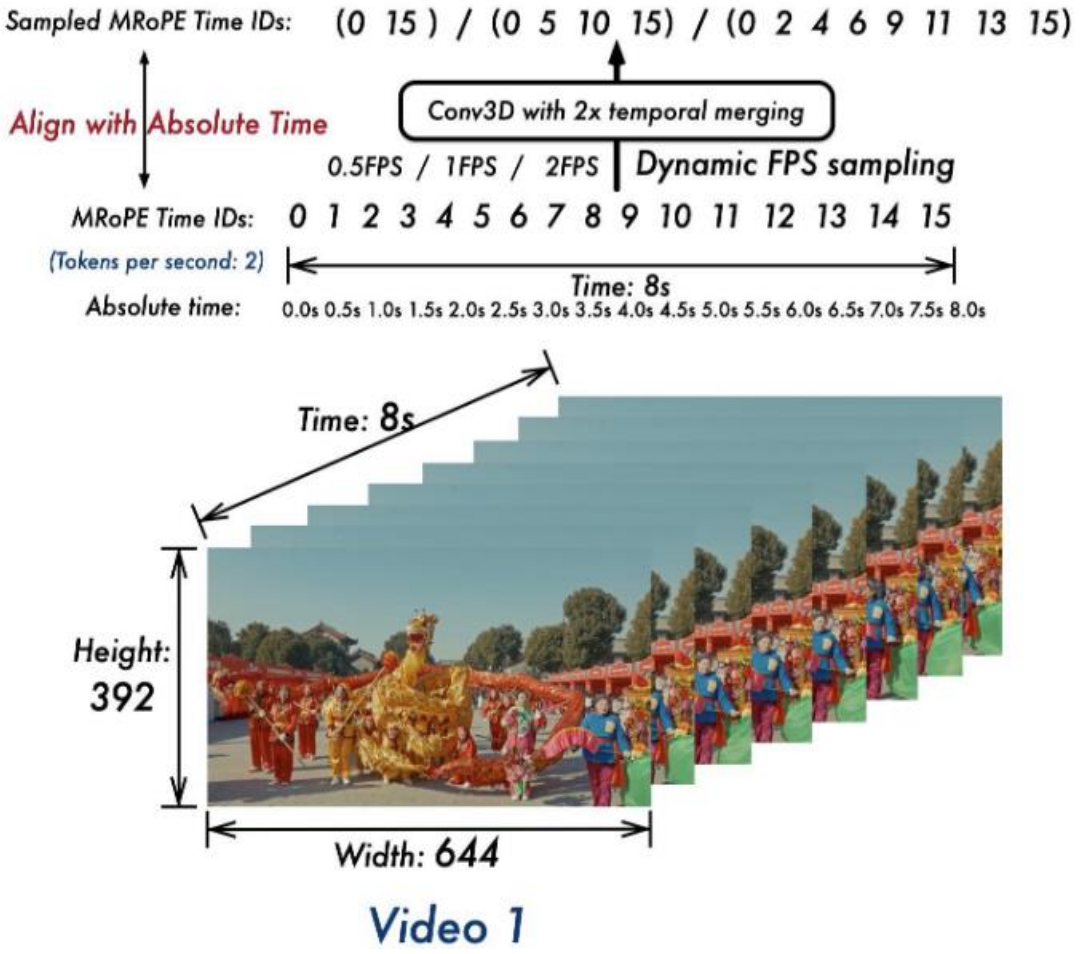
Patch K/V	Temporal position	Height position	Width Position

[1] Su et al., *RoFormer: Enhanced Transformer with Rotary Position Embedding, 2021*

Georgia Tech

Video Encoding

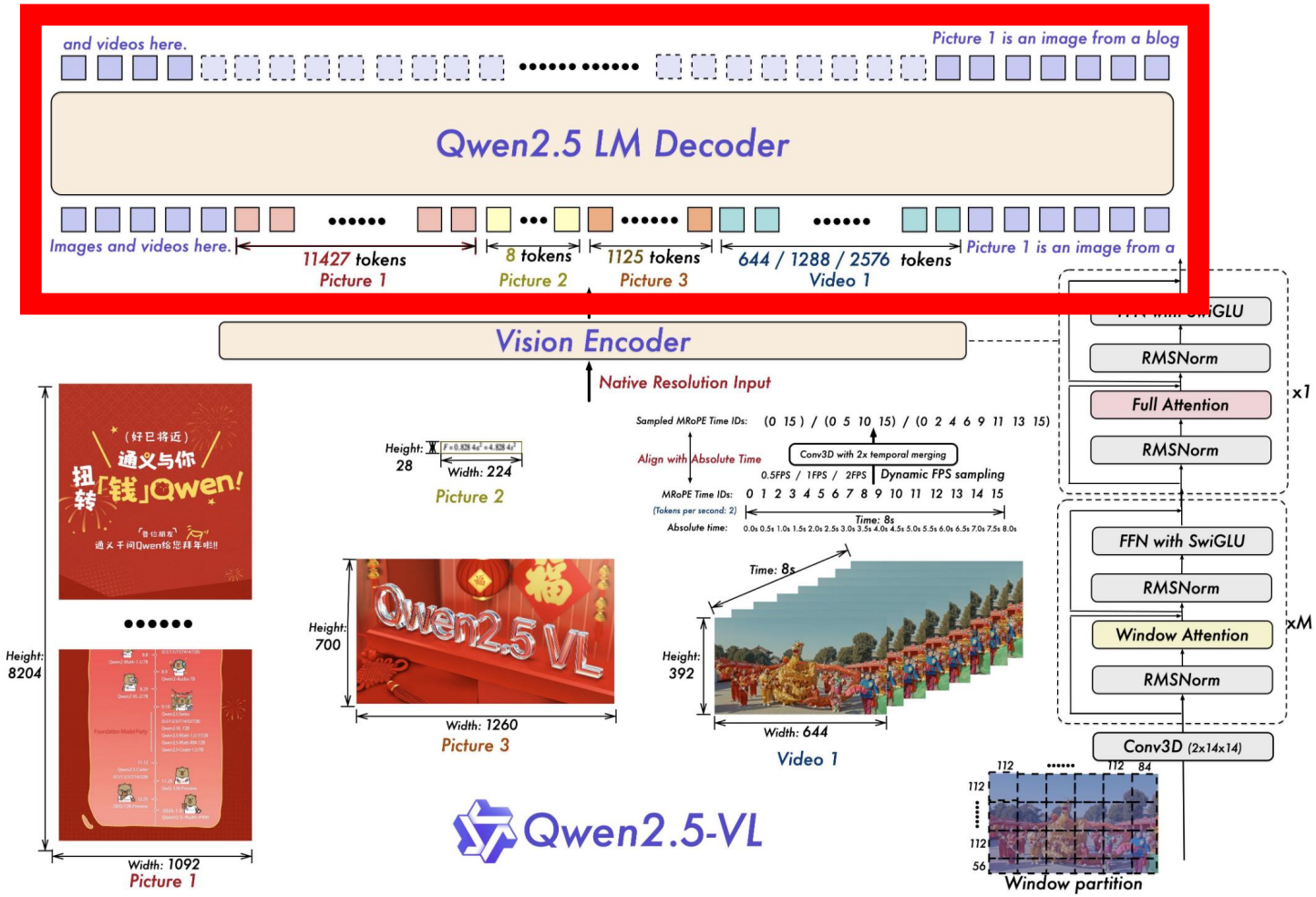
- Patch size: **2 x** 14 x 14
  - Bind two consecutive timesteps
- Dynamic FPS sampling
  - Vision Encoder handles various FPS
- Temporal position
  - aligned with absolute time
  - Qwen2-VL: simple frame index



Georgia Tech

LLM

- Initialized with Qwen2.5[1]
- We omit the details of LLM



[1] Qwen Team, *Qwen2.5 Technical Report*, 2025

Georgia Tech

Model Configuration

Configuration	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-72B
Vision Transformer (ViT)			
Hidden Size	1280	1280	1280
# Layers	32	32	32
# Num Heads	16	16	16
Intermediate Size	3456	3456	3456
Patch Size	14	14	14
Window Size	112	112	112
Full Attention Block Indexes	{7, 15, 23, 31}	{7, 15, 23, 31}	{7, 15, 23, 31}
Vision-Language Merger			
In Channel	1280	1280	1280
Out Channel	2048	3584	8192
Large Language Model (LLM)			
Hidden Size	2048	3,584	8192
# Layers	36	28	80
# KV Heads	2	4	8
Head Size	128	128	128
Intermediate Size	4864	18944	29568
Embedding Tying	✓	✗	✗
Vocabulary Size	151646	151646	151646
# Trained Tokens	4.1T	4.1T	4.1T

Training Overview

Pre-Training

1. Visual Pre-Training
2. Multimodal Pre-Training
3. Long-Context Pre-Training

Post-Training

1. Supervised Fine-Tuning
   - Domain-Tailored Filtering
   - Rejection Sampling
2. Direct Preference Optimization

Data Processing Overview

Pre-Training

Interleaved Image-Text Data

Document Omni-Parsing Data

OCR Data

Video Data

Agent Data

Post-Training

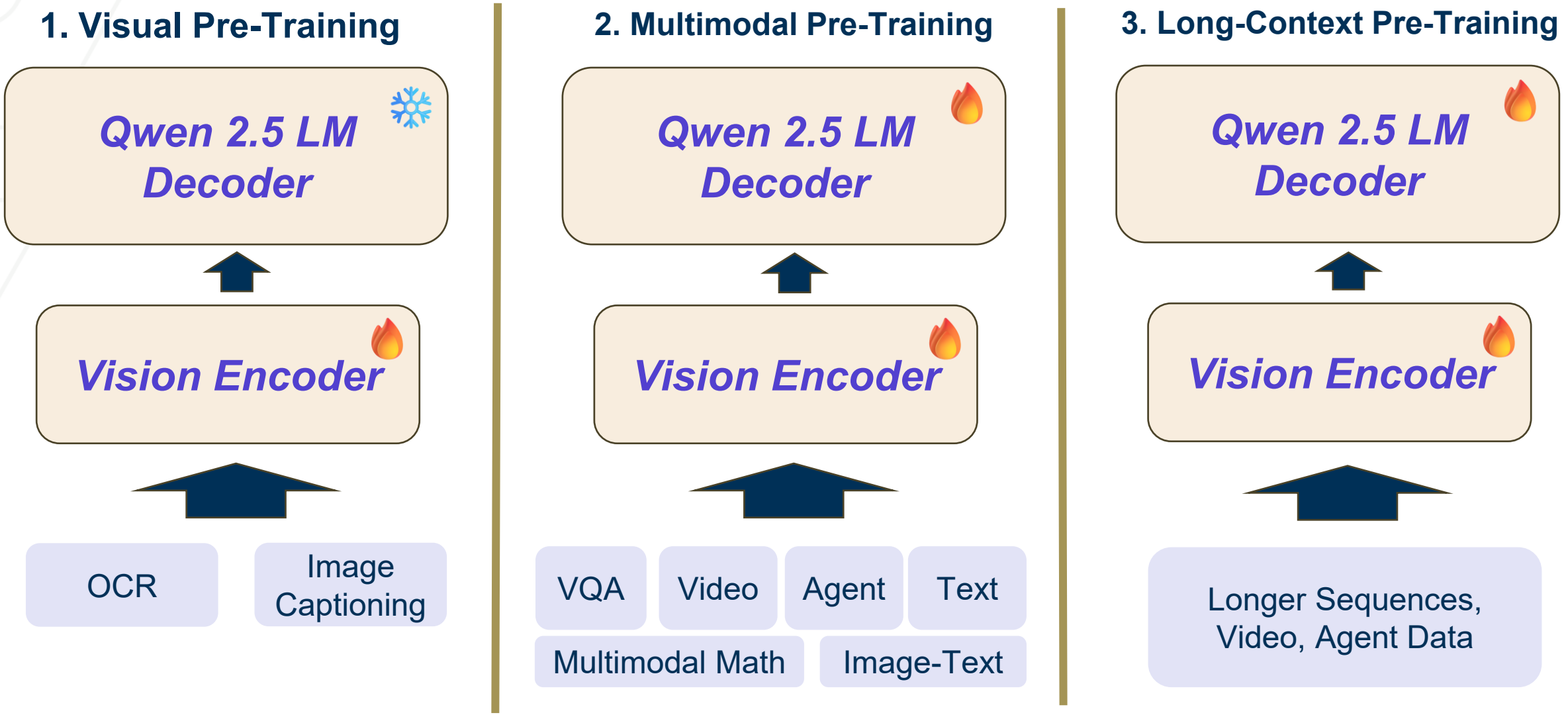
Instruction Data

VQA Data

Image Captioning

...

Pre-Training



Pre-Training Data

Interleaved Image-Text Data

- Standard data cleaning (Wei Li et al. 2024)
- Four-stage scoring system
  (1) text-only quality
  (2) image-text relevance
  (3) image-text complementarity
  (4) information density balance

Pre-Training Data

Interleaved Image-Text | Document Omni-Parsing | OCR | Video | Agent

Document Omni-Parsing

- Convert document data into HTML format
- Include the coordinates for each module

QwenVL HTML Format

```
<html><body>
# paragraph
<p data-bbox="x1 y1 x2 y2"> content </p>
# table
<style>table{id} style</style><table data-bbox="x1 y1 x2 y2" class="table{id}"> table content
</table>
# chart
<div class="chart" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><table> chart content
</table></div>
# formula
<div class="formula" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /> <div> formula
content </div></div>
# image caption
<div class="image caption" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> image
caption </p></div>
# image ocr
<div class="image ocr" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> image ocr
</p></div>
# music sheet
<div class="music sheet" format="abc notation" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1
x2 y2" /> <div> music sheet content </div></div>
# chemical formula content
<div class="chemical formula" format="smile" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1
x2 y2" /> <div> chemical formula content </div></div>
</html></body>
```

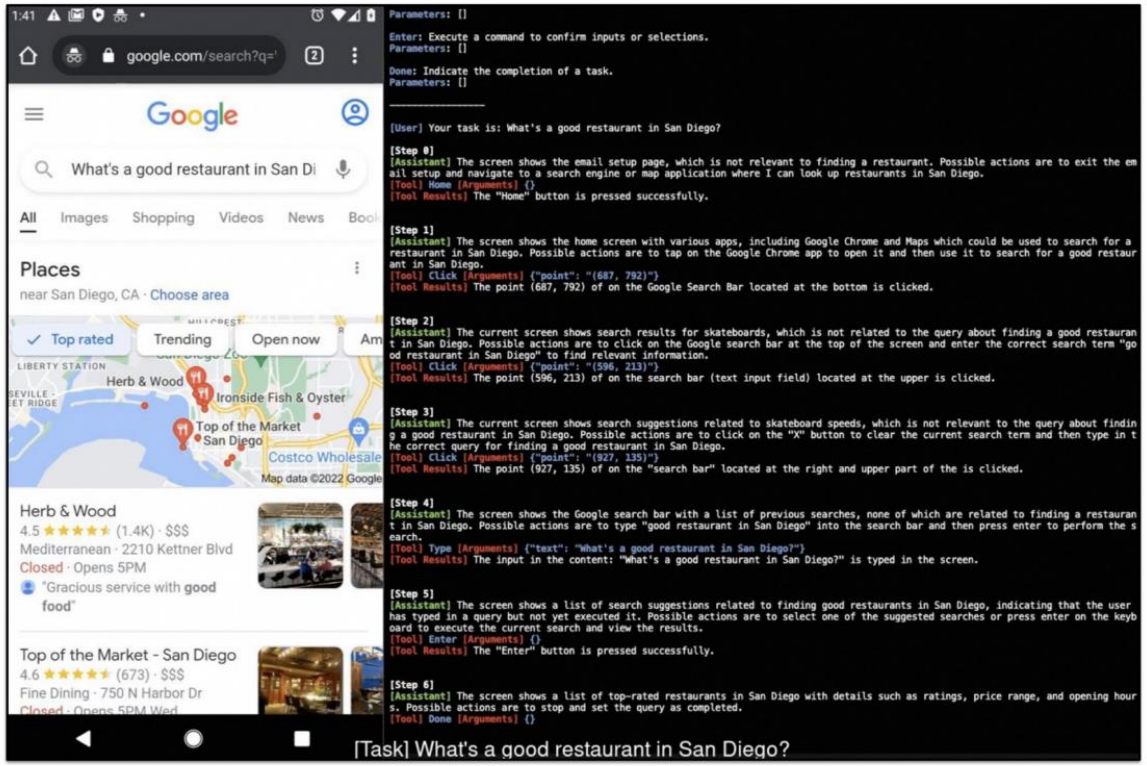
Pre-Training Data

Interleaved Image-Text | Document Omni-Parsing | OCR | Video | Agent

Agent Data

- Collect screenshots on mobile, web, and desktop platforms
- For decision-making, unify the operations across platforms into a function call
- Given a GT operation, explanation for reasoning content is generated by human and model annotators

UI Interaction



Pre-Training

Stages	Visual Pre-Training	Multimodal Pre-Training	Long-Context Pre-Training
Data	Image Caption Knowledge OCR	+ Pure text Interleaved Data VQA, Video Grounding, Agent	+ Long Video Long Agent Long Document
Tokens	1.5T	2T	0.6T
Sequence length	8192	8192	32768
Training	ViT	ViT & LLM	ViT & LLM

Post-Training Data

1. Supervised Fine-Tuning(SFT)

- 2 million entries
- Evenly distributed between text data(50%) and multimodal data(50%)

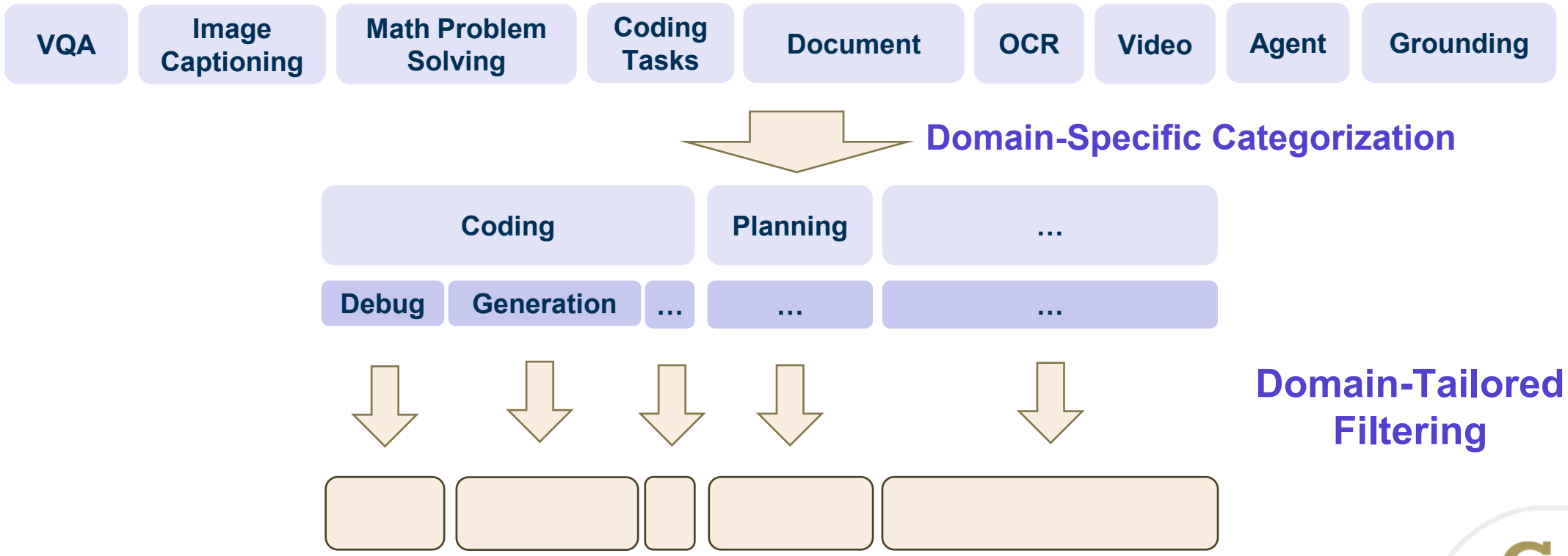
VQA	Image Captioning	Math Problem Solving	Coding Tasks	Document	OCR	Video	Agent	Grounding

2. Direct Preference Optimization(DPO)

Post-Training Data

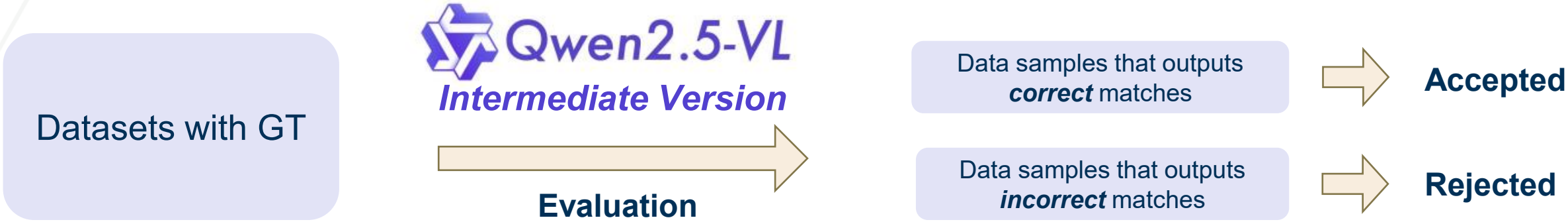
1. Supervised Fine-Tuning(SFT)

- 2 million entries
- Evenly distributed between text data(50%) and multimodal data(50%)



Post-Training Data

1. Supervised Fine-Tuning(SFT)
 - Domain-Specific Categorization and Domain-Tailored Filtering
 - ***Rejection Sampling***



Georgia Tech

Post-Training Data

1. Supervised Fine-Tuning(SFT)
 - Domain-Specific Categorization and Domain-Tailored Filtering
 - ***Rejection Sampling***

Does it learn easy dialogues and problems only?

Georgia Tech

Post-Training Data

1. Supervised Fine-Tuning(SFT)
 - Domain-Specific Categorization and Domain-Tailored Filtering
 - Rejection Sampling

2. Direct Preference Optimization(DPO)
 - Only on image-text and pure text data

Experiments and Results

Table 3: **Performance of Qwen2.5-VL and State-of-the-art.**

Datasets	Previous Open-source SoTA	Claude-3.5 Sonnet-0620	GPT-4o 0513	InternVL2.5 78B	Qwen2-VL 72B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
College-level Problems								
$MMMU_{val}$ (Yue et al., 2023)	70.1 Chen et al. (2024d)	68.3	69.1	70.1	64.5	**70.2**	58.6	53.1
$MMMU\text{-}Pro_{overall}$ (Yue et al., 2024)	48.6 Chen et al. (2024d)	51.5	**51.9**	48.6	46.2	51.1	38.3	31.56
Math								
$MathVista_{mini}$ (Lu et al., 2024)	72.3 Chen et al. (2024d)	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
$MATH\text{-}Vision_{full}$ (Wang et al., 2024d)	32.2 Chen et al. (2024d)	-	30.4	32.2	25.9	**38.1**	25.1	21.2
$MathVerse_{mini}$ (Zhang et al., 2024c)	51.7 Chen et al. (2024d)	-	50.2	51.7	-	**57.6**	49.2	47.6
General Visual Question Answering								
MegaBench (Chen et al., 2024b)	47.4 MiniMax et al. (2025)	52.1	**54.2**	45.6	46.8	51.3	36.8	28.9
$MMBench\text{-}EN_{test}$ (Liu et al., 2023d)	88.3 Chen et al. (2024d)	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
$MMBench\text{-}CN_{test}$ (Liu et al., 2023d)	88.5 Chen et al. (2024d)	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
$MMBench\text{-}V1.1\text{-}EN_{test}$ (Liu et al., 2023d)	87.4 Chen et al. (2024d)	80.9	83.1	87.4	86.1	**88.4**	82.6	77.4
MMStar (Chen et al., 2024c)	69.5 Chen et al. (2024d)	65.1	64.7	69.5	68.3	**70.8**	63.9	55.9
MME_{sum} (Fu et al., 2023)	**2494** Chen et al. (2024d)	1920	2328	**2494**	2483	2448	2347	2157
MuirBench (Wang et al., 2024a)	63.5 Chen et al. (2024d)	-	68.0	63.5	-	**70.7**	59.6	47.7
$BLINK_{val}$ (Fu et al., 2024c)	63.8 Chen et al. (2024d)	-	**68.0**	63.8	-	64.4	56.4	47.6
$CRPE_{relation}$ (Wang et al., 2024h)	78.8 Chen et al. (2024d)	-	76.6	78.8	-	**79.2**	76.4	73.6
$HallBench_{avg}$ (Guan et al., 2023)	**58.1** Wang et al. (2024f)	55.5	55.0	57.4	**58.1**	55.2	52.9	46.3
MTVQA (Tang et al., 2024)	**31.9** Chen et al. (2024d)	25.7	27.8	31.9	30.9	31.7	29.2	24.8
$RealWorldQA_{avg}$ (X.AI, 2024)	78.7 Chen et al. (2024d)	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
$MME\text{-}RealWorld_{en}$ (Zhang et al., 2024f)	62.9 Chen et al. (2024d)	51.6	45.2	62.9	-	**63.2**	57.4	53.1
$MMVet_{turbo}$ (Yu et al., 2024)	74.0 Wang et al. (2024f)	70.1	69.1	72.3	74.0	**76.2**	67.1	61.8
MM-MT-Bench (Agrawal et al., 2024)	7.4 Agrawal et al. (2024)	7.5	**7.72**	-	6.59	7.6	6.3	5.7

Experiments and Results

Table 3: **Performance of Qwen2.5-VL and State-of-the-art.**

Datasets	Previous Open-source SoTA	Claude-3.5 Sonnet-0620	GPT-4o 0513	InternVL2.5 78B	Qwen2-VL 72B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
College-level Problems								
$MMMU_{val}$ (Yue et al., 2023)	70.1 Chen et al. (2024d)	68.3	69.1	70.1	64.5	**70.2**	58.6	53.1
$MMMU\text{-}Pro_{overall}$ (Yue et al., 2024)	48.6 Chen et al. (2024d)	51.5	**51.9**	48.6	46.2	51.1	38.3	31.56
Math								
$MathVista_{mini}$ (Lu et al., 2024)	72.3 Chen et al. (2024d)	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
$MATH\text{-}Vision_{full}$ (Wang et al., 2024d)	32.2 Chen et al. (2024d)	-	30.4	32.2	25.9	**38.1**	25.1	21.2
$MathVerse_{mini}$ (Zhang et al., 2024c)	51.7 Chen et al. (2024d)	-	50.2	51.7	-	**57.6**	49.2	47.6
General Visual Question Answering								
MegaBench (Chen et al., 2024b)	47.4 MiniMax et al. (2025)	52.1	**54.2**	45.6	46.8	51.3	36.8	28.9
$MMBench\text{-}EN_{test}$ (Liu et al., 2023d)	88.3 Chen et al. (2024d)	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
$MMBench\text{-}CN_{test}$ (Liu et al., 2023d)	88.5 Chen et al. (2024d)	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
$MMBench\text{-}V1.1\text{-}EN_{test}$ (Liu et al., 2023d)	87.4 Chen et al. (2024d)	80.9	83.1	87.4	86.1	**88.4**	82.6	77.4
MMStar (Chen et al., 2024c)	69.5 Chen et al. (2024d)	65.1	64.7	69.5	68.3	**70.8**	63.9	55.9
MME_{sum} (Fu et al., 2023)	**2494** Chen et al. (2024d)	1920	2328	**2494**	2483	2448	2347	2157
MuirBench (Wang et al., 2024a)	63.5 Chen et al. (2024d)	-	68.0	63.5	-	**70.7**	59.6	47.7
$BLINK_{val}$ (Fu et al., 2024c)	63.8 Chen et al. (2024d)	-	**68.0**	63.8	-	64.4	56.4	47.6
$CRPE_{relation}$ (Wang et al., 2024h)	78.8 Chen et al. (2024d)	-	76.6	78.8	-	**79.2**	76.4	73.6
$HallBench_{avg}$ (Guan et al., 2023)	**58.1** Wang et al. (2024f)	55.5	55.0	57.4	**58.1**	55.2	52.9	46.3
MTVQA (Tang et al., 2024)	**31.9** Chen et al. (2024d)	25.7	27.8	31.9	30.9	31.7	29.2	24.8
$RealWorldQA_{avg}$ (X.AI, 2024)	78.7 Chen et al. (2024d)	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
$MME\text{-}RealWorld_{en}$ (Zhang et al., 2024f)	62.9 Chen et al. (2024d)	51.6	45.2	62.9	-	**63.2**	57.4	53.1
$MMVet_{turbo}$ (Yu et al., 2024)	74.0 Wang et al. (2024f)	70.1	69.1	72.3	74.0	**76.2**	67.1	61.8
MM-MT-Bench (Agrawal et al., 2024)	7.4 Agrawal et al. (2024)	7.5	**7.72**	-	6.59	7.6	6.3	5.7

Performance degradation

Experiments and Results

Table 4: **Performance on pure text tasks of the 70B+ Instruct models and Qwen2.5-VL.**

Datasets	Llama-3.1-70B	Llama-3.1-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-VL-72B
General Tasks					
MMLU-Pro	66.4	**73.3**	64.4	71.1	71.2
MMLU-redux	83.0	86.2	81.6	**86.8**	85.9
LiveBench-0831	46.6	53.2	41.5	52.3	**57.0**
Mathematics & Science Tasks					
GPQA	46.7	**51.1**	42.4	49.0	49.0
MATH	68.0	73.8	69.0	**83.1**	83.0
GSM8K	95.1	**96.8**	93.2	95.8	95.3
Coding Tasks					
HumanEval	80.5	**89.0**	86.0	86.6	87.8
MultiPL-E	68.2	73.5	69.2	75.1	**79.5**
Alignment Tasks					
IFEval	83.6	86.0	77.6	84.1	**86.3**

Experiments and Results

Retained performance on text-only evaluation, due to pure text data(50%) in train datasets

Table 4: **Performance on pure text tasks of the 70B+ Inst**

Datasets	Llama-3.1-70B	Llama-3.1-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-VL-72B
General Tasks					
MMLU-Pro	66.4	**73.3**	64.4	71.1	71.2
MMLU-redux	83.0	86.2	81.6	**86.8**	85.9
LiveBench-0831	46.6	53.2	41.5	52.3	**57.0**
Mathematics & Science Tasks					
GPQA	46.7	**51.1**	42.4	49.0	49.0
MATH	68.0	73.8	69.0	**83.1**	83.0
GSM8K	95.1	**96.8**	93.2	95.8	95.3
Coding Tasks					
HumanEval	80.5	**89.0**	86.0	86.6	87.8
MultiPL-E	68.2	73.5	69.2	75.1	**79.5**
Alignment Tasks					
IFEval	83.6	86.0	77.6	84.1	**86.3**

Experiments and Results

Table 5: **Performance of Qwen2.5-VL and other models on OCR, chart, and document understanding benchmarks.**

Datasets	Claude-3.5 Sonnet	Gemini 1.5 Pro	GPT 4o	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
OCR-related Parsing Tasks							
CC-OCR	62.5	73.0	66.9	64.7	**79.8**	77.8	74.5
$\text{OmniDocBench}_{\text{edit en/zh}}\downarrow$	0.330/0.381	0.230/**0.281**	0.265/0.435	0.275/0.324	**0.226**/0.324	0.308/0.398	0.409/0.543
OCR-related Understanding Tasks							
$\text{AI2D}_{\text{w. M.}}$	81.2	88.4	84.6	**89.1**	88.7	83.9	81.6
$\text{TextVQA}_{\text{val}}$	76.5	78.8	77.4	83.4	83.5	**84.9**	79.3
$\text{DocVQA}_{\text{test}}$	95.2	93.1	91.1	95.1	**96.4**	95.7	93.9
$\text{InfoVQA}_{\text{test}}$	74.3	81.0	80.7	84.1	**87.3**	82.6	77.1
$\text{ChartQA}_{\text{test Avg.}}$	**90.8**	87.2	86.7	88.3	89.5	87.3	84.0
$\text{CharXiv}_{\text{RQ/DQ}}$	**60.2**/84.3	43.3/72.0	47.1/84.5	42.4/82.3	49.7/**87.4**	42.5/73.9	31.3/58.6
SEED-Bench-2-Plus	71.7	70.8	72.0	71.3	**73.0**	70.4	67.6
OCRBench	788	754	736	854	**885**	864	797
$\text{VCR}_{\text{En-Hard-EM}}$	41.7	28.1	73.2	-	79.8	**80.5**	37.5
OCR-related Comprehensive Tasks							
$\text{OCRBench\_v2}_{\text{en/zh}}$	45.2/39.6	51.9/43.1	46.5/32.2	49.8/52.1	**61.5/63.7**	56.3/57.2	54.3/52.1

Less performance drop compared to multimodal problems and math problems

Experiments and Results

Table 6: **Performance of Qwen2.5-VL and other models on grounding.**

Datasets	Gemini 1.5 Pro	Grounding DINO	Molmo 72B	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
Refcoco_{val}	73.2	90.6	-	93.7	92.7	90.0	89.1
Refcoco_{testA}	72.9	93.2	-	95.6	94.6	92.5	91.7
Refcoco_{testB}	74.6	88.2	-	92.5	89.7	85.4	84.0
Refcoco+_{val}	62.5	88.2	-	90.4	88.9	84.2	82.4
Refcoco+_{testA}	63.9	89.0	-	94.7	92.2	89.1	88.0
Refcoco+_{testB}	65.0	75.9	-	86.9	83.7	76.9	74.1
Refcocog_{val}	75.2	86.1	-	92.7	89.9	87.2	85.2
Refcocog_{test}	76.2	87.0	-	92.2	90.3	87.2	85.7
ODinW	36.7	55.0	-	31.7	43.1	37.3	37.5
PointGrounding	-	-	69.2	-	67.5	67.3	58.3

Table 7: **Performance of Qwen2.5-VL and other models on counting.**

Datasets	Gemini 1.5-Pro	GPT-4o	Claude-3.5 Sonnet	Molmo-72b	InternVL2.5-78B	Qwen2.5-VL-72B
CountBench	85.5	87.9	89.7	91.2	72.1	93.6

Experiments and Results

Significantly outperforming results on long video understanding, due to MRoPE and dynamic FPS sampling

Table 8: **Performance of Qwen2.5-VL and other models on video benchmarks.**

Datasets	Gemini 1.5-Pro	GPT-4o	Qwen2.5-VL-72B	Qwen2.5-VL-7B	Qwen2.5-VL-3B
Video Understanding Tasks					
Video-MME$_{\text{w/o sub.}}$	**75.0**	71.9	73.3	65.1	61.5
Video-MME$_{\text{w sub.}}$	**81.3**	77.2	79.1	71.6	67.6
Video-MMMU	53.9	**61.2**	60.2	47.4	-
MMVU$_{\text{val}}$	65.4	**67.4**	62.9	50.1	-
MVBench	60.5	64.6	**70.4**	69.6	67.0
MMBench-Video	1.30	1.63	**2.02**	1.79	1.63
LongVideoBench$_{\text{val}}$	64.0	**66.7**	60.7	56.0	54.2
LVBench	33.1	30.8	**47.3**	45.3	43.3
EgoSchema$_{\text{test}}$	71.2	72.2	**76.2**	65.0	64.8
PerceptionTest$_{\text{test}}$	-	-	**73.2**	70.5	66.9
MLVU$_{\text{M-Avg}}$	-	64.6	**74.6**	70.2	68.2
TempCompass$_{\text{Avg}}$	67.1	73.8	**74.8**	71.7	64.4
Video Grounding Tasks					
Charades-STA$_{\text{mIoU}}$	-	35.7	**50.9**	43.6	38.8

Long Video

Experiments and Results

Table 9: **Performance of Qwen2.5-VL and other models on GUI Agent benchmarks.**

Benchmarks	GPT-4o	Gemini 2.0	Claude	Aguvis-72B	Qwen2-VL-72B	Qwen2.5-VL-72B
ScreenSpot	18.1	84.0	83.0	**89.2**	-	87.1
ScreenSpot Pro	-	-	17.1	23.6	1.6	**43.6**
Android Control High_{EM}	20.8	28.5	12.5	66.4	59.1	**67.36**
Android Control Low_{EM}	19.4	60.2	19.4	84.4	59.2	**93.7**
$\text{AndroidWorld}_{\text{SR}}$	34.5% (SoM)	26% (SoM)	27.9%	26.1%	6% (SoM)	**35%**
$\text{MobileMiniWob++}_{\text{SR}}$	61%	42% (SoM)	61% (SoM)	66%	50% (SoM)	**68%**
OSWorld	5.03	4.70	**14.90**	10.26	2.42	8.83

Outperform other models even when Set-of-Mark(SoM) is applied to their inputs

Strengths

- Can handle images at their native resolution
- Improved long video comprehension by dynamic FPS sampling
- SOTA on some document parsing benchmark
- Strong visual agent capabilities
- Improved computational efficiency by applying RMSNorm and SwiGLU

Weaknesses & Limitations

- Lacks long-from video performance comparison with the previous version, Qwen 2-VL

Qwen3-Omni Technical Report

Qwen Team, Alibaba Group

Sep 2025

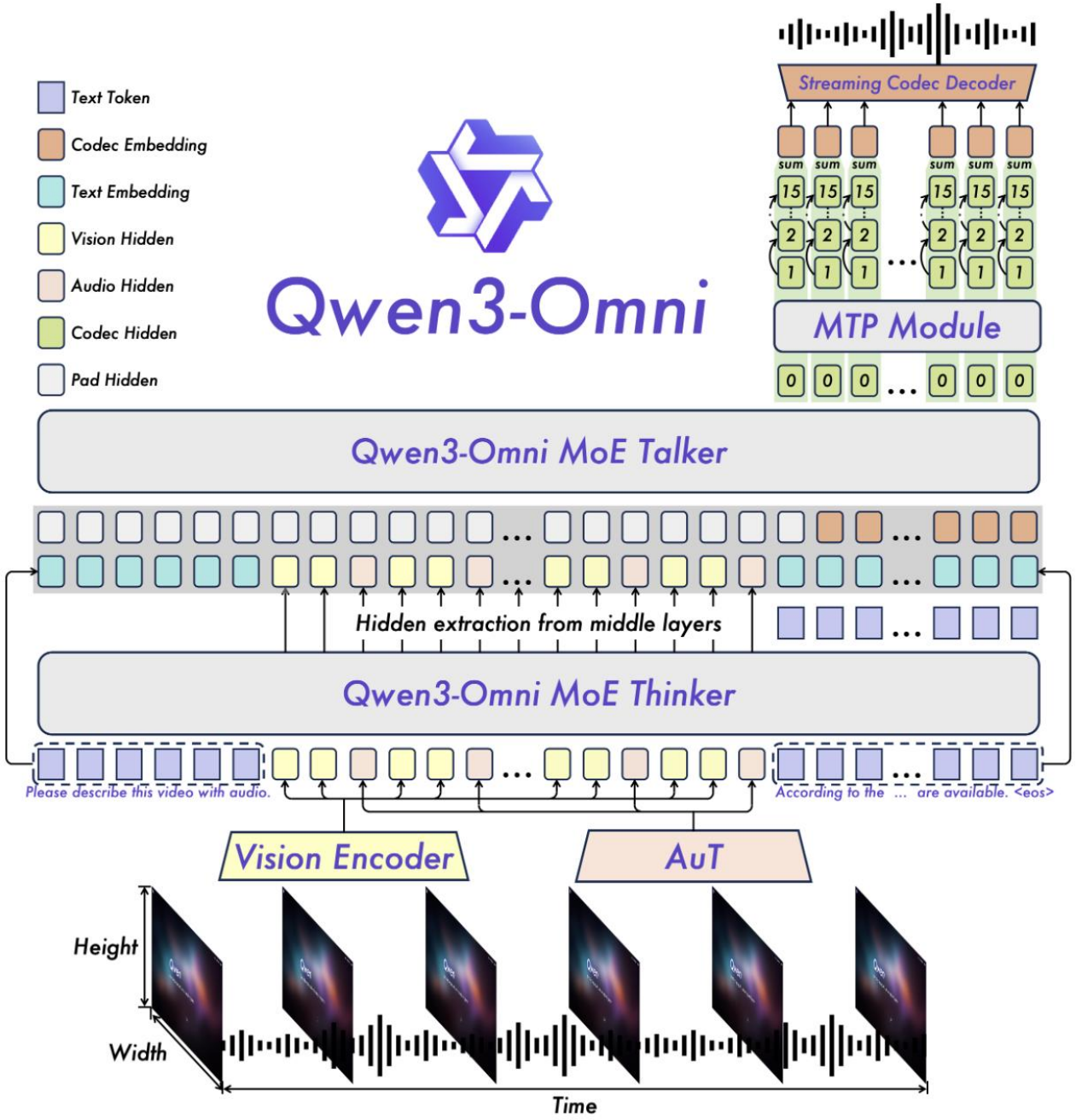
Jaehyeon Son, Junhyun Kim

Georgia Tech

Architecture

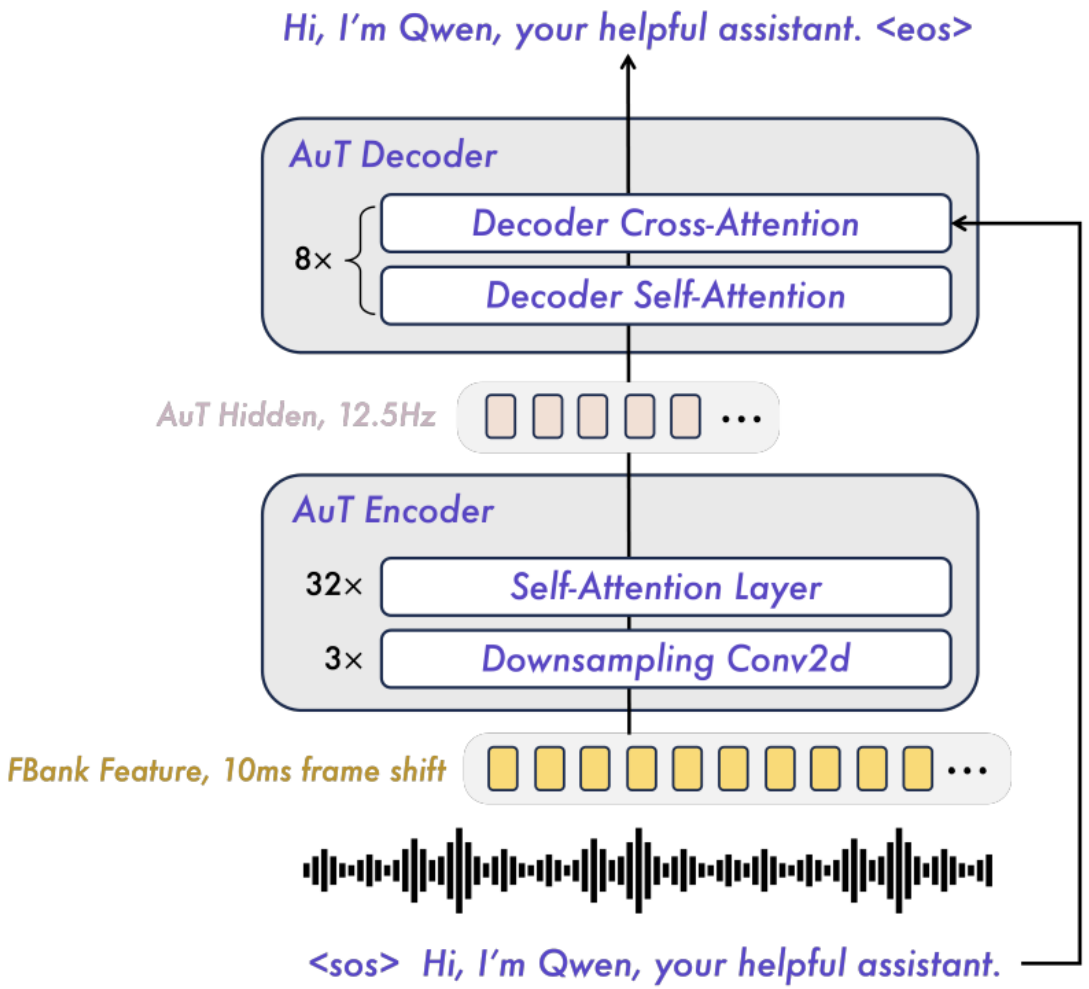
- Audio Transformer (AuT)
- Mixture of Experts (MoE)

- Thinker-Talker (Qwen 2.5-Omni)
- TMRoPE (Time-aligned MRoPE) (Qwen 2.5-Omni)



Architecture

- Audio Transformer (AuT)
  - The filter bank features are downsampled using Conv2D blocks.
  - Token rate is reduced to 12.5 Hz in hidden layer.
  - 0.6B parameters



Results

Table 11: **AudioVisual → Text performance of Qwen3-Omni-Instruct and other non-reasoning baselines. The highest scores are shown in bold.**

Datasets	Previous Open-source SoTA	Gemini-2.5-Flash	Qwen2.5-Omni	Qwen3-Omni-30B-A3B -Instruct	Qwen3-Omni-Flash -Instruct
WorldSense	47.1(Yang et al., 2025b)	50.9	45.4	54.0	**54.1**

Table 12: **AudioVisual → Text performance of Qwen3-Omni-30B-A3B-Thinking and other reasoning baselines. The highest scores are shown in bold.**

Datasets	Previous Open-source SoTA	Gemini-2.5-Flash -Thinking	Qwen3-Omni-30B-A3B -Thinking	Qwen3-Omni-Flash -Thinking
DailyOmni	69.8(Tang et al., 2025)	72.7	75.8	**76.2**
VideoHolmes	55.6(Tang et al., 2025)	49.5	**57.3**	**57.3**

Thank you

Georgia Tech