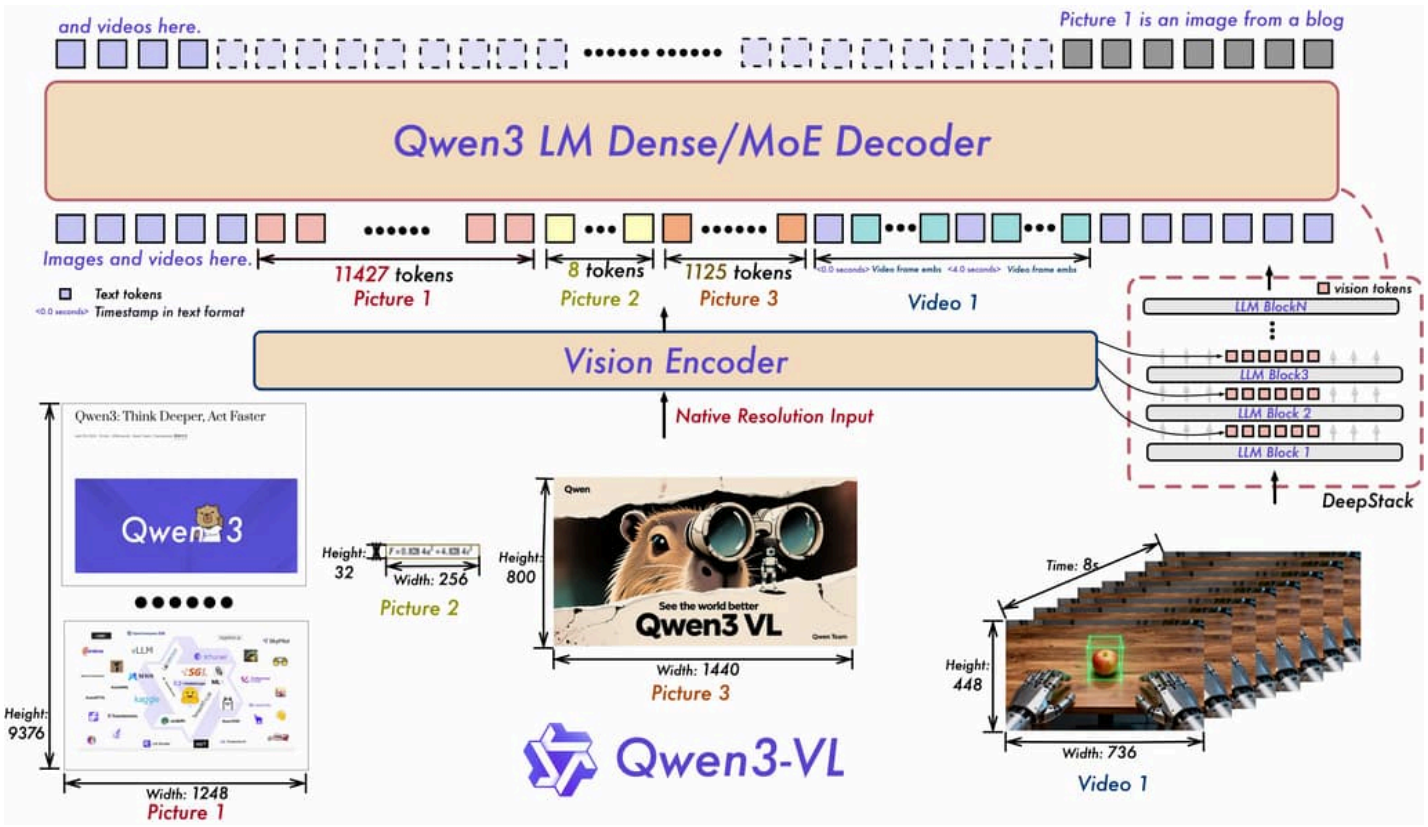
[Paper Review] Qwen3-VL Technical Report (Video Understanding)

arxiv.org

https://arxiv.org/pdf/2511.21631



Qwen3-VL은 구조적 설계 측면에서 **세 가지 중대한 개선**을 이루었음

1. **Long-range Video Understanding**
2. **Fine-grained Visual Understanding**
3. **Precise Temporal Reasoning**

1. Long-range Video Understanding

▼ Preliminary

Positional Encoding : Transformer는 입력 토큰의 “순서”를 스스로 알 수 없기 때문에, 토큰이 문장 내에서 어느 위치에 있는지 알려줌

기법	방법	효과	한계
PE (Absolute Positional Encoding)	토큰 임베딩에 **sin/cos 기반 절대 위치 벡터를 더함** `X_p = E_p + PE(p)`	- 토큰의 절대적 위치(앞/뒤)를 알 수 있음	- Attention에서 중요한 상대 거리(p−q)가 자연스럽게 반영되지 않음
RoPE (Rotary Position Embedding)	입력에 더하지 않고, **Q/K를 위치 p에 따라 회전** `Q̃_p = R(ωp)Q_p` `K̃_p = R(ωp)K_p` (ω: 주파수 벡터, p: 위치, ωp: 각도 벡터)	- Attention score 계산 시 `(R(p)Q)ᵀ(R(q)K) = QᵀR(p−q)K` → Q와 K의 내적 사이에 ‘각도 차이(ω(p−q))’가 생기므로, 상대 위치가 자연스럽게 반영됨	- 기본적으로 **1D 시퀀스(텍스트)** 전용 구조 - 토큰 사이의 상대적 각도가 너무 커지면 멀리 떨어진 위치들이 비슷하게 보일 수 있음
MRoPE (Multi-Dimensional RoPE)	RoPE를 2D/3D에 확장.(t, h, w) 차원을 **블록 단위**로 나누어 축별로 RoPE 적용	- t/h/w 각 축을 독립적으로 인코딩 가능 - t축 RoPE는 t축 차원에서 적용, 다른 축도 마찬가지	- 연속된 블록 배치 때문에 **주파수 편향** 발생 (예: t 블록이 뒤에 있으면 **고주파만** 차지 → 장기 시간 정보 손실)
Interleaved-MRoPE	블록을 없애고 t/h/w 차원을 **교차(interleave)** 배치: `t, h, w, t, h, w, ...`	- t/h/w가 **저·중·고 주파수 전체에 고르게 분포** - 특히 시간축(t)이 저주파(장기 변화), 중주파(중간 패턴), 고주파(세밀한 변화)모두 획득 → **긴 비디오 이해 성능 향상**	- 구현 복잡도 증가

- **Interleaved-MRoPE**

- 기존의 MRoPE(Multi-dimensional Rotary Position Embedding)는 시간(t), 높이(h), 너비(w)의 축을 블록 단위로 나누어 차원에 할당
  - 모든 시간 정보가 고주파 차원(high-frequency dimensions)에 집중되는 한계를 가지고 있었으며, 특히 **긴 비디오 데이터의 시간적 맥락 이해**에서 성능 저하가 발생함
- **Interleaved-MRoPE**를 도입하여 균형잡힌 위치 임베딩을 수행함
  - t, h, w를 교차적으로 배치하여 주파수 대역 전반에 고르게 분산시킴으로써, 시간·공간 축 전체에서 균형 잡힌 위치 인코딩을 제공
- 이미지 인식 능력을 유지하면서도, 장시간 비디오 분석과 같은 과제에서 현저히 강화된 성능을 발휘함

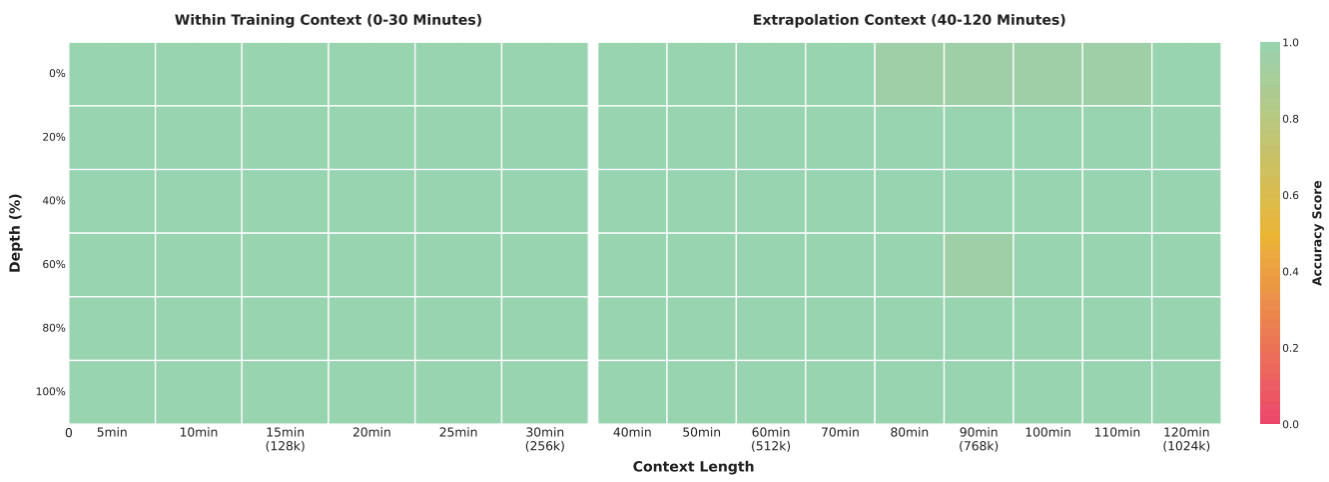


Figure 3: Needle-in-a-Haystack performance heatmap for Qwen3-VL-235B-A22B-Instruct across varying video durations and needle positions. Each cell shows accuracy (%) for locating and answering questions about the inserted "needle" frame.

- **long-form video modeling**
  - 논문은 **비디오·이미지·텍스트가 섞여 있는 초장문 입력을 그대로 처리할 수 있는 모델**을 만들고자 함
  - 기본 context window(훈련 시 실제 사용 길이) 자체를 256K로 확장했고, 그 위로 RoPE extrapolation(YaRN)을 써서 훈련되지 않은 초장문인 1M 토큰까지도 안정적으로 처리하도록 설계
    - YaRN: RoPE를 더 길고 자연스럽게 사용할 수 있게 만드는 기술로, 긴 문서를 읽고도 앞·뒷부분 관계를 잘 이해하게 함
  - **4단계 pre-training**

Table 1: Training setup and hyperparameters across different stages for Qwen3-VL.

Stage	Objective	Training	Token Budget	Sequence Length
S0	Vision-Language Alignment	Merger	67B	8,192
S1	Multimodal Pre-Training	All	~1T	8,192
S2	Long-Context Pre-Training	All	~1T	32,768
S3	Ultra-Long-Context Adaptation	All	100B	262,144

    - **S0:** 8K, 모달리티 align (vision–language 정렬만 업데이트, vision encoder-LLM freeze)
    - **S1:** 8K, full multimodal pretraining
    - **S2:** 32K, long-context 비중 확대 (long docs + long videos)
    - **S3: 262K**, ultra-long context 적응
    - captioning/OCR(저수준 인식 정렬) → VQA·visual grounding(고수준 reasoning)
  - 256K 안에 비디오 하나를 넣으려면 **프레임 수 × 프레임당 visual token 수**가 너무 커질 수 있음
    - 최대 프레임 2,048장, 프레임당 visual token을 약 640~768 개로 제한
    - MLP Merger가 2×2 패치 그룹을 하나의 visual token으로 압축
    - 훈련에서는 256K까지만 실제로 보고,그 이후의 길이—40~120분(≈1M tokens)—은 YaRN을 적용해 확장함

2. **Fine-grained Visual Understanding**

- **Multi-Level DeepStack**

- ViT 여러 레이어의 feature를 LLM 레이어로 직접 주입
- 저수준(fine-grained) + 중간 수준(지역적 구조) + 고수준(semantic) 시각정보 모두 reasoning에 전달
- 결과적으로 InfoVQA, DocVQA 등 **문서·차트·세밀 인식 벤치에서 일관된 성능 향상**을 보임

Table 12: **Ablation on DeepStack.** We conduct the ablation study on the DeepStack using an internal 15B-A2B LLM, with all experiments pretrained on 200 billion tokens. We directly evaluate these pretrained models on the validation sets, without any post-training.

Method	AVG	AI2D	OCRB	TVQA	InfoVQA	ChartQA	DocVQA	MMMU	MMStar	RLWDQA	MMB_{EN}	MMB_{CN}
Baseline	74.7	81.8	81.0	80.6	71.9	81.5	89.5	52.9	55.5	67.7	81.0	78.1
DeepStack	76.0	83.2	83.6	80.5	74.2	83.3	91.1	54.1	57.7	68.1	81.2	78.5

3. **Precise Temporal Reasoning**

- **Text-Timestamp Alignment**
  - 기존에는 비디오 프레임의 temporal position ID를 **초 단위 절대 시간에 직접 매핑한 뒤, 이를 MRoPE의 시간 축 pos로 사용하는 방식**을 썼음
    - 긴 비디오에서는 pos ID가 **너무 크고 희소해져서** long-range temporal context 이해가 부족
    - 다양한 fps에 대해 학습하려면 **프레임 레이트별로 균형 잡힌 대규모 데이터**가 필요
  - Qwen3-VL은 절대 시간을 RoPE에 넣지 않고, **각 비디오 temporal patch 앞에 "텍스트 형태의 타임스탬프 토큰"을 붙여서 시간 정보를 언어로 표현하는 전략**을 채택
    - 각 프레임 앞에 `<3.0 seconds>` 또는 `<HH:MM:SS>` 같은 문자열을 prefix로 붙이고, 모델은 이를 일반 텍스트 토큰처럼 인코딩하여 해석하며, RoPE는 단지 프레임 순서 등의 위치 정보만 담당하도록 설계
  - **pos ID 폭주와 fps 의존성을 줄여** long-video에서도 안정적인 temporal reasoning을 가능하게 함
    - fps가 바뀌어도 '언어적 시간 표현'만 바뀌고 RoPE pos는 안정적으로 유지되기 때문

- **Massive multimodal temporal data**
  - **Dense Caption**
    - 긴 영상을 **짧은 구간들 분할→ 구간별 캡션 생성 → 다시 이어붙여 타임스탬프 달린 긴 스토리 캡션**으로 제작
    - 일어난 사건을 한 줄로 요약하는 문장(이벤트 레벨 요약) + 화면에 보이는 구체적 요소(사람, 물체, 배경, 행동 등)를 같아 쓰도록 함
    - "시간 흐름을 따라가는 사건 요약" + "각 구간의 시각적 디테일"을 동시에 담은 데이터
  - **시공간(spatio-temporal) video grounding**
    - 어느 프레임 구간에서 어떤 객체/행동이 발생하는지를 모델이 추적할 수 있게 하는 데이터임
    - "파란 셔츠를 입은 사람이 5~8초 구간에서 컵을 집어 들고, 8~10초에 마신다"
  - **다양한 long video corpus**
    - 튜토리얼·강의형 instructional 영상, 영화/드라마 같은 cinematic films, 1인칭 egocentric 영상 등 여러 도메인에서 데이터를 수집
    - 데이터 비율을 조정해 균형 잡힌 데이터 분포를 맞춤

- **Performance of Qwen3-VL-235B-A22B and top-tier models on visual benchmarks**
  - **데이터셋**
    - **MVBENCH, Video-MME:** 다양한 비디오 분석 능력을 종합적으로 평가
    - **MVLU:** 영화/드라마 영상의 스토리 이해 능력 평가
    - **LVBench:** 긴 비디오 이해, **Charades:** 특정 행동이 일어나는 시점 파악
    - **MMMU:** 다양한 학술 문제 기반 비디오 이해, **VMVU:** 멀티뷰 비디오 이해
  - **실험 결과**
    - **Charades, MMMU, VMVU와 같이 난이도가 높은 데이터셋은 thinking 모드가 유리**
    - **thinking:** opus 4.1 이상의 성능, **instruct:** GPT-5 수준의 성능

	Benchmark	Qwen3-VL 235B-A22B		Gemini 2.5 Pro		OpenAI GPT-5		Claude Opus 4.1	
		thinking	instruct	thinking	budget-128	high	minimal	thinking	non-thinking
STEM Puzzle	MMMU	80.6	78.7	81.7*	80.9	**84.2***	74.4*	78.4	77.2
	MMMU-Pro	69.3	68.1	68.8*	71.2	**78.4***	62.7*	64.8	60.7
	$\text{MathVista}_{\text{mini}}$	**85.8**	84.9	82.7*	77.7	81.3	50.9	75.5	74.5
	MathVision	**74.6**	66.5	73.3*	66.0	70.9	45.8	64.3	57.7
	$\text{MathVision}_{\text{WP}}$	**63.8**	57.0	63.2	56.9	62.8	40.1	54.0	46.4
	We-Math	74.8	67.5	**80.6**	74.5	73.8	51.8	65.2	60.2
	$\text{MathVerse}_{\text{mini}}$	**85.0**	72.5	82.9	65.9	84.1	43.0	70.6	68.1
	DynaMath	82.8	79.4	80.0	78.5	**85.4**	74.0	75.1	72.0
	Math-VR	**66.8**	65.0	64.7*	54.3	58.1	21.7	54.3	38.0
	ZeroBench	**4**	2	3	1	2	2	3	1
	VlmsAreBlind	79.5	80.4	**86.1**	78.5	80.5	53.4	77.8	72.2
	LogicVista	**72.2**	65.8	72.0	68.7	71.8	46.3	67.3	63.5
	VisuLogic	**34.4**	29.9	31.6	26.9	28.5	27.2	27.9	27.2
	VisualPuzzles	57.2	54.7	**60.9**	56.9	57.3	47.9	48.8	47.6
General VQA	MMBench-EN	88.8	89.3	**90.1***	88.4	83.8	81.3	79.4	83.0
	MMBench-CN	88.6	88.9	**89.7***	86.4	83.5	79.9	84.9	74.3
	RealWorldQA	81.3	79.2	78.0*	76.0	**82.8**	77.3	69.9	68.5
	MMStar	**78.7**	78.4	77.5*	78.5	76.4	65.2	72.1	71.0
	SimpleVQA	61.3	63.0	**65.4**	66.9	61.8	56.7	56.7	55.7
Alignment	HallusionBench	**66.7**	63.2	63.7*	60.9	65.7	53.7	60.4	55.1
	MM-MT-Bench	**8.5**	8.5	8.4*	7.6	7.6	7.5	7.8	7.9
	MIA-Bench	**92.7**	91.3	92.3	91.3	92.4	92.6	91.2	90.0
Document Understanding	DocVQA_{test}	**96.5**	97.1	92.6	94.0	91.5	89.6	92.5	89.2
	InfoVQA_{test}	**89.5**	89.2	84.2	82.9	79.0	69.9	69.4	60.9
	$\text{AI2D}_{\text{w. M.}}$	89.2	89.7	**90.9**	90.0	89.7	84.1	86.4	84.4
	ChartQA_{test}	**90.3**	90.3	83.3	62.6	59.7	59.1	86.2	83.9
	OCRBench	**875**	920	866	872	810	787	764	750
	$\text{OCRBench\_v2}_{\text{en}}$	**66.8**	67.1	54.3	55.2	53.0	48.2	48.4	47.2
	$\text{OCRBench\_v2}_{\text{zh}}$	**63.5**	61.8	48.5	53.1	43.2	37.7	43.7	38.0
	CC-OCR	**81.5**	82.2	77.2	76.8	68.3	66.1	69.1	66.0
	$\text{OmniDocBench}_{\text{en}}$	**0.155**	0.143	0.347	0.206	0.356	0.174	0.194	-
	$\text{OmniDocBench}_{\text{zh}}$	**0.207**	0.207	0.238	0.249	0.472	0.389	0.293	-
	CharXiv(DQ)	90.5	89.4	**94.4**	87.8	89.2	79.5	88.5	87.8
	CharXiv(RQ)	66.1	62.1	67.9	62.9	**81.1***	57.8	63.6	60.2
	$\text{MMLongBench}_{\text{Doc}}$	**56.2**	57.0	55.6	51.2	51.5	42.4	54.5	48.1
2D/3D Grounding	RefCOCO-avg	**92.1**	91.9	74.6*	-	66.8	-	-	-
	CountBench	**93.7**	93.0	91.0*	91.0	91.7	87.8	93.1	91.9
	ODinW-13	**43.2**	48.6	33.7*	34.5	-	-	-	-
	ARKitScenes	**53.7**	56.9	-	-	-	-	-	-
	Hypersim	**11.0**	13.0	-	-	-	-	-	-
	SUNRGBD	**34.9**	39.4	29.7	-	-	-	-	-
Embodied/Spatial Understanding	ERQA	52.5	51.3	55.3	50.3	**65.7***	42.0*	34.8	28.0
	VSI-Bench	**60.0**	62.7	-	-	-	-	-	-
	EmbSpatialBench	**84.3**	83.1	79.1	73.3	82.9	75.1	69.2	66.0
	RefSpatialBench	**69.9**	65.5	36.5	35.6	23.8	23.1	-	-
	RoboSpatialHome	**73.9**	69.4	47.5	49.2	53.5	43.6	-	-
Multi-Image	BLINK	67.1	70.7	70.6*	70.0	**71.0**	62.8	64.1	62.9
	MUIRBENCH	**80.1**	73.0	77.2	74.0	77.5	66.5	-	-
Video Understanding	MVBench	75.2	76.5	69.9	65.8	**75.3**	64.6	61.4	59.0
	$\text{Video-MME}_{\text{w/o sub.}}$	79.0	79.2	**85.1**	80.6	84.7	77.3	75.6	73.3
	$\text{MLVU}_{\text{M-Avg}}$	83.8	84.3	85.6	81.2	**86.2**	78.3	73.5	71.2
	LVBench	63.6	67.7	**73.0**	69.0	-	-	-	-
	$\text{Charades-STA}_{\text{mIoU}}$	**63.5**	64.8	-	-	-	-	-	-
	VideoMMMU	80.0	74.7	83.6*	79.4	**84.6***	61.6*	76.2	70.1
	MMVU	71.1	68.1	**74.9**	72.2	73.0	68.1	66.4	61.4
Perception with Tool	V*	**85.9**	93.7+	83.8	72.7	72.8	56.7	-	-
	HRBench4K	84.3	85.4+	**87.3**	84.8	-	-	-	-
	HRBench8K	76.6	82.4+	**85.4**	80.1	-	-	-	-
Multi-Modal Coding	Design2Code	**93.4**	92.0	89.2	90.3	92.5	88.9	88.5	85.3
	ChartMimic	78.4	80.5	83.9	79.9	62.1	41.4	**85.2**	82.9
	UniSVG	65.8	69.8	70.0	67.9	71.7	74.5	**73.0**	72.5
Multi-Modal Agent	ScreenSpot Pro	**61.8**	62.0	-	-	-	-	-	-
	OSWorldG	**68.3**	66.7	45.2	-	-	-	-	-
	AndroidWorld	**62.0**	63.7	-	-	-	-	-	-
	OSWorld	**38.1**	31.6	-	-	-	-	-	44.4
	WindowsAA	**32.1**	28.9	-	-	-	-	-	-