

InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models

licence: Apache2.0

1. 프로젝트 개요

- **InternVL** 시리즈의 **3세대 모델**

- 기존 “텍스트→멀티모달 전환” 방식 대신 처음부터 멀티모달·언어 능력을 동시에 학습하도록 설계
- 단일 사전학습 단계에서 다양한 이미지·비디오·텍스트 데이터와 대규모 순수 텍스트 코퍼스를 함께 사용

2. 통합 사전학습 패러다임의 특징

| 문제점 (기존 MLLM) | InternVL3 의 해결책 |
|---|---|
| 텍스트 LLM 후처리 단계에서 정렬(alignment) 오류 다발 | 사전학습 단계에서 멀티모달·언어 동시 학습으로 구조적 정렬 문제 최소화 |
| 멀티스테이지 파이프라인으로 인한 복잡성·최적화 난이도 | 단일 파이프라인 → 모델 구조·훈련 코드 간소화 |

3. 성능·확장성 강화를 위한 핵심 기술

1. V2PE (Variable Visual Position Encoding)

- 입력 이미지/토큰 길이를 가변적으로 지원 → 긴 멀티모달 컨텍스트 처리 능력 향상

2. 고급 후처리 기법

- **SFT(Supervised Fine-Tuning)**: 태스크 특화 지도 학습
- **MPO(Mixed Preference Optimization)**: 인스트럭션·피드백 기반 선호도 혼합 최적화

3. Test-Time Scaling 전략 & 최적화된 인프라

- 인퍼런스 시 동적 스케일링으로 처리량 보존
- 분산 학습 인프라 최적화 → 대규모 모델(78B)도 효율적으로 훈련

4. 실험 결과 하이라이트

- InternVL3-78B

- **MMMU 72.2** 점 → 오픈소스 **MLLM SOTA** 달성
- 상용 최상위 모델 (*ChatGPT-4o, Claude 3.5 Sonnet, Gemini 2.5 Pro*) 과 경쟁적 성능
- 순수 언어 태스크에서도 높은 정확도 유지

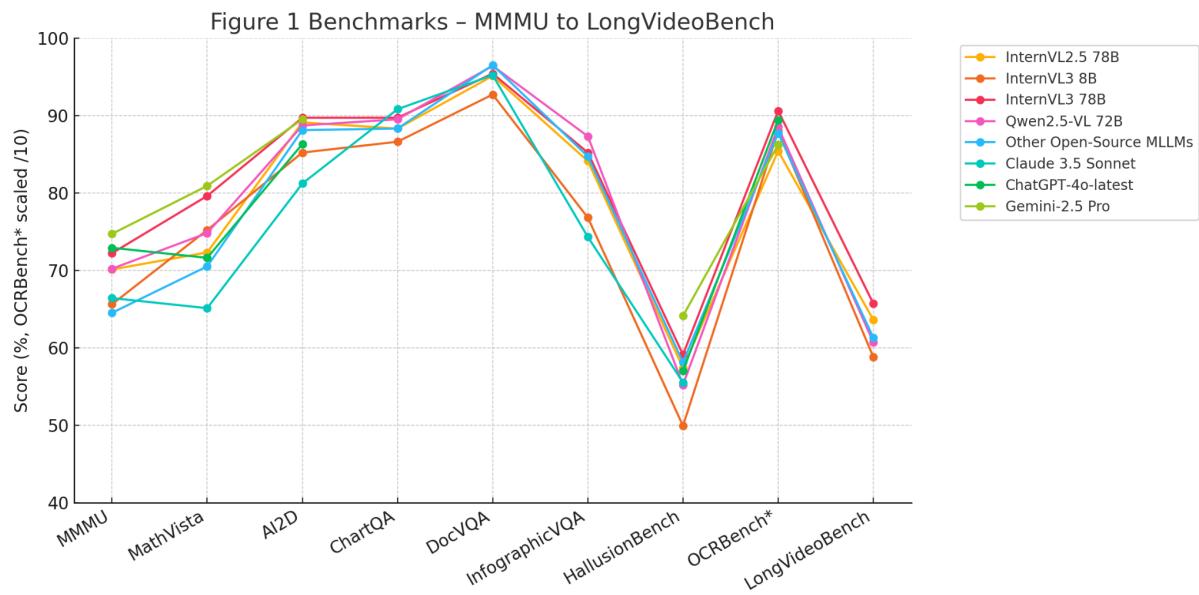
5. 오픈사이언스 기여

- 학습 데이터셋 및 모델 가중치 전면 공개
 - 후속 연구·상용화 실험에 활용 가능
 - 멀티모달 AI 생태계 확장 촉진

| | InternVL2.5 78B | InternVL3 8B | InternVL3 78B | Qwen2.5-VL 72B | Other Open-Source MLLMs | Claude-3.5 Sonnet | ChatGPT- 4o-latest | Gemini-2.5 Pro |
|---|--------------------|-----------------|---------------------|-------------------|-------------------------------|----------------------|-----------------------|-------------------|
| Model Weights | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| Training Data | ✗ | ✓ | ✓ | ✗ | - | ✗ | ✗ | ✗ |
| MMMU Multi-discipline | 70.1% | 65.6% | 72.2% (2.1↑) | 70.2% | 64.5% | 66.4% | 72.9% | 74.7% |
| MathVista Math | 72.3% | 75.2% | 79.6% (7.3↑) | 74.8% | 70.5% | 65.1% | 71.6% | 80.9% |
| AI2D Diagrams | 89.1% | 85.2% | 89.7% (0.6↑) | 88.7% | 88.1% | 81.2% | 86.3% | 89.5% |
| ChartQA Charts | 88.3% | 86.6% | 89.7% (1.4↑) | 89.5% | 88.3% | 90.8% | - | - |
| DocVQA Documents | 95.1% | 92.7% | 95.4% (0.3↑) | 96.4% | 96.5% | 95.2% | - | - |
| InfographicVQA infographics | 84.1% | 76.8% | 85.2% (1.1↑) | 87.3% | 84.7% | 74.3% | - | - |
| HallusionBench Hallucination | 57.4% | 49.9% | 59.1% (1.7↑) | 55.2% | 58.1% | 55.5% | 57.0% | 64.1% |
| OCRBench OCR | 854 | 880 | 906 (52↑) | 885 | 877 | - | 894 | 862 |
| LongVideoBench Video | 63.6% | 58.8% | 65.7% (2.1↑) | 60.7% | 61.3% | - | - | - |

그림 1. InternVL 시리즈와 다른 최신 MLLM들의 멀티모달 성능 비교.

InternVL 시리즈는 멀티모달 역량에서 꾸준한 단계적 향상을 보여 왔습니다. 새로 공개된 **InternVL3**는 기존 오픈소스 MLLM들을 현저히 능가하며, 최첨단 폐쇄형 상용 모델들과 비교해도 여전히 매우 경쟁력 있는 성능을 보입니다.



1. Introduction

1. 배경 & 동기

- 최근 멀티모달 대형 언어 모델(**MLLM**)은 다양한 과제에서 인간 수준, 혹은 그 이상의 성과를 보이며 AGI에 한 걸음 더 다가섰다는 평가를 받음
- 하지만 대부분의 선도 MLLM(오픈·클로즈드 소스 모두)은 텍스트-전용 **LLM**을 다단계 파이프라인으로 “사후(**multistage**) 적응” 하여 시각 입력을 지원하게 만드는 구조를 취함

2. 기존 접근의 한계

| 문제 | 설명 |
|--------------|---|
| 모달리티 정렬 부담 | 텍스트 사전학습 → 시각 정렬 단계로 이어지는 과정에서 언어·시각 간 간극(alignment gap)이 빈번히 발생 |
| 복잡·고비용 파이프라인 | OCR 등 특수 도메인 보조 데이터, 층별 파라미터-프리징, 다중 파인튜닝 스케줄 등 리소스-집약적 전략 필요 |

3. InternVL3 : 네이티브 멀티모달 프리트레이닝

- 핵심 아이디어 : 텍스트-전용 LLM을 “나중에” 시각으로 확장하는 대신, 사전학습 단계에서부터 텍스트 + 다양한 멀티모달 데이터를 함께 투입해 언어·시각 능력을 동시 습득
- 결과적으로 추가 브리징 모듈·후속 정렬 절차가 불필요, 파이프라인이 단순화되고 학습 효율이 향상됨.

4. 핵심 기술 혁신

1. V2PE (Variable Visual Position Encoding)

- 시각 토큰에 가변 위치 증분 δ 를 부여해 긴 멀티모달 컨텍스트 처리 능력 강화

2. 고급 후처리

- SFT(지도식 파인튜닝) + MPO(혼합 선호 최적화)로 대화·추론 성능 상승

3. 테스트-타임 스케일링 & 최적화된 학습 인프라

- 효율적 분산 학습·추론을 지원하는 **InternEVO** 프레임워크 확장 적용

5. 주요 성과 하이라이트

- 폭넓은 벤치마크에서 전 세대(InternVL2.5)를 포함한 오픈소스 MLLM을 일관되게 능가하고,
- **MMMU 72.2** 점으로 오픈소스 SOTA 달성, 클로즈드 상용 모델(ChatGPT-4o, Claude 3.5 Sonnet, Gemini 2.5 Pro 등)과 경쟁적 성능 유지

6. 의의

“처음부터 멀티모달”이라는 통합 사전학습 패러다임으로
복잡·고비용이던 기존 다단계 파이프라인을 단순화하고,
언어·시각 능력을 자연스럽게 결합한 새로운 표준을 제시한다.

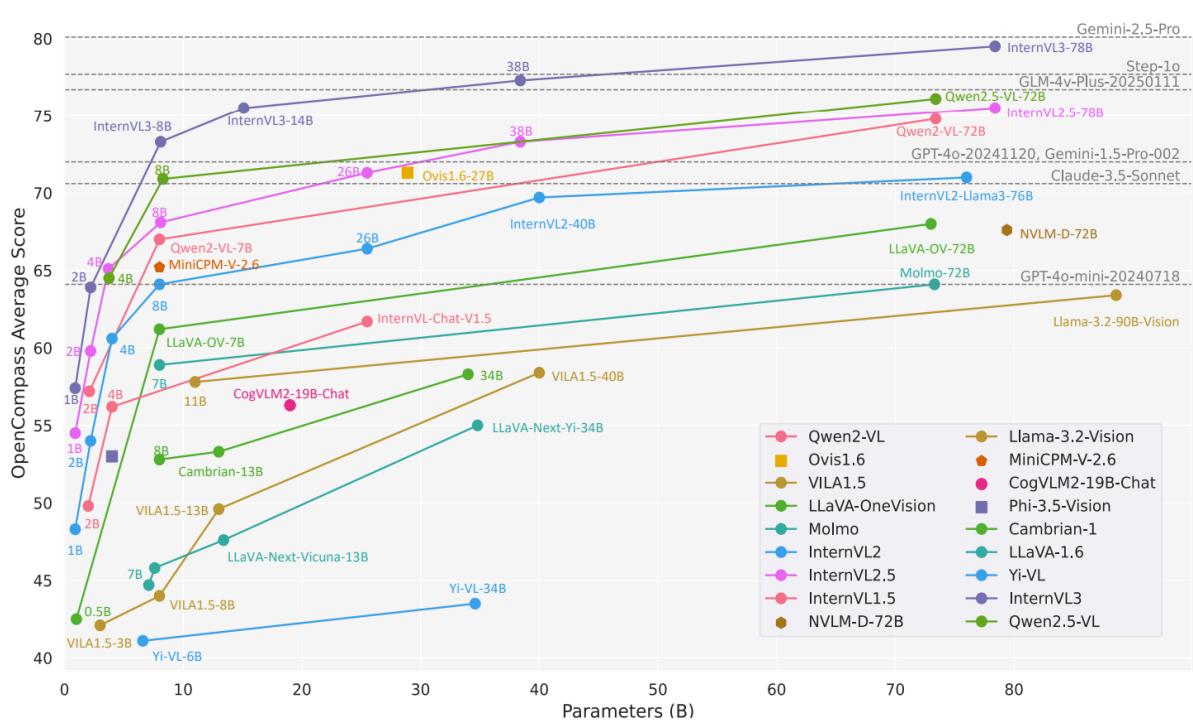


Figure 2는 [OpenCompass](#) 멀티모달 학술 리더보드에서 여러 MLLM(멀티모달 대형 언어 모델)의 성능을 비교한 결과를 보여 줍니다.

- 업그레이드된 **InternVL 시리즈(InternVL3)**는 뛰어난 멀티모달 처리 능력을 증명하며, **Qwen2.5-VL** 계열뿐 아니라 **Step-1o**, **GLM-4v-Plus**, **GPT-4o** 같은 비공개 모델을 모두 큰 폭으로 앞질렀습니다.
- 특히 **InternVL3-78B**는 최신 최고 수준 모델로 꼽히는 **Gemini-2.5-Pro**와도 견줄 만큼 강력한 경쟁력을 유지합니다.

📦 2 InternVL3 핵심 정리

2.1 모델 아키텍처

- **ViT-MLP-LLM** 파라다임 유지: 시각 인코더(InternViT-300M or InternViT-6B) → 2-층 MLP → LLM(Qwen 2.5 시리즈·InternLM3-8B)로 구성
- LLM은 베이스 가중치만 사용(사전 인스트럭션 튜닝 X) → 계산 비용 절감, 언어 능력 보존
- **Pixel Unshuffle**로 448×448 이미지 한장을 256개의 비주얼 토큰으로 축소, 고해상도 처리 효율 향상
- **V2PE(Variable Visual Position Encoding)**: 시각 토큰에 $\delta (< 1)$ 만큼의 작은 위치 증가값을 적용해 **32 K** 토큰까지 긴 멀티모달 컨텍스트 지원

InternVL3 사전학습 모델 라인업

| 모델 | 파라미터 | 비전 인코더 | 언어 모델 | OpenCompass 점수 |
|----------------------|---------------|----------------|---------------|----------------|
| InternVL3-1B | 0.9 B | InternViT-300M | Qwen2.5-0.5 B | 57.4 |
| InternVL3-2B | 1.9 B | InternViT-300M | Qwen2.5-1.5 B | 63.9 |
| InternVL3-8B | 8.1 B | InternViT-300M | Qwen2.5-7 B | 73.3 |
| InternVL3-9B | 9.2 B | InternViT-300M | InternLM3-8 B | 72.4 |
| InternVL3-14B | 15.1 B | InternViT-300M | Qwen2.5-14 B | 75.5 |
| InternVL3-38B | 38.4 B | InternViT-6B | Qwen2.5-32 B | 77.3 |
| InternVL3-78B | 78.4 B | InternViT-6B | Qwen2.5-72 B | 79.5 |

💡 InternVL3 논문 수식 (1) ~ (4)

✓ 요약

- (1) 학습 샘플의 멀티모달 토큰 시퀀스 정의
- (2) 모달리티별 함수를 이용한 재귀적 포지션 계산
- (3) 텍스트·시각 토큰에 증분 차등 적용(**1 vs δ**) → 정렬·컨텍스트 문제 완화

- (4) δ 를 가변적으로 선택해 텍스트 위치 공간을 효율적으로 사용

① 수식 (1) $x = (x_1, x_2, \dots, x_L)$

- 의미
 - 하나의 학습 샘플을 이루는 멀티모달 토큰 시퀀스를 나타냅니다.
 - x_i 는 텍스트 토큰 임베딩, 시각 패치 임베딩(이미지·비디오) 등 모달리티별 표현 모두를 포함합니다.
 - 왜 필요한가?
 - 텍스트·이미지·비디오가 한 줄의 시퀀스로 통합되면 *Transformer*의 자연스러운 자기회귀(**autoregressive**) 학습이 가능해집니다.
-

② 수식 (2) $p_i = \theta, p_i = f_{\text{pos}}(p_{\{i-1\}}, x_i) (i \geq 2)$

- 의미
 - 각 토큰 x_i 의 포지션 인덱스 p_i 를 재귀적으로 계산합니다.
 - f_{pos} 는 토큰 종류(텍스트 vs 시각)에 따라 다르게 동작하는 함수입니다.
 - 핵심 포인트
 - 기존 MLLM은 단순히 +1씩 증가하지만, 여기서는 모달리티별 맞춤 증분을 주어 정렬 오류를 줄입니다.
-

③ 수식 (3) $p_i = p_{\{i-1\}} + \{ 1 (\text{텍스트}), \delta (\text{시각}) \}$

- 의미
 - 텍스트 토큰이면 기존 방식(증분 1) 유지,
 - 시각 토큰이면 더 작은 증분 δ 로 이동해 텍스트 위치 공간을 “빽빽하게” 쓰지 않도록 합니다.
- 장점

- 시각 토큰이 긴 연속 블록으로 들어와도 텍스트 위치 범위를 침범하지 않아,
 - 긴 컨텍스트(> 32k 토큰)에서도 위치 인덱스 오버플로우 위험이 낮음.
 - 텍스트 상대 위치는 그대로 보존돼 언어 이해력 저하를 막습니다.
-

④ 수식 (4) $\delta \in \Delta = \{ 1, 1/2, 1/4, \dots, 1/256 \}$

- 의미
 - 훈련 시: 이미지마다 δ 를 무작위로 선택하여 다양한 길이 시나리오를 학습합니다.
 - 추론 시: 입력 시퀀스 길이에 맞춰 δ 를 동적으로 조정—
 - 짧은 입력엔 $\delta \approx 1$ 로 일반 PE처럼,
 - 긴 입력엔 $\delta \ll 1$ 로 컨텍스트 윈도 창을 절약.
- 결과적 효과
 - 하나의 모델이 다양한 해상도·시퀀스 길이에 유연하게 대응 → 스케일러블 멀티모달 처리.

이 네 식이 합쳐져 **V2PE**를 구성하며, InternVL3가 보다 긴 입력과 복합 모달리티를 안정적으로 다룰 수 있도록 돕습니다.

2.2 네이티브 멀티모달 프리트레이닝

- 언어(**50 B**) + 멀티모달(**150 B**) \Rightarrow 총 **200 B** 토큰을 한 번에 학습해 언어·시각 능력을 동시에 획득
- 표준 좌→우 오토리그레시브 목표, 다만 손실은 텍스트 토큰에만 계산하여 언어 정밀도 유지
- 데이터 구성
 - 기존 InternVL2.5 코퍼스 + GUI / 툴 사용 / 3D Scene / Video 등 실세계 태스크 확장

- 부족한 서사성을 보완하기 위해 대규모 순수 텍스트 자료 추가, 수학·추론 성능 강화

수식 (5) ~ (8) 해설 — InternVL3 논문의 Native Multimodal Pre-Training 부분

핵심 포인트 요약

1. **(5)** 전체 손실: 멀티모달 시퀀스 전체에 대해 자가회귀 방식으로 학습.
2. **(6)** 텍스트-전용 손실: 그래디언트를 텍스트 토큰에만 걸어 이미지·비디오 토큰을 조건으로 활용.
3. **(7)** 가중치 설계: 길이 편향을 줄이기 위해 $1/l^1/\sqrt{l}$ **square averaging**을 사용.
4. **(8)** 공동 최적화: 텍스트·멀티모달 데이터를 하나의 데이터 풀로 보고 하나의 모델을 최적화 → 복잡한 사후 정렬 필요 없음.

■ 수식 (5) : 전체(Full) 자가회귀 학습 목표

$$\mathcal{L}_{\text{full}}(\theta) = - \sum_{i=2}^L w_i \log p_\theta(x_i \mid x_1, \dots, x_{i-1})$$

| 요소 | 의미 |
|-------------------------|--|
| $x = (x_1, \dots, x_L)$ | 텍스트·이미지·비디오 토큰이 뒤섞인 멀티모달 시퀀스 |
| p_θ | 파라미터 θ 를 가진 Transformer가 다음 토큰을 예측하는 확률 |
| w_i | 토큰별 손실 가중치(뒤의 수식 (7)에서 정의) |
| 특징 | 원쪽→오른쪽(autoregressive) 방식으로 모든 토큰에 대해 그래디언트를 전달해 언어·시각 정보를 동시에 학습 |

■ 수식 (6) : 텍스트-전용(Text-only) 손실

$$\mathcal{L}_{\text{text-only}}(\theta) = - \sum_{i=2}^L [x_i \in \text{Text}] w_i \log p_\theta(x_i \mid x_1, \dots, x_{i-1})$$

- 시각 토큰은 예측 대상이 아님 → 조건 컨텍스트로만 사용되어 언어 디코딩을 돋는다.
- 결과적으로 모델은 “이미지를 참고해 텍스트를 생성”하는 능력을 자연스럽게 획득한다.

■ 수식 (7) : 가중치 w_i 설계 — 길이 편향 완화

$$w_i = \begin{cases} 1/l^0 & (\text{token averaging}) \\ 1/l^{0.5} & (\text{square averaging}) \\ 1/l^1 & (\text{sample averaging}) \end{cases}$$

| 전략 | 직관 | 장·단점 |
|-----------------------------------|-------------|-------------------------|
| Token averaging (1) | 토큰마다 동일 가중치 | 길이가 긴 시퀀스에 그래디언트 집중 |
| Sample averaging ($1/l$) | 샘플마다 동일 총합 | 짧은 응답에 편향 |
| Square averaging ($1/\sqrt{l}$) | 중간 절충 | 두 편향을 완화 → 논문은 이 방식을 채택 |

■ 수식 (8) : 멀티태스크 공동 최적화 목표

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{x \sim \mathcal{D}_{\text{multi}}} [\mathcal{L}_{\text{text-only}}(\theta)]$$

- $\mathcal{D}_{\text{multi}}$: 대규모 텍스트 전용·멀티모달(이미지·비디오) 데이터의 합집합
- 모든 파라미터를 한 번에 업데이트
 - 기존 “LLM을 먼저 훈련 → 시각 적응” 파이프라인과 달리, 언어·시각 기능을 처음부터 동시 학습함으로써 별도 정렬 단계를 제거하고 일관된 표현 공간을 형성

이 네 수식은 **InternVL3**의 “처음부터 멀티모달” 사전학습 전략을 수학적으로 뒷받침하며, 언어·시각 표현을 자연스럽게 결합하도록 설계된 핵심 메커니즘입니다.

2.3 사후 학습(Post-Training)

| 단계 | 핵심 아이디어 | 효과 |
|------------|---|-------------------------------------|
| SFT | JPEG 압축·멀티모달 패킹·스퀘어 로스 가중치 + 21.7 M 고품질 샘플 | 대화·도구 사용·GUI 조작 등 대화형 성능 ↑ |
| MPO | DPO + BCO + LM Loss를 결합해 선호·품질·생성을 동시 최적화 | 7개 추론 벤치마크에서 최대 +4.5 점 향상 |

InternVL3 MPO 학습 손실 수식 (9)–(13) 해설

(원문 *InternVL3* 논문 2.3절 “Mixed Preference Optimization” 중)

🔑 요약

- (9) **MPO** 총 손실: 선호·품질·생성 세 축을 통합.
- (10) **DPO** 선호 손실: “선택 > 거부” 상대적 선호를 학습.
- (11) 품질 손실을 선택/거부로 분리해 절대 품질을 평가.
- (12) 선택 응답이 기준 $\delta\backslash\delta$ 이상이 되도록 강화.
- (13) 거부 응답은 기준 이하로 떨어지게 억제.

수식 (9) 총 손실 정의

$$\mathcal{L} = w_p \mathcal{L}_p + w_q \mathcal{L}_q + w_g \mathcal{L}_g$$

- 의미 : MPO 단계에서 모델이 동시에
 1. 선호(**preference**),
 2. 응답 품질(**quality**),
 3. 텍스트 생성(**generation**)
 능력을 학습하도록 세 가지 손실을 가중 합산한 전체 목적 함수입니다.
- w_p, w_q, w_g : 각 손실 항에 부여되는 **가중치 하이퍼파라미터**로, 학습 중 세 항의 상대적 중요도를 조절합니다.

수식 (10) 선호 손실 \mathcal{L}_p (DPO 기반)

$$\mathcal{L}_p = - \log \sigma \left(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_0(y_c|x)} - \beta \log \frac{\pi_\theta(y_r|x)}{\pi_0(y_r|x)} \right)$$

- DPO(Direct Preference Optimization) 을 그대로 차용.
- x : 사용자 질문,
- y_c / y_r : 선택된(chosen) / 거부된(rejected) 응답.
- π_θ : 현재 정책 모델,
- π_0 : 초기 준거(reference) 모델.
- β : KL 규제 계수로, 업데이트 폭을 안정화합니다.
- 로지스틱 함수 σ 는 선택 > 거부 관계를 확률적으로 강제하며, 차이가 클수록 손실이 작아져 선호 방향으로 학습됩니다.

수식 (11) 품질 손실 \mathcal{L}_q 분해

$$\mathcal{L}_q = \mathcal{L}_q^+ + \mathcal{L}_q^-$$

- 응답 개별 품질의 “절대적” 우수·열등 여부를 학습하도록, 선택 응답과 거부 응답을 각각 평가하는 두 항으로 분리합니다.

수식 (12) 선택 응답 품질 손실 \mathcal{L}_q^+

$$\mathcal{L}_q^+ = - \log \sigma \left(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_0(y_c|x)} - \delta \right)$$

- δ : 보상 이동 상수(reward shift) 의 이동 평균으로, 훈련 초기에 값이 요동치는 것을 완화해 안정적인 학습을 돋습니다.
- 모델이 선택 응답의 품질 점수 $\geq \delta$ 가 되도록 유도합니다.

수식 (13) 거부 응답 품질 손실 \mathcal{L}_q^-

$$\mathcal{L}_q^- = - \log \sigma \left(- \left(\beta \log \frac{\pi_\theta(y_r|x)}{\pi_0(y_r|x)} - \delta \right) \right)$$

- 부호가 반전되어 있어 거부 응답의 품질 점수 $\leq \delta$ 로 수렴하도록 만듭니다.
- 결국 $\mathcal{L}_q^+, \mathcal{L}_q^-$ 은 절대 품질 판별을 학습해 모델이 좋은 답을 올리고 나쁜 답을 억제하도록 보강합니다.

이 구조 덕분에 InternVL3는 선호 일관성과 절대 품질을 동시에 최적화하며, 언어·멀티모달 추론 능력 모두에서 성능을 끌어올릴 수 있습니다.

2.4 테스트-타임 스케일링

- Best-of-N(N = 8) 샘플링 후 **VisualPRM-8B** 비평 모델로 최고 응답 선택 → 수학·추론 벤치마크 추가 상승

▶ 수식 (14) — Visual Process Reward Model

$$c_i \sim M(y_i | I, q, s_{\leq i}) \quad (14)$$

① 수식의 의미

- **VisualPRM-8B** 모델 M 이 이미지 I , 질문 q , 그리고 현재까지의 풀이 단계 집합 $s_{\leq i}$ 를 조건으로 i -번째 단계의 정답 여부를 추정한다는 식입니다.
- 결과 c_i 는 두 값 중 하나를 택합니다.
 - $+$: 올바른 단계
 - $-$: 잘못된 단계
- 학습 시에는 각 단계별로 $+$ / $-$ 레이블을 예측하도록 모델을 훈련하며, 추론 시에는 $+$ 를 생성할 확률을 단계 점수로 사용합니다.

② 기호 설명

| 기호 | 설명 |
|---------|-----------------------------------|
| c_i | i -번째 단계의 정답 레이블 ($+$ 또는 $-$) |
| $M(\#)$ | VisualPRM 모델의 조건부 확률 분포 |
| y_i | 모델이 예측하려는 정답 토큰 (레이블) |
| I | 문제에 제시된 이미지 |

| | |
|--------------|--|
| q | 해당 이미지와 함께 주어지는 질문 |
| $s_{\leq i}$ | 첫 단계 s_0 부터 현재 단계 s_i 까지의 누적 풀이 과정 |

3 문맥 — VisualPRM의 역할

1. 다단계 채점(critic) 모델

- InternVL3는 추론·수학 문제에서 **Best-of-N** 전략을 쓰는데, VisualPRM이 각 후보 해설의 단계별 품질을 평가해 최종 해설을 고릅니다.

2. 멀티턴 Chat 포맷

- (I , q , s_0)를 첫 턴에 주고, 이후 턴마다 새로운 단계 s_i 입력 → 모델이 c_i 예측.

3. 스코어 산출

- 추론 단계에서 “+” 확률을 단계 점수로 삼아 전체 해설의 평균 점수를 계산합니다.

4 왜 중요한가?

- 세밀한 단계별 보상을 통해 모델이 장문 추론 체인의 논리적 일관성을 학습할 수 있게 합니다.
- 결과적으로 InternVL3는 수학·추론 벤치마크에서 높은 성능을 달성하는 데 큰 기여를 했습니다.

2.5 학습 인프라 & 스케일링

- **InternEVO 2.0:** ZeRO 기반 프레임워크를 ViT·MLP·LLM 모듈 단위 **Sharding**으로 확장, 데이터·텐서·시퀀스·파이프라인 병렬 자유 조합
- 동적 로드밸런싱으로 ViT·LLM 간 계산 편차 해소, 32 K 컨텍스트 지원
- 동일 자원 기준 훈련 속도 **50 % ~ 200 %** 가속(InternVL2.5 대비)

3 Experiments

3.1 전체 MLLM 비교 (Overall Comparison)

- 폭넓은 벤치마크 우위 : MMMU 72.2 점으로 오픈소스 최고, AI2D·ChartQA 등 일부 항목은 GPT-4o · Claude 3.5보다도 앞섰다
- 향상 원인
 1. 네이티브 멀티모달 사전학습
 2. **V2PE·MPO** 등 후처리 기법
 3. 확대된 고품질 데이터 코퍼스

3.2 멀티모달 추론 & 수학

| 주요 지표 | InternVL3-78B | 개선폭 |
|------------|----------------------|-----------------|
| MMMU | 72.2 | +2 ~ 5 pt 기준 대비 |
| MathVista | 79.0 | +4 ~ 7 pt |
| LogicVista | 46.1 | +6 pt 이상 |

- 스케일링 효과 : 파라미터 증가와 동시에 모든 수리·논리 벤치마크가 상승
- **Best-of-N(Test-time scaling)** : Bo-8 채택 시 MathVerse-VisionOnly 최대 +6 pt 상승, 소형 모델도 이득

- 실제 환경(**Real-world**) 벤치마크: RealWorldQA [27], MME-RealWorld [151], WildVision [86], RBench [62]
 - 일부 결과는 각 벤치마크 논문과 OpenCompass 리더보드 [26]에서 인용됨.
-

3.5 실 세계 이해 (**Real-World Comprehension**)

| 모델 | RealWorldQA | WildVision(win) | R-Bench |
|------------|-------------|-----------------|-------------|
| 1B | 58.2 | 43.8 | 60.4 |
| 78B | 78.0 | 73.6 | 77.4 |

- GPT-4o 대비 RealWorldQA +2.6 pt, MME-RW에서도 우세
-

3.6 종합 멀티모달 평가 (**MME · MMBench** 등)

- **MME 합계 2 549.8 (78B)**, MMBench(EN/CN) 89.0 / 88.7 → Qwen2.5-VL-72B보다 높은 일관 성능
- 스케일 별로 꾸준히 우상향, 멀티태스크 균형 양호

- HallusionBench 59.1 (78B) · CRPE 79.2로 이전 세대 대비 개선
 - 관찰 : 대규모에서도 일부 MMHai 감소 등 미세한 편차 존재 → 데이터/학습 전략 추가 정교화 필요
-

3 Experiments 섹션 요약 (2/2)

범위 3.8 ~ 3.14 – 앞선 3.1 ~ 3.7과 동일한 형식으로 정리했습니다.

3.8 시각적 그라운딩 (Visual Grounding)

모델 RefCOCO / + / g 평균

InternVL3-2B 86.7

InternVL3-14B 89 ± 0.5

InternVL3-78B 91.4 (↑)

- 스케일 효과: 1B → 14B까지 꾸준히 상승, 그러나 78B에서 증가폭이 둔화 –
추가적인 지시·바운딩박스 데이터가 부족했기 때문으로 추정
- 여전히 오픈소스 최고 성능이지만, 전작 InternVL2.5-78B (92.3)와의 격차는 과제별
데이터 특화 여부에 따라 달라짐.

| Model Name | RefCOCO | | | RefCOCO+ | | | RefCOCOg | | Overall |
|---------------------------|---------|--------|--------|----------|--------|--------|----------|------|---------|
| | val | test-A | test-B | val | test-A | test-B | val | test | |
| Grounding-DINO-L [74] | 90.6 | 93.2 | 88.2 | 82.8 | 89.0 | 75.9 | 86.1 | 87.0 | 86.6 |
| UNINEXT-H [133] | 92.6 | 94.3 | 91.5 | 85.2 | 89.6 | 79.8 | 88.7 | 89.4 | 88.9 |
| ONE-PEACE [122] | 92.6 | 94.2 | 89.3 | 88.8 | 92.2 | 83.2 | 89.2 | 89.3 | 89.8 |
| Qwen2.5-VL-3B [6] | 89.1 | 91.7 | 84.0 | 82.4 | 88.0 | 74.1 | 85.2 | 85.7 | 85.0 |
| InternVL3-1B | 85.8 | 90.1 | 81.7 | 76.6 | 84.1 | 69.2 | 82.8 | 82.6 | 81.6 |
| InternVL3-2B | 89.8 | 92.6 | 86.4 | 84.0 | 89.2 | 76.5 | 87.6 | 87.2 | 86.7 |
| Shikra-7B [12] | 87.0 | 90.6 | 80.2 | 81.6 | 87.4 | 72.1 | 82.3 | 82.2 | 82.9 |
| Ferret-v2-13B [144] | 92.6 | 95.0 | 88.9 | 87.4 | 92.1 | 81.4 | 89.4 | 90.0 | 89.6 |
| CogVLM-Grounding [123] | 92.8 | 94.8 | 89.0 | 88.7 | 92.9 | 83.4 | 89.8 | 90.8 | 90.3 |
| MM1.5 [143] | — | 92.5 | 86.7 | — | 88.7 | 77.8 | — | 87.1 | — |
| Qwen2-VL-7B [121] | 91.7 | 93.6 | 87.3 | 85.8 | 90.5 | 79.5 | 87.3 | 87.8 | 87.9 |
| Qwen2.5-VL-7B [6] | 90.0 | 92.5 | 85.4 | 84.2 | 89.1 | 76.9 | 87.2 | 87.2 | 86.6 |
| TextHawk2 [140] | 91.9 | 93.0 | 87.6 | 86.2 | 90.0 | 80.4 | 88.2 | 88.1 | 88.2 |
| InternVL2-8B [19] | 87.1 | 91.1 | 80.7 | 79.8 | 87.9 | 71.4 | 82.7 | 82.7 | 82.9 |
| InternVL2.5-8B [18] | 90.3 | 94.5 | 85.9 | 85.2 | 91.5 | 78.8 | 86.7 | 87.6 | 87.6 |
| InternVL3-8B | 92.5 | 94.6 | 88.0 | 88.2 | 92.5 | 81.8 | 89.6 | 90.0 | 89.6 |
| InternVL3-9B | 91.8 | 93.2 | 86.6 | 86.4 | 91.0 | 79.9 | 88.0 | 88.5 | 88.2 |
| InternVL3-14B | 92.0 | 94.4 | 87.8 | 87.4 | 92.1 | 81.5 | 88.6 | 89.3 | 89.1 |
| Qwen2-VL-72B [121] | 93.2 | 95.3 | 90.7 | 90.1 | 93.8 | 85.6 | 89.9 | 90.4 | 91.1 |
| Qwen2.5-VL-72B [6] | 92.7 | 94.6 | 89.7 | 88.9 | 92.2 | 83.7 | 89.9 | 90.3 | 90.3 |
| InternVL2-Llama3-76B [19] | 92.2 | 94.8 | 88.4 | 88.8 | 93.1 | 82.8 | 89.5 | 90.3 | 90.0 |
| InternVL2.5-78B [18] | 93.7 | 95.6 | 92.5 | 90.4 | 94.7 | 86.9 | 92.7 | 92.2 | 92.3 |
| InternVL3-38B | 93.2 | 95.1 | 90.2 | 89.8 | 93.2 | 85.2 | 91.4 | 91.5 | 91.2 |
| InternVL3-78B | 93.4 | 95.4 | 90.3 | 90.1 | 93.8 | 85.3 | 91.5 | 91.5 | 91.4 |

Table 6: Comparison of visual grounding performance. We evaluate InternVL’s visual grounding capability on RefCOCO, RefCOCO+, and RefCOCOg datasets [56, 88]. Parts of the results are collected from [121].

3.9 멀티모달 다국어 이해

- **MMMB + Multilingual MMBench + MTVQA 평균**
 - **InternVL3-78B 88.7 점 → Qwen2-VL-72B (85.8)보다 우수**
- 1B-2B 소형 모델에서도 전세대 대비 1-2 pt 개선 → 네이티브 멀티모달 사전학습이 다국어 표현·시각 정합성에 기여.

| Model Name | Video-MME (wo / w sub) | MVBench | MMBench-Video (val) | MLVU (M-Avg) | LongVideoBench (val total) | CG-Bench (long / clue acc.) | Overall |
|---------------------------|---------------------------|---------|------------------------|-----------------|-------------------------------|--------------------------------|---------|
| InternVL2-1B [19] | 42.9 / 45.4 | 57.5 | 1.14 | 51.6 | 43.3 | — | — |
| InternVL2.5-1B [18] | 50.3 / 52.3 | 64.3 | 1.36 | 57.3 | 47.9 | — | — |
| InternVL3-1B | 51.0 / 53.0 | 63.1 | 1.3 | 53.0 | 48.1 | 24.8 / 39.1 | 46.9 |
| Qwen2-VL-2B [121] | 55.6 / 60.4 | 63.2 | — | — | — | — | — |
| Qwen2.5-VL-3B [7] | 61.5 / 67.6 | 67.0 | 1.63 | 68.2 | 43.3 | — | — |
| InternVL2-2B [19] | 46.2 / 49.1 | 60.2 | 1.30 | 54.3 | 46.0 | — | — |
| InternVL2.5-2B [18] | 51.9 / 54.1 | 68.8 | 1.44 | 61.4 | 52.0 | — | — |
| InternVL3-2B | 58.9 / 61.4 | 70.4 | 1.42 | 64.2 | 55.4 | 30.8 / 50.7 | 54.9 |
| VideoChat2-HD [64] | 45.3 / 55.7 | 62.3 | 1.22 | 47.9 | — | — | — |
| MiniCPM-V-2.6 [135] | 60.9 / 63.6 | — | 1.70 | — | 54.9 | — | — |
| LLaVA-OneVision-7B [60] | 58.2 / — | 56.7 | — | — | — | — | — |
| Qwen2-VL-7B [121] | 63.3 / 69.0 | 67.0 | 1.44 | — | 55.6 | — | — |
| Qwen2.5-VL-7B [7] | 65.1 / 71.6 | 69.6 | 1.79 | 70.2 | 45.3 | — | — |
| InternVL2-8B [19] | 56.3 / 59.3 | 65.8 | 1.57 | 64.0 | 54.6 | — | — |
| InternVL2.5-8B [18] | 64.2 / 66.9 | 72.0 | 1.68 | 68.9 | 60.0 | — | — |
| InternVL3-8B | 66.3 / 68.9 | 75.4 | 1.69 | 71.4 | 58.8 | 38.6 / 55.2 | 61.4 |
| InternVL3-9B | 66.7 / 68.9 | 74.3 | 1.69 | 70.8 | 62.5 | 41.1 / 58.0 | 62.3 |
| InternVL3-14B | 70.4 / 73.0 | 76.6 | 1.73 | 73.3 | 63.9 | 44.1 / 60.6 | 64.9 |
| InternVL2-26B [19] | 57.0 / 60.2 | 67.5 | 1.67 | 64.2 | 56.1 | — | — |
| InternVL2.5-26B | 66.9 / 69.2 | 75.2 | 1.86 | 72.3 | 59.9 | — | — |
| Oryx-1.5-32B [78] | 67.3 / 74.9 | 70.1 | 1.52 | 72.3 | — | — | — |
| Qwen2.5-VL-32B [7] | 70.5 / 77.9 | — | 1.93 | — | — | — | — |
| VILA-1.5-40B [71] | 60.1 / 61.1 | — | 1.61 | 56.7 | — | — | — |
| InternVL2-40B [19] | 66.1 / 68.6 | 72.0 | 1.78 | 71.0 | 60.6 | — | — |
| InternVL2.5-38B [18] | 70.7 / 73.1 | 74.4 | 1.82 | 75.3 | 63.3 | — | — |
| InternVL3-38B | 72.7 / 75.0 | 76.9 | 1.81 | 77.8 | 67.3 | 46.9 / 62.8 | 67.5 |
| GPT-4V/4T [1] | 59.9 / 63.3 | 43.7 | 1.53 | 49.2 | 59.1 | — | — |
| GPT-4o-20240513 [97] | 71.9 / 77.2 | — | 1.63 | 64.6 | 66.7 | — | — |
| GPT-4o-20240806 [97] | — | — | 1.87 | — | — | 41.8 / 58.3 | — |
| Gemini-1.5-Pro [102] | 75.0 / 81.3 | — | 1.30 | — | 64.0 | 40.1 / 56.4 | — |
| VideoLLaMA2-72B [23] | 61.4 / 63.1 | 62.0 | — | — | — | — | — |
| LLaVA-OneVision-72B [60] | 66.2 / 69.5 | 59.4 | — | 66.4 | 61.3 | — | — |
| Qwen2-VL-72B [121] | 71.2 / 77.8 | 73.6 | 1.70 | — | — | 41.3 / 56.2 | — |
| Qwen2.5-VL-72B [7] | 73.3 / 79.1 | 70.4 | 2.02 | 74.6 | 60.7 | — | — |
| InternVL2-Llama3-76B [19] | 64.7 / 67.8 | 69.6 | 1.71 | 69.9 | 61.1 | — | — |
| InternVL2.5-78B [18] | 72.1 / 74.0 | 76.4 | 1.97 | 75.7 | 63.6 | 42.2 / 58.5 | 66.0 |
| InternVL3-78B | 72.7 / 75.7 | 78.7 | 1.81 | 79.5 | 65.7 | 48.4 / 65.3 | 68.3 |

Table 8: Comparison of video understanding performance. We evaluate InternVL’s video understanding capabilities across 6 benchmarks. For Video-MME [38], MMBench-Video [35], MLVU [154], and LongVideoBench [129], we test with four different settings: 16, 32, 48, and 64 frames, and report the maximum results. For MVBench [65], we conduct testing using 16 frames. For CG-Bench [2], we use 32 frames.

3.11 GUI 그라운딩

- **ScreenSpot:** InternVL3-72B 88.7 % → GPT-4o 18.1 % 대비 대폭 상회
- **ScreenSpot-V2:** 72B 90.9 %, 38B 88.3 %, 8B 81.4 % → 모델 크기 ↑ → 복잡한 UI 배치 파악 능력 ↑
- Aguvis-72B만 소폭 앞설 뿐, 오픈소스 최고 레벨 유지.

| Method | GPT-4o | Gemini 2.0 | Claude | Aguvis-72B | Qwen2.5-VL-72B | UI-TARS-72B | InternVL3-8B | -38B | -72B |
|---------------|--------|------------|--------|-------------|----------------|-------------|--------------|------|-------------|
| ScreenSpot | 18.1 | 84.0 | 83.0 | 89.2 | 87.1 | 88.4 | 79.5 | 85.6 | 88.7 |
| ScreenSpot-V2 | — | — | — | — | — | 90.3 | 81.4 | 88.3 | 90.9 |

Table 9: Performance of InternVL3 and other models on GUI grounding benchmarks.

3.12 공간 추론 (Spatial Reasoning)

- **VSI-Bench** 종합
 - 8B 42.1 → **38B 48.9 / 78B 48.4** – GPT-4o·Gemini-1.5 Pro보다 높은 점수
- 서브테스크: 객체 개수 71.2, 상대 거리 55.9 등 세밀한 **3D** 장면 이해에서도 강세.

| Model Name | Obj.count | Abs.Dist. | Obj.size | Room Size | Rel.Dist. | Rel.Dir. | Route Plan | Appr.Order | Overall |
|----------------------------|-----------|-----------|----------|-----------|-----------|----------|------------|------------|---------|
| GPT-4o [97] | 46.2 | 5.3 | 43.8 | 38.2 | 37.0 | 41.3 | 31.5 | 28.5 | 34.0 |
| Gemini-1.5 Pro [102] | 56.2 | 30.9 | 64.1 | 43.6 | 51.3 | 46.3 | 36.0 | 34.6 | 45.4 |
| VILA-1.5-8B [71] | 17.4 | 21.8 | 50.3 | 18.8 | 32.1 | 34.8 | 31.0 | 24.8 | 28.9 |
| LongVA-7B [145] | 38.0 | 16.6 | 38.9 | 22.2 | 33.1 | 43.3 | 25.4 | 15.7 | 29.2 |
| LLaVA-NeXT-Video-7B [150] | 48.5 | 14.0 | 47.8 | 24.2 | 43.5 | 42.4 | 34.0 | 30.6 | 35.6 |
| LLaVA-OneVision-7B [60] | 47.7 | 20.2 | 47.4 | 12.3 | 42.5 | 35.2 | 29.4 | 24.4 | 32.4 |
| InternVL3-8B | 68.1 | 39.0 | 48.4 | 33.6 | 48.3 | 36.4 | 27.3 | 35.4 | 42.1 |
| InternVL3-38B | 71.7 | 50.2 | 46.1 | 41.7 | 53.5 | 38.6 | 28.9 | 60.7 | 48.9 |
| LLaVA-NeXT-Video-72B [150] | 48.9 | 22.8 | 57.4 | 35.3 | 42.4 | 36.7 | 35.0 | 48.6 | 40.9 |
| LLaVA-OneVision-72B [60] | 43.5 | 23.9 | 57.6 | 37.5 | 42.5 | 39.9 | 32.5 | 44.6 | 40.2 |
| InternVL3-78B | 71.2 | 53.7 | 44.4 | 39.5 | 55.9 | 39.5 | 28.9 | 54.5 | 48.4 |

Table 10: Performance of InternVL3 and other models on VSI-Bench.

3.13 언어 능력 평가

| 벤치마크 | Qwen2.5-72B-Chat | InternVL3-78B |
|-----------|------------------|---------------------|
| MMLU | 84.4 | 86.9 |
| CMMLU | 87.4 | 89.9 |
| C-Eval | 88.1 | 89.5 |
| GSM8K | 88.2 | 90.5 |
| HumanEval | 87.2 | 82.3 (근소 열세) |

- **25 %** 순수 텍스트 데이터를 포함한 네이티브 사전학습 + **SFT**가 언어 지표 전반 향상을 견인

이는 단일 사전학습 + 정교한 사후 튜닝만으로도 SOTA 혹은 상용 모델급 경쟁력을 확보할 수 있음을 보여줍니다.

4 Conclusion

| 요점 | 설명 |
|---------------------|--|
| InternVL3 소개 | InternVL 시리즈의 세 번째 모델로, 처음부터 멀티모달·언어 능력을 동시에 학습하는 ‘네이티브 멀티모달 사전학습’ 방식을 채택했습니다. |
| 후처리 복잡성 해소 | 전통적인 “텍스트 LLM → 멀티모달 전환” 파이프라인에서 발생하던 정렬(Alignment)·최적화 난제를 사전학습 단계에서 해결해 학습 과정을 단순화했습니다. |
| 핵심 기술 요소 | 1) V2PE (Variable Visual Position Encoding) → 긴 이미지/비디오 입력도 처리 2) SFT (지도 파인튜닝) & MPO (혼합 선호 최적화) → 실 사용 태스크 적합도 향상 3) Test-Time Scaling → 추론 단계에서 출력 품질을 추가로 끌어올림 |
| 성능 하이라이트 | 78B 파라미터 모델이 MMMU 72.2 점을 기록해 오픈소스 MLLM 최고 성능을 달성, Gemini-2.5 Pro 같은 상용 모델과의 격차를 크게 줄였습니다. |
| 언어 능력 유지 | 멀티모달 성능이 향상되면서도 순수 텍스트 과제에서의 언어 이해·생성 능력을 그대로 보존했습니다. |
| 오픈사이언스 선언 | 학습 데이터와 모델 가중치를 공개해 커뮤니티 주도의 연구·개발을 촉진하겠다고 명시했습니다. |

한마디 요약

InternVL3는 “사전학습 단계에서 멀티모달을 원천적으로 통합”한 덕분에, 복잡한 후처리 없이도 오픈소스 최고 수준의 멀티모달 성능과 탄탄한 언어 능력을 동시에 확보했으며, 모델과 데이터를 공개해 생태계 발전에 기여하겠다는 의지를 밝히고 있습니다.