

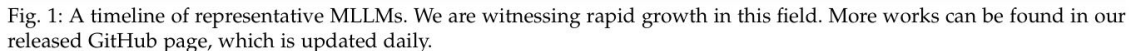
Deepseek Multimodal 모델 논문 리뷰 (Janus, Janus-Pro)



Preliminary Information

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

Shukang Yin*, Chaoyou Fu*†, Sirui Zhao*, Ke Li,
Xing Sun, Tong Xu, and Enhong Chen, *Fellow, IEEE*



Yizhang Jin^{1,2,*}, Jian Li^{1,*}, Yexin Liu³, Tianjun Gu⁴, Kai Wu¹, Zhengkai Jiang¹,
Muyang He³, Bo Zhao³, Xin Tan⁴, Zhenye Gan¹, Yabiao Wang¹, Chengjie Wang¹,
Lizhuang Ma²

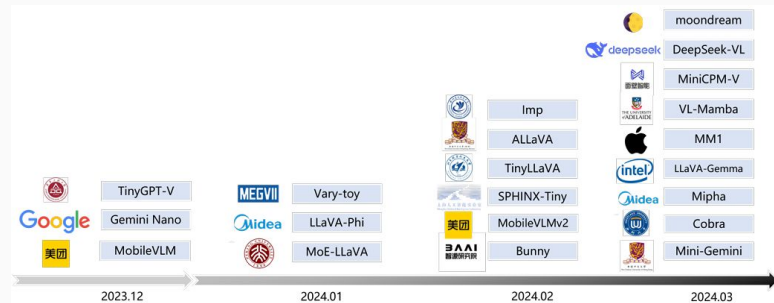
¹Youtu Lab, Tencent, ²SJTU, ³BAAI, ⁴ECNU

Figure 1: The timeline of efficient MLLMs.

Multimodal Model(1)

멀티모달 모델에서 입력과 출력의 형태에 따른 예시

Multimodal Model vs Multimodal LLM

참고. CLIP(Contrastive Language-Image Pre-training): 생성보다는 이해, 분류, 검색에 특화된 모델, 이미지와 텍스트를 같은 임베딩 공간으로 매핑

모델	주요 특징	입력 (Input)	출력 (Output)	주요 차이점
Janus (DeepSeek, 2024)	텍스트+이미지 이해 & 생성 통합 모델 <small>cf. Chameleon</small>	텍스트 + 이미지	텍스트 + 이미지	시각적 인코딩을 이해와 생성으로 분리하여 최적화
GPT-4V (OpenAI, 2023)	GPT-4 기반의 비전 멀티모달 모델	텍스트 + 이미지	텍스트	이미지 생성 불가능, 이해만 가능
Flamingo (DeepMind, 2022)	Few-shot 학습이 가능한 이미지-텍스트 모델	텍스트 + 이미지	텍스트	대화형 AI에 최적화, 이미지 생성 불가능
BLIP-2 (Salesforce, 2023)	강력한 이미지 캡셔닝 & 질문 응답 모델	텍스트 + 이미지	텍스트	이미지 생성 불가능, 이미지 설명에 강함
Kosmos-2 (Microsoft, 2023)	멀티모달 학습을 강화한 비전-언어 모델	텍스트 + 이미지	텍스트	강력한 시각적 질문 응답, 생성보다는 이해 중심
IDEFICS (Hugging Face, 2023)	오픈소스 기반 비전-언어 모델	텍스트 + 이미지	텍스트	멀티모달 대화에 특화, 이미지 생성 불가능
DALL·E 3 (OpenAI, 2023)	<small>cf. stable diffusion, midjourney</small> 텍스트에서 고해상도 이미지 생성	텍스트	이미지	이미지 이해 불가능, 생성만 가능
DeepFloyd IF (Stability AI, 2023)	Diffusion 기반 텍스트-이미지 생성	텍스트	이미지	텍스트에서 이미지 생성, 이해 불가능

※ Janus와 견줄만한 모델로 VL-GPT 있었지만 현재는 종료된 상태이고 SEED-X로 이어져 오픈 소스로 공개되어 있음.



TencentAILab-CVC
Tencent AI Lab - Computer Vision Center

Multimodal Model(2)

Multimodal 방식에 따른 분류

방식	대표 모델	주요 특징	장점	단점	적용 분야
Autoregressive (AR)	DALL·E 2, SEED-X	Transformer 기반 자기회귀 방식	텍스트 이해력 우수, 문맥 고려 가능	생성 속도 느림, 긴 문장에서 오류 가능	텍스트-이미지 생성, 대화형 AI
Diffusion	Stable Diffusion, Imagen	노이즈 제거 방식 (Denoising-based)	초고해상도 이미지 생성 가능, 자연스러운 출력	계산량 많음, 생성 속도 느림	초고해상도 이미지 생성, 예술 AI
Two-Stream	Flamingo, Kosmos-2	인코더-디코더 분리형 구조	학습 효율적, 다양한 모달리티 처리 가능	생성 성능은 AR/Diffusion보다 낮음	멀티모달 이해, 질문응답 시스템
Autoencoding (AE)	MAE, BEiT	비지도 학습 기반 인코딩	데이터 효율적 활용 가능, 대량의 학습 가능	직접적인 생성 불가능	이미지-텍스트 이해, 사전 학습 모델
Retrieval-Augmented (RAG)	BLIP, REVEAL, KATANA	벡터 DB에서 관련 정보 검색 후 생성	정밀한 정보 제공, 환각 (Hallucination) 최소화	창의적인 생성 어려움, 기존 데이터 의존	이미지-텍스트 검색, 정보 기반 대화 AI
Hybrid (AR + Diffusion)	DeepFloyd IF, DALL·E 3	AR 기반 문맥 이해 + Diffusion 기반 고해상도 생성 조합	텍스트 조건 반영하면서도 고품질 이미지 생성 가능	연산 비용 큼, 복잡한 모델 구조	고품질 텍스트-이미지 생성, 창작 AI

※ Janus도 Autoregressive 기반

Multimodal 이해와 생성

Image Understanding Tasks

What is my cat doing?



MLLM (LLaVA)

The cat is trying to catch a fish

Example for image understanding tasks, solved by MLLM

Image Generation Tasks

A cute cat



Diffusion Model
(Stable Diffusion, DALL-E 3)



Example for an image generation task

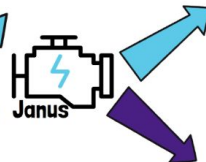
Unified Model

What is my cat doing?



The cat is trying to catch a fish

A cute cat



Unified handling of understanding and generation tasks with Janus

<https://aipapersacademy.com/janus-pro/>

Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation

통합된 멀티모달 이해 및 생성을 위한 시각적 인코딩
분리

Abstract

- 멀티모달 이해 (multi-modal understanding)와 생성 (generation)을 통합하는
Autoregressive 프레임워크

- 기존 모델들

- 이미지 이해와 생성에 단일 visual encoder(예. Chameleon) 적용

META에서 개발한 멀티모달 모델

→ 요구되는 정보의 **세분화 수준 (granularity)**이 다르기 때문에 multimodal understanding

측면에서 최적의 성능내기 어려움

- 이해 작업에서 비주얼 인코더의 목적은 이미지에서 고수준의 의미론적 정보(예: 객체 범주나 시각적 속성)를

추출

예시. 이 사진에 있는 동물은 무엇인가요?라는 질문에 답하기 위해서는 이미지에서 동물의 종류, 특징, 주변

환경의 맥락 생성 작업에서는 주로 이미지의 지역적 세부 사항을 생성하고 전체적 일관성을 유지하는 데 초점,

미세하고 고수준의 의미 정보를 추출

구조와 텍스트 세부 사항을 표현할 수 있는 저차원 인코딩이 필요

예시. "해변에서 서핑하는 강아지" 이미지를 생성할 때는 강아지의 털 질감, 파도의 세부적인 모양, 모래의

결합

저수준의 시각적 세부 사항이 중요

- 이를 해결하기 위해,

비주얼 인코딩을 별도의 경로로 분리하면서 단일의 통합된 트랜스포머 아키텍처를 활용

(멀티모달 이해와 생성 구성 요소 모두 각자에게 가장 적합한 인코딩 방법을 독립적으로 선택)

이전의 통합 모델을 능가하고 작업별 특화 모델의 성능과 비슷하거나 더 뛰어난 것으로 나타남

1. Introduction

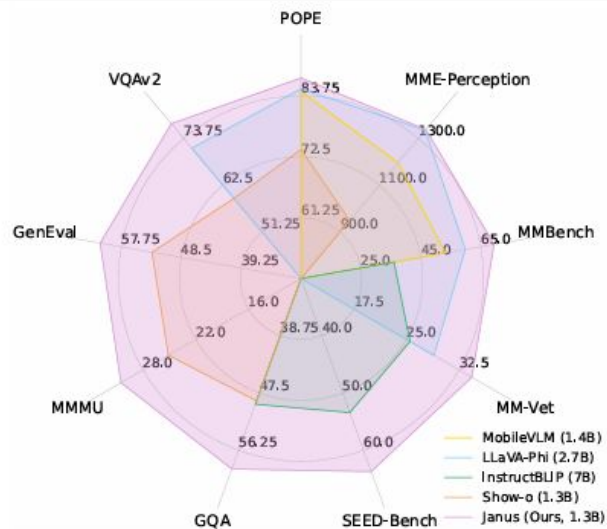
1. Introduction(1)

- Multimodal large model의 발전
 - understanding: LLaVA design을 따름(Vision encoder를 연결 고리로 LLM이 이미지를 이해)
 - visual generation: diffusion 기반 접근 방식이 가장 주목, 최근에는 autoregressive 방식이 발전
- Researcher들은 understanding과 generation을 결합시키려고 노력해왔으나 문제점이 존재
 - Emu는 LLM의 출력을 pretrained diffusion 모델의 condition으로 사용하고, 그 후 diffusion 모델을 통해 이미지를 생성
 - 하지만, 시각적 생성 기능이 외부 확산 모델에 의해 처리되기 때문에 진정한 통합 모델로 간주될 수 없음(multimodal LLM이 직접 이미지를 생성하는 능력이 부족)
 - 이해와 생성 작업에 단일 transformer 적용
 - (이미지 생성에 instruction-following을 개선, 잠재 생성 능력 향상, 모델 중복 회피)
 - 이해와 생성 작업에 필요한 표현(representation)이 서로 다름
 - (두 작업의 표현을 동일한 공간에서 통합하면 충돌과 절충이 발생)

1. Introduction(2)

- 기존의 문제점을 해결하기 위해 Janus를 제안
 - **multimodal 이해와 생성을 위한 시각적 인코딩을 독립적으로 분리**하는 통합 다중 모달 프레임워크
 - 장점
 - 이해와 생성의 서로 다른 세분화된 요구 사항에서 발생하는 충돌 완화 및 절충 필요성 제거
 - 유연성, 확장성: 분리 후에는 이해와 생성 작업 모두 각 도메인에 특화된 최신 인코딩 기술을 채택 가능, 포인트 클라우드, EEG 신호, 또는 오디오 데이터와 같은 추가 입력 유형을 수용, 독립적인 인코더가 특징을 추출한 다음 통합된 트랜스포머를 사용하여 이를 처리
- Janus가 비슷한 매개변수 크기를 가진 기존의 통합 모델들을 다중 모달 이해와 생성 벤치마크 모두에서 능가하며, state-of-the-art 결과를 달성
 - Multimodal understanding 벤치마크인 MMBench, SEED-Bench, POPE에서 Janus(1.3B)는 각각 69.4, 63.7, 87.0의 점수를 달성하여 LLaVA-v1.5(7B)와 Qwen-VL-Chat(7B)을 능가
 - Visual generation 벤치마크인 MSCOCO-30K와 GenEval에서 Janus는 8.53의 FID 점수와 61%의 정확도를 달성하여 DALL-E 2와 SDXL과 같은 텍스트-이미지 생성 모델들을 뛰어넘음

1. Introduction(3)



(a) Benchmark Performance.



(b) Visual Generation Results.

Figure 1 | Multimodal understanding and vision generation results from our Janus. Janus outperforms the previous state-of-the-art unified multimodal models as well as some task-specific multimodal understanding models, while also demonstrating strong visual generation capabilities. The image resolution is 384×384 . Best viewed on screen.

2. Related Work

2.1 Visual Generation

- Autoregressive model은 트랜스포머를 활용하여 **discrete visual tokens**를 예측
 - 시각적 데이터를 토큰화하고 **GPT** 스타일과 유사한 방법으로 예측 수행
- **BERT** 스타일 마스킹 방법으로 마스킹 부분 예측(합성 효율 향상, 비디오 생성에 적용)
- **Continuous diffusion** 모델은 **visual generation**에 인상적인 성능을 보여주었고 확률적 방법을 생성에 적용하여 이산적인 방법을 보완해옴

2.2 Multimodal Understanding

- pretrained LLM을 활용하여 Multimodal Large Language Model(MLLM)은 multimodal 정보를 이해하고 처리해왔음
(pretrained diffusion 모델을 활용하여 MLLM은 이미지 생성까지 확장)

→ 요약: MLLM은 직접적인 이미지 생성 능력이 없기 때문에, 도구 활용 측면에서 diffusion 모델은 MLLM의
출력을 조건으로하여 이미지를 생성

※ 전체 시스템의 Visual 생성 능력은 외부 확산 모델에 의해 제한되어 확산 모델 단독으로 직접 사용하는 것보다
성능이 떨어짐

2.3 Unified Multimodal Understanding and Generation

- 통합된 **multimodal** 이해와 생성 모델은 서로 다른 **modalities** 간에 **reasoning**과 **generation**의 자연스러운 연결을 가능하게 함
- 전통적인 방식은 **Autoregressive** 방식을 사용하였던지 **diffusion** 모델을 사용하였던지 단일 **visual representation**을 사용
 - Chameleon은 **VQ Tokenizer**를 사용하여 이미지 인코딩→ 이해와 생성의 요구사항 사이에서 **trade-off**에 직면하여 **suboptimal** 결과를 초래

반면, **Janus**는 서로 다른 작업이 다양한 수준의 정보를 필요할 수 있을 것이므로, 이해와 생성을 위한 **visual representation**을 명시적으로 분리

3. Janus: A Simple, Unified and Flexible Multimodal Framework

3.1 Architecture(1)

- 순수 텍스트 이해, 멀티모달 이해, 시각적 생성을 위해 독립적인 인코딩 방법을 적용하여 입력 원본을 특징으로 변환하고, 이를 통합된 자기회귀 트랜스포머로 처리

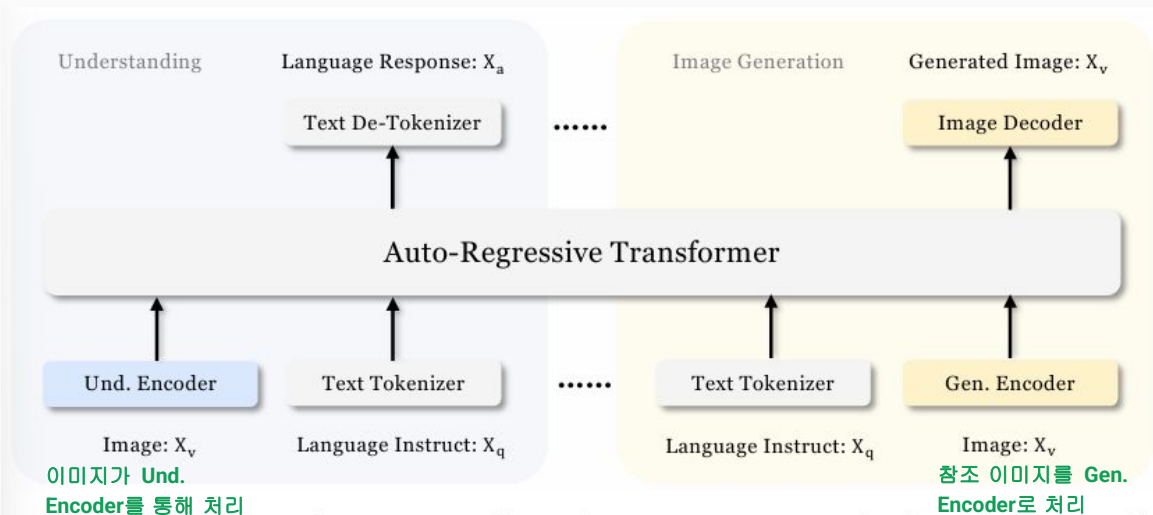


Figure 2 | **Architecture of our Janus.** Different from previous approaches [77, 85] that typically assume visual understanding and generation require the same visual encoder, our Janus decouples visual encoding for visual understanding and visual generation. “Und. Encoder” and “Gen. Encoder” are abbreviations for “Understanding Encoder” and “Generation Encoder”, respectively. Best viewed in color.

1. 텍스트 이해를 위해 LLM의 내장 토큰라이저를 사용하여 텍스트를 이산적 ID로 변환하고 각 ID에 해당하는 특징 표현을 획득
2. 멀티모달 이해를 위해 SigLIP 인코더를 사용하여 이미지에서 고차원 의미 특징을 추출. 이러한 특징들은 2D 그리드에서 1D 시퀀스로 평탄화되며, **understanding** 어댑터를 사용하여 이미지 특징을 LLM의 입력 공간으로 매핑
3. 시각적 생성 작업을 위해 VQ 토큰라이저를 사용하여 이미지를 이산적 ID로 변환. ID 시퀀스가 1D로 평탄화된 후, **generation** 어댑터를 사용하여 각 ID에 해당하는 **codebook** 임베딩을 LLM의 입력 공간으로

3.1 Architecture(1)

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

A Survey on Multimodal Large Language Models

Shukang Yin*, Chaoyou Fu*, Sirui Zhao*, Ke Li,
Xing Sun, Tong Xu, and Enhong Chen, *Fellow, IEEE*

참고용.

기존 모델들은 이런 식으로 했다...

이미지가 LLM에 어떤 식으로 전달되는지...

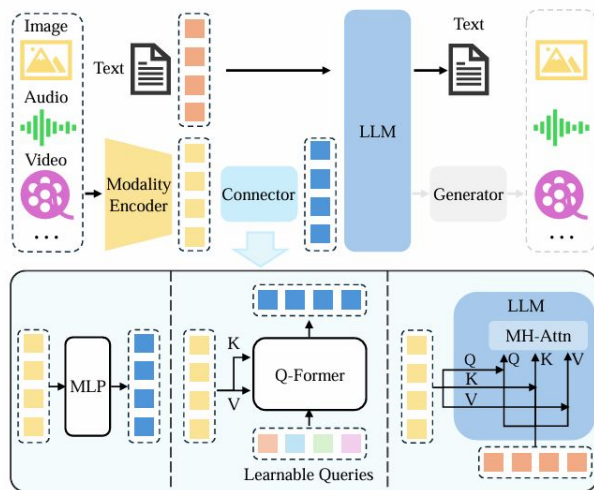
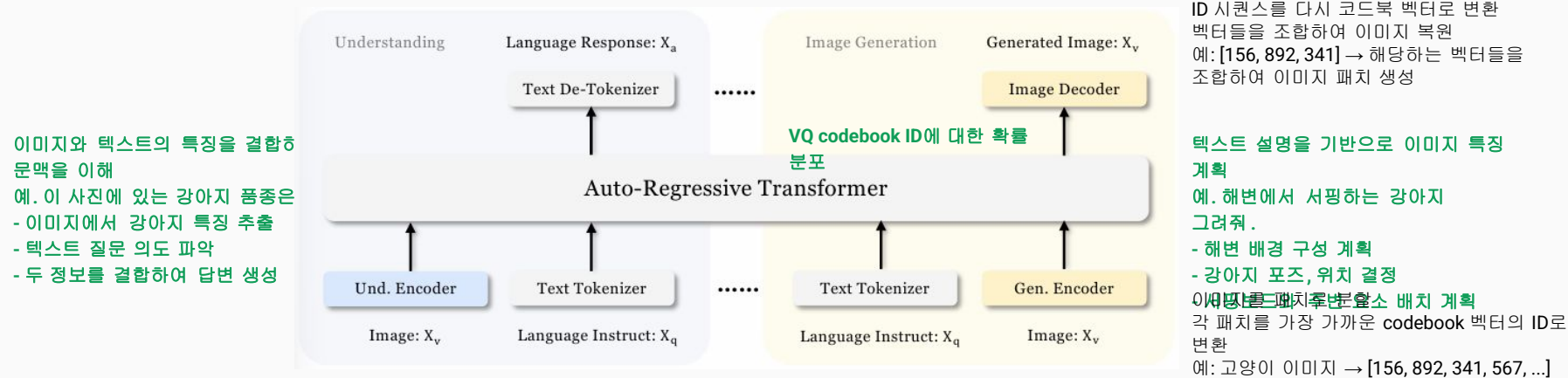


Fig. 2: An illustration of typical MLLM architecture. It includes an encoder, a connector, and a LLM. An optional generator can be attached to the LLM to generate more modalities besides text. The encoder takes in images, audios or videos and outputs features, which are processed by the connector so that the LLM can better understand. There are broadly three types of connectors: projection-based, query-based, and fusion-based connectors. The former two types adopt token-level fusion, processing features into tokens to be sent along with text tokens, while the last type enables a feature-level fusion inside the LLM.

3.1 Architecture(2)

특징 시퀀스들을 연결하여 LLM에 입력하여 처리



순수 텍스트 이해와 멀티모달 이해 작업에서는

LLM의 built-in prediction head를 텍스트 예측에 활용

모델의 출력층(Head)이 특정 예측 작업을 수행하도록 설계된 구조

전체 모델은 특정한 설계된 어텐션 마스크 없이

Auto-Regressive Transformer 프레임워크를 따름

시각적 생성 작업에서는 Randomly Initialized Janus에서 별도 구현

prediction Head를 이미지 예측에 사용

이미지 생성 작업에서는 이러한 방식이 기존에 학습된 텍스트 기반 Prediction Head보다 더 효과적

(VQ tokenizer의 discrete ID를 예측해야 하므로 다른 형태의 prediction head가 필요)

3.2 Training Procedure(1)

- Janus의 학습 과정은 3 단계로 구성됨

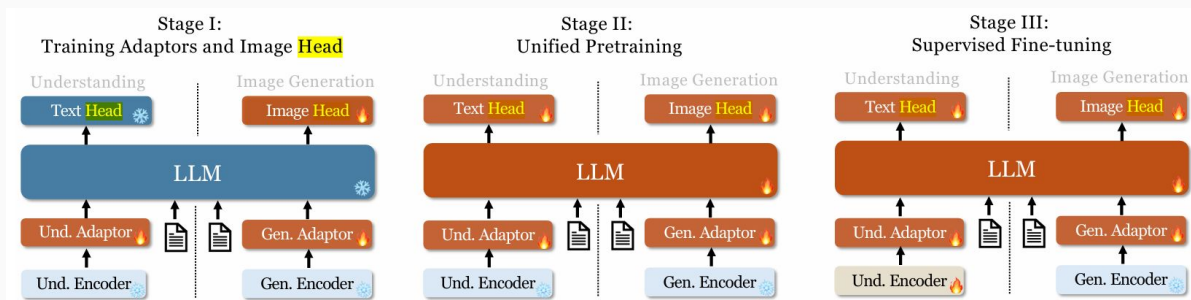
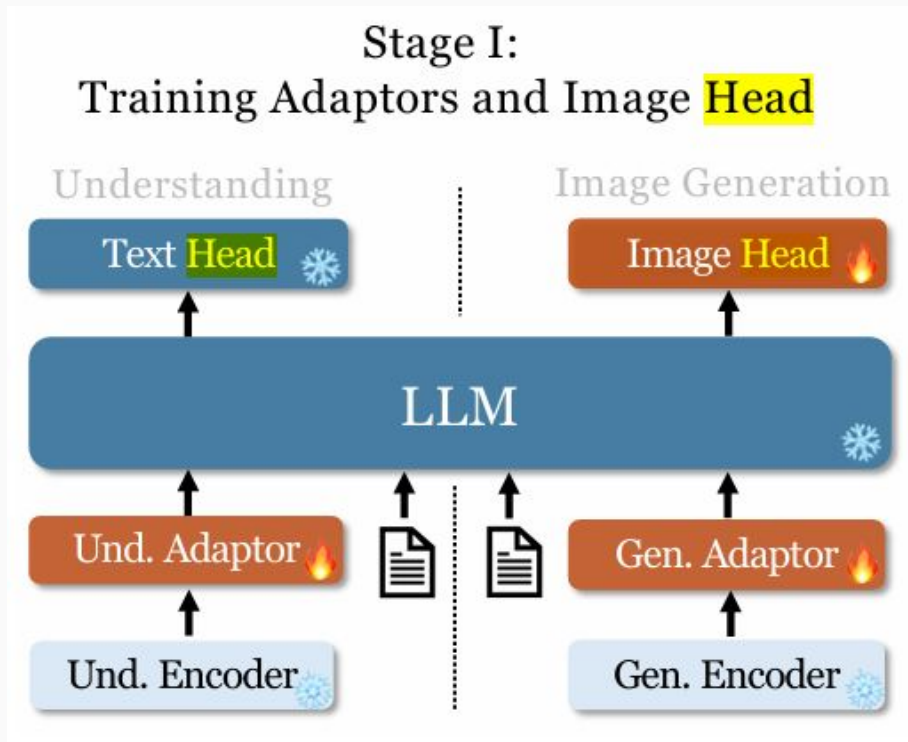


Figure 3 | **Our Janus adopts a three-stage training procedure.** We use flame symbols/flame symbols in the diagram to indicate the module updates/does not update its parameters.

3.2 Training Procedure(2)

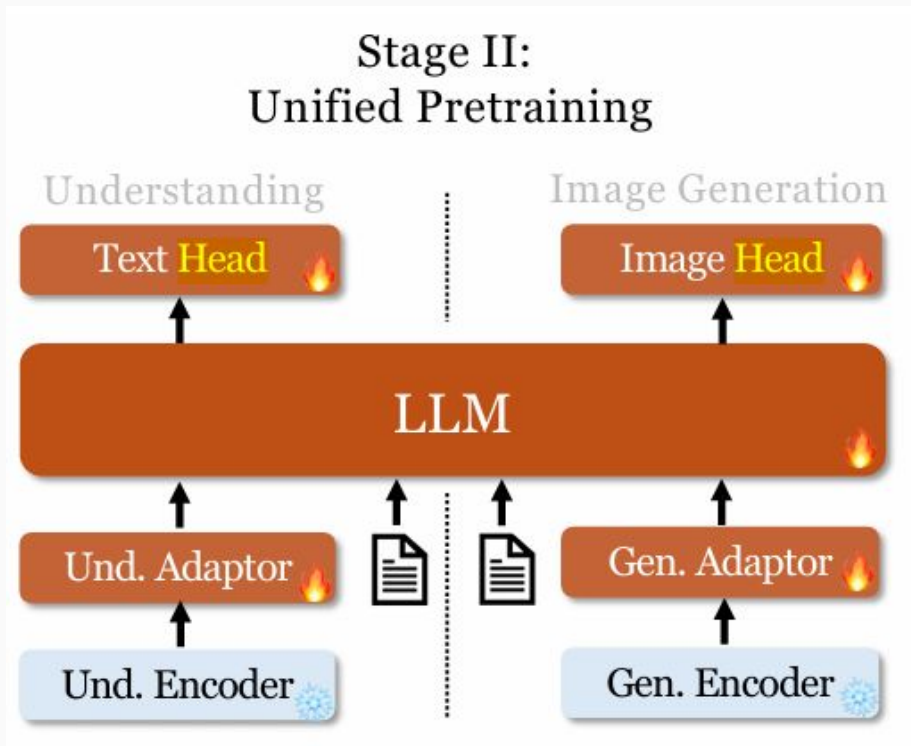


임베딩 공간 내에서 시각적 요소와 언어적 요소 간의 개념적 연결을 만드는 과정

LLM이 이미지에 표시된 개체들을 이해하고 기초적인 시각적 생성 능력을 갖게됨

시각적인코더와 LLM을 고정(frozen)하고 understanding adaptor, generation adaptor, 이미지 헤드 내의 학습 가능한 파라미터들만 업데이트

3.2 Training Procedure(3)



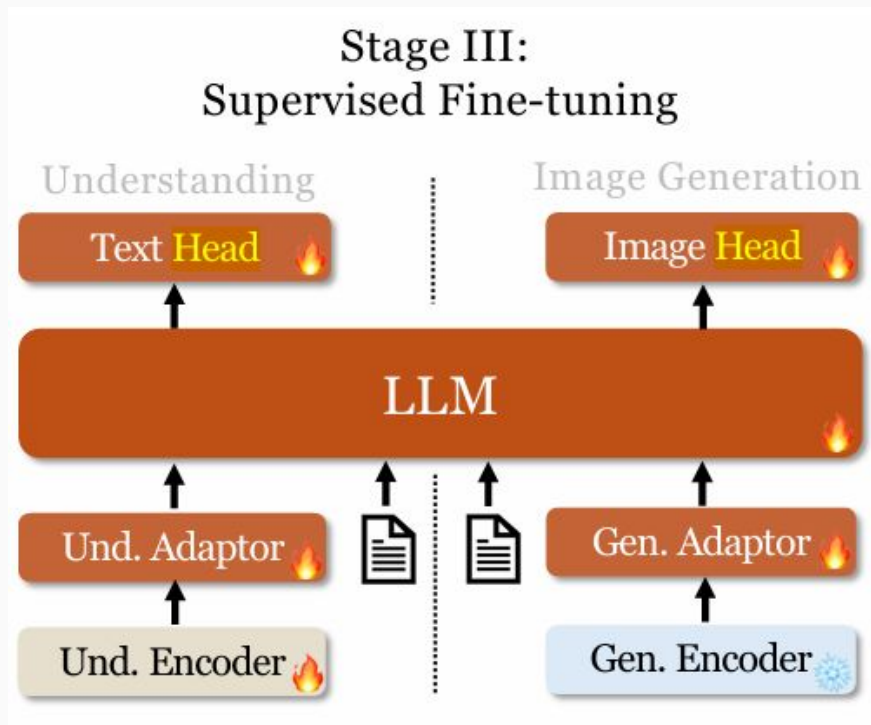
Janus가 멀티모달 이해와 생성을 모두 학습할 수 있도록 멀티모달 코퍼스로 통합 사전 학습을 수행

LLM을 언프리즈(unfreeze)하고 모든 유형의 학습 데이터(순수 텍스트 데이터, 멀티모달 이해 데이터, 시각적 생성 데이터)를 활용

Pixart에서 영감을 받아, ImageNet-1k를 사용한 단순한 시각적 생성 학습으로 시작하여 모델이 기본적인 픽셀 의존성을 파악

이후 일반적인 텍스트-투-이미지 데이터를 통해 모델의 오픈 도메인 시각적 생성 능력을 향상

3.2 Training Procedure(4)



모델의 instruction-following 및 대화 능력을 향상시키기 위해 instruction tuning data로 사전 학습된 모델을 미세 조정

generation encoder를 제외한 모든 파라미터를 미세 조정

시스템과 사용자 프롬프트를 마스킹하면서 답변을 supervising하는 데 중점

Janus가 멀티모달 이해와 생성 모두에서 숙련도를 갖추도록 하기 위해, 특정 작업을 위한 별도의 모델을 미세 조정하지 않음
이미지 생성에만 fine-tuning하지 않는다는 뜻
순수 텍스트 대화 데이터, 멀티모달 이해 데이터, 시각적 생성 데이터를 혼합하여 사용함으로써 다양한 시나리오에서의 다재다능성을 보장

3.3 Training Objective

- Janus는 autoregressive model이고 학습에 cross-entropy loss를 적용함

$$\mathcal{L} = - \sum_{i=1} \log P_{\theta}(x_i | x_{<i})$$

가중치 θ 에 의해 모델링되는 조건부 확률

- pure text understanding and multimodal understanding task \rightarrow text sequence loss 계산

예시.

```
P_θ(퀴엠타강아지는) = 0.8  
log(0.8) = -0.223  
Loss = -(-0.223) = 0.223
```

- visual generation tasks \rightarrow image sequence에 대해서만 loss 계산

예시.

첫 번째 토큰 예측

```
P_θ(ID_234|"해변의 강아지") = 0.7  
log(0.7) = -0.357
```

두 번째 토큰 예측

```
P_θ(ID_567|ID_234, "해변의 강아지") = 0.6  
log(0.6) = -0.511
```

최종 loss

```
Total Loss = -(-0.357 + -0.511) = 0.868
```

모델의 단순 구조를 위해 서로 다른 작업에 서로 다른 가중치를 할당하지 않음

예시.

최종 loss = ($\alpha \times$ 텍스트 이해 loss) + ($\beta \times$ 멀티모달 이해 loss) + ($\gamma \times$ 이미지 생성 loss)

동일 가중치의 경우: 최종 손실 = ($1.0 \times$ 텍스트 이해 loss) + ($1.0 \times$ 멀티모달 이해 loss) + ($1.0 \times$ 이미지 생성 loss)

3.4 Inference

- 추론에 next-token prediction 방식 적용 이전 출력을 입력으로 사용하여 토큰을 차례대로 생성
 - pure text understanding and multimodal understanding task → 예측된 분포에서 순차적으로 토큰을 샘플링하는 표준 방식 따름

예시. 입력: [강아지 이미지] + "이 사진의 동물은" → 1단계: 다음 토큰 예측 분포 → 2단계: 다음 토큰 예측 분포 → 최종 출력: "이 사진의 동물은 강아지입니다"

"강아지": 0.9	"입니다": 0.7
"고양이": 0.05	"가": 0.2
"토끼": 0.05	"이": 0.1

→ 확률에 따라 "강아지" 샘플링. → 확률에 따라 "입니다" 샘플링

- visual generation tasks → 이전 연구들과 유사하게 classifier-free guidance(CFG)를 활용
이미지 생성 과정에서 생성 품질을 향상시키는 기술
각 토큰에 대한 logit lg 계산

○ $lg = lu + s(lc - lu)$

lc : 조건부 logit, lu : 비조건부 logit, s : classifier-free guidance를 위한 스케일(evaluation 기본값은 5)

예시.

1. 조건부 예측 (lc)

- 프롬프트 "해변에서 노는 강아지"를 고려한 예측
- ID_234: 0.8 (하늘)
- ID_567: 0.1 (구름)
- ID_890: 0.1 (기타)

2. 비조건부 예측 (lu)

- 프롬프트 없이 예측
- ID_234: 0.4
- ID_567: 0.3
- ID_890: 0.3

text-to-image data에 대해 10% 확률로 PAD 토큰 처리

3. CFG 적용 ($s=5$)

$$lg = lu + 5(lc - lu)$$

lg 에 따라 이미지 토큰을 선택

CFG의 스케일 값(s)이 클수록 프롬프트의 영향력이 커져 더 명확한 이미지가 생성되지만, 너무 크면 부자연스러운 결과

3.5 Possible Extensions

- 이해와 생성을 위한 **separate encoders**를 특징으로 하며, 이는 **간단하고 확장하기 쉬움**
 - Multimodal Understanding
 - EVA-CLIP, InternViT 등과 같이 시각 생성 작업 처리 능력을 걱정할 필요 없이 더 강력한 비전 인코더를 선택 가능
 - 고해상도 이미지를 처리하기 위해 동적 고해상도 기술을 사용 가능,
pixel shuffle 연산을 사용하여 계산 비용을 절약하기 위해 토큰을 더 압축
 - Visual Generation
 - MoVQGAN과 같이 인코딩 후 더 많은 이미지 세부 사항을 보존하기 위해 더 세밀한 인코더를 선택 가능
 - 확산 손실과 같이 시각적 생성을 위해 특별히 설계된 손실 함수를 사용 가능
 - 시각적 생성 중 누적된 오류를 줄이기 위해 **AR**(인과적 어텐션)과 병렬(양방향 어텐션) 방법의 조합을 사용 가능
 - Support for Additional Modalities
 - 3D 포인트 클라우드, 촉각, EEG와 같은 다양한 모달리티를 수용하기 위해 추가 인코더와 쉽게 통합 가능
 - Janus가 강력한 **multimodal generalist model**이 될 수 있는 잠재력을 갖게 함

4. Experiments

- 모델 아키텍처, 학습 데이터셋, 평가 벤치마크를 포함하는 실험 설정
- Janus의 성능 타 모델들과 벤치 마크에서 성능 비교
- 효과를 검증하기 위해 광범위한 **ablation studies**
- 정성적 결과를 제시

4.1 Implementation Details(1)

- 구현 세부 사항
 - Base language model: 최대 4096 시퀀스 길이를 지원하는 DeepSeek-LLM (1.3B)
 - vision encoder(for understanding tasks): SigLIP-Large-Patch16-384 @GOOGLE, 이미지와 텍스트를 함께 처리 할 수 있는 모델
 - generation encoder: 16,384 크기의 codebook, 이미지를 16배 downsampling 384x384 이미지 처리, 제로샷 이미지 분류, 이미지-텍스트 검색, image gen 적용
 - understanding 어댑터와 generation 어댑터: two-layer MLPs 가로, 세로 메모리/연산량/정보 압축, 픽셀 공간보다 압축된 잠재 공간에서 VQ 코딩을 통해 LLM이 처리할 수 있는 텍스트 임베딩 공간으로 변환
 - 이미지 resize: 384x384 pixels
 - multimodal understanding data에 대해 이미지의 긴 쪽을 384 크기 조정하고 짧은 쪽을 배경색(RGB: 127, 127, 127)으로 패딩하여 384 크기에 맞춤
 - visual generation data에 대해 짧은 쪽을 384로 크기 조정하고 긴 쪽을 384로 자름
 - 학습 효율성을 높이기 위해 학습 중에 시퀀스 패킹을 사용 합승 과정에서 배치 처리의 효율성을 높이는 기술, 하나의 배치에 여러 시퀀스를 패킹
 - single training step에서 지정된 비율에 따라 모든 데이터 유형을 혼합 배치 1: [이미지 14, 이미지 2, 이미지 3] (384x384) → 할당에 따라 데이터 (40%, 이미지 + 텍스트), 시각적 생성 데이터 (30%, 텍스트, “~를 그려주세요”)
 - PyTorch 기반의 경량화되고 효율적인 분산 학습 프레임워크인 HAI-LLM을 사용하여 학습되고 평가
 - 전체 학습 과정은 각각 8개의 Nvidia A100(40GB) GPU가 장착된 16개 노드로 구성된 클러스터에서 7일이 소요

4.1 Implementation Details(2)

- 점진적인 학습률 감소, 웜업 적용, 배치 크기 조정, 가중치 감쇠 도입 등의 전략을 통해 모델을 효과적으로 학습

Table 1 | **Detailed hyperparameters of our Janus.** Data ratio refers to the ratio of multimodal understanding data, pure text data, and visual generation data.

Hyperparameters	Stage 1	Stage 2	Stage 3
Learning rate	1.0×10^{-3}	1×10^{-4}	2.0×10^{-5}
LR scheduler	Cosine	Constant	Constant
Weight decay	0.0	0.0	0.1
Gradient clip	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)		
Warm-up steps	300	5,000	0
Training steps	10,000	180,000	24,000
Batch size	256	512	256
Data Ratio	1 : 0 : 1	2 : 3 : 5	7 : 3 : 10

정규화

초반에 높은 학습률로 빠르게 학습, 후반에 작은 학습률

초반에 cosine 학습률 스케줄러로 점진적 감소, 이후 일정한 학습률

최종 단계에서 0.1로 적용하여 과적합(overfitting)을 방지, 손실 함수에 가중치 크기에 대한

Gradient explosion을 방지

Weight Decay를 더 효과적으로 적용, 더 빠르게 변화에 적응(cf. $\beta_1=0.9, \beta_2=0.99$, 느리게 반응)

초기에 웜업 스텝을 설정하여 학습을 안정적으로 시작

Stage 2에서 가장 많은 훈련 스텝을 사용하여 메인 학습

Stage 2에서 배치 크기를 가장 크게 설정하여 학습 효율성을 높임

각 단계별로 다른 데이터 분포를 사용하여 모델의 학습 데이터

구성을 조절

$$w_t = w_t - \alpha \frac{m_t}{\sqrt{w_t} + \epsilon} - \lambda w_t$$

4.2 Data Setup(1)


- 사전 학습(pretraining) 및 지도 학습 기반 미세 조정(supervised finetuning) 데이터셋 세부사항

- Stage I 학습 데이터 구성

- Multimodal understanding

- ShareGPT4V의 125만개 image-text paired captions 사용


- <image><text> 형식

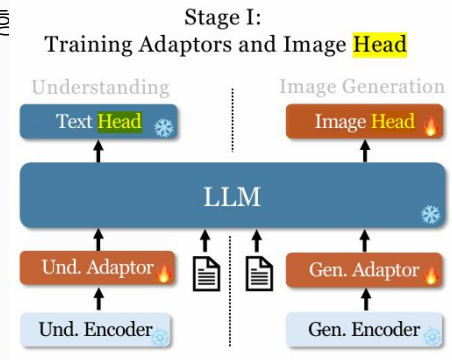
- <image>:  (한 폭의 그림)
- <text>: "This is a beautiful sunset over the mountains."

- Visual generation

- Image-Net-1k의 120만개 샘플 활용

- <category_name><image> 형식으로 text-to-image 데이터(이미지와 레이블이 짝을 이루는 데이터셋)

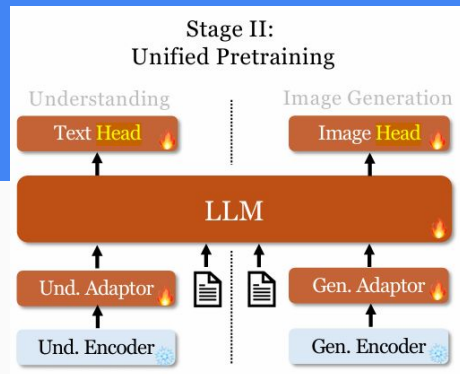
- <category_name>: "Golden Retriever"
- <image>:  (해당 카테고리에 속하는 강아지 사진)



4.2 Data Setup(2)

○ Stage II 학습 데이터 구성

- 텍스트 전용 데이터 (Text-only data)
 - DeepSeek-LLM 제공하는 training text corpus
- 이미지-텍스트 혼합 데이터 (Interleaved image-text data)
 - WikiHow(단계별 가이드 제공하는 문서, 이미지 조합) 및 WIT(위키피디아 문서 추출 이미지와 텍스트) 데이터셋을 활용
- 이미지 캡션 데이터 (Image caption data)
 - 출처: (논문에 명시)
 - 오픈소스 multimodal model을 활용하여 image recaptioning
 - question-answer pairs 형식: "<image>Describe the image in detail.<caption>"
- 테이블 및 차트 데이터 (Table and chart data)
 - DeepSeek-VL에서 제공하는 테이블 및 차트 데이터를 활용
 - <question><answer> 형식



이미지 : 🏠 (화분에 물을 주는 손)

텍스트 : "Step 1: Gently pour water into the pot to hydrate the plant."

4.2 Data Setup(3)

- Stage II 학습 데이터 구성 continue...
 - 시각적 생성 데이터 (Visual Generation Data)
 - 출처: (논문에 명시)
 - 200만개 내부 데이터도 사용
 - 일부 데이터셋은 **aesthetic scores** 및 **image size** 기준으로 필터링하여 20%만 남김
 - 학습시 첫 번째 문장만 사용하는 확률을 25%로 설정하여 모델이 **short description** 생성하는 능력 강화
 - ImageNet 샘플은 초기 12만 스텝 동안만 사용, 이후 다른 데이터셋의 이미지가 후반부 6만 스텝에서 등장
→ 기본적인 **pixel dependencies**를 먼저 학습한 후, 더 복잡한 **scene understanding**으로 발전하도록 유도
 - <caption><image> 형식

4.2 Data Setup(3)

- Stage III 학습 데이터 구성
 - 텍스트 이해 (Text Understanding)
 - 출처: (논문 명시)
 - 멀티모달 이해 (Multimodal Understanding):
 - 출처: instruct tuning data from (논문 명시)
 - 시각적 생성 (Visual Generation)
 - 출처: Stage II에서 사용된 일부 데이터
 - 400만개 in-house data

※ Instruct Tuning 데이터 형식: “User: <입력 메시지> \n Assistant:<Response>”

multi-turn dialogues를 위해 위 형식 반복

4.3 Evaluation Setup

- Multimodal Understanding 능력 평가

- image-based vision-language benchmarks에서 평가
 - VQAv2, GQA, POPE, MME, SEED, MMB, MM-Vet, MMMU

→ 모델이 이미지와 텍스트를 함께 이해하고 처리하는 능력을 다양한 벤치마크에서 평가

- Visual Generation 능력 평가

- MSCOCO-30K, MJHQ-30K
 - 생성된 이미지와 30,000개의 고품질 이미지를 비교하여 Fréchet Inception Distance (FID) 지표를 사용해 평가
 - 이미지 생성 모델의 전반적인 효과성(효율성)을 측정하는 데 사용
- GenEval
 - Text-to-Image Generation을 평가하는 고난이도 벤치마크
 - 모델의 종합적인 생성 능력(comprehensive generative abilities)을 평가하도록 설계
 - 개별 인스턴스 수준(instance-level)에서 모델의 구성 능력(compositional capabilities)을 분석

→ 모델이 고품질 이미지를 얼마나 효과적으로 생성하는지, 그리고 이미지 기반으로 텍스트를 얼마나 정확하게 생성하는지를 평가

4.4 Comparison with State-of-the-arts(1)

- Multimodal Understanding 성능
 - SOTA의 Understanding only 모델과 Unified 모델과 비교

Table 2 | Comparison with state-of-the-arts on multimodal understanding benchmarks. “Und.” and “Gen.” denote “understanding” and “generation”, respectively. Models using external pretrained diffusion model are marked with †.

Type	Model	# LLM Params	POPE†	MME-P†	MMB†	SEED†	VQAv2 _(test) †	GQA†	MMM†	MM-Vet†
Und. Only	LLaVA-v1.5-Phi-1.5 [86]	1.3B	84.1	1128.0	-	-	75.3	56.5	30.7	-
	MobileVLM [14]	1.4B	84.5	1196.2	53.2	-	-	56.1	-	-
	MobileVLM-V2 [15]	1.4B	84.3	1302.8	57.7	-	-	59.3	-	-
	MobileVLM [14]	2.7B	84.9	1288.9	59.6	-	-	59.0	-	-
	MobileVLM-V2 [15]	2.7B	84.7	1440.5	63.2	-	-	61.1	-	-
	LLaVA-Phi [96]	2.7B	85.0	1335.1	59.8	-	71.4	-	-	28.9
	LLaVA [51]	7B	76.3	809.6	38.7	33.5	-	-	-	25.5
	LLaVA-v1.5 [50]	7B	85.9	1510.7	64.3	58.6	78.5	62.0	35.4	31.1
	InstructBLIP [16]	7B	-	-	36.0	53.4	-	49.2	-	26.2
	Qwen-VL-Chat [3]	7B	-	1487.5	60.6	58.2	78.2	57.5	-	-
	IDEFICS-9B [41]	8B	-	-	48.2	-	50.9	38.4	-	-
	Emu3-Chat [83]	8B	85.2	-	58.5	68.2	75.1	60.3	31.6	-
	InstructBLIP [16]	13B	78.9	1212.8	-	-	-	49.5	-	25.6
Und. and Gen.	DreamLLM† [21]	7B	-	-	-	-	72.9	-	-	36.6
	LaVIT† [36]	7B	-	-	-	-	66.0	46.8	-	-
	Emu† [75]	13B	-	-	-	-	52.0	-	-	-
	NEXT-GPT† [84]	13B	-	-	-	-	66.7	-	-	-
	Show-o [86]	1.3B	73.8	948.4	-	-	59.3	48.7	25.1	-
	Gemini-Nano-1 [78]	1.8B	-	-	-	-	62.7	-	26.3	-
	LWM [52]	7B	75.2	-	-	-	55.8	44.8	-	9.6
	VILA-U [85]	7B	85.8	1401.8	-	59.0	79.4	60.8	-	33.5
	Chameleon [77]	7B	-	-	-	-	-	-	22.4	8.3
	Janus (Ours)	1.3B	87.0	1338.0	69.4	63.7	77.3	59.1	30.5	34.3

- Janus가 멀티모달 이해와 시각적 생성을 위한 시각적 인코딩 (Visual Encoding)을 분리 (Decoupling)하여, 두 작업 간의 충돌을 완화했기 때문에 우수한 성능을 기록하였다고 얘기함

4.4 Comparison with State-of-the-arts(2)

- Visual Generation 성능

- GenEval, COCO-30K, MJHQ-30K 벤치마크에서 성능 평가

Table 3 | Evaluation of text-to-image generation ability on GenEval benchmark. “Und.” and “Gen.” denote “understanding” and “generation”, respectively. Models using external pretrained diffusion model are marked with †.

Type	Method	# Params	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall†
Gen. Only	LlamaGen [73]	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [67]	1.4B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [67]	0.9B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-α [9]	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [67]	0.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [66]	6.5B	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [83]	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [62]	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
Und. and Gen.	SEED-X† [29]	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o [86]	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	LWM [52]	7B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	Chameleon [77]	34B	-	-	-	-	-	-	0.39
	Janus (Ours)	1.3B	0.97	0.68	0.30	0.84	0.46	0.42	0.61

→ Janus가 지시(instruction)를 따르는 능력이 뛰어남을 보여줌

※ FID(Fréchet Inception Distance) cf. MSE, SSIM

- 생성된 이미지와 실제 이미지 간의 품질 차이를 측정하는 지표(GAN, Diffusion 모델 평가에 많이 사용)

- 생성된 이미지(Generated Images)와 실제 이미지(Real Images)의 특징(feature) 분포 차이를 측정

Table 4 | Evaluation of text-to-image generation ability on MSCOCO-30K and MJHQ-30K benchmark. “Und.” and “Gen.” denote “understanding” and “generation”, respectively. Models using external pretrained diffusion model are marked with †.

Type	Model	# Params	COCO-30K↓	MJHQ-30K↓
Gen. Only	DALL-E [65]	12B	27.50	-
	GLIDE [59]	5B	12.24	-
	LDM [67]	1.4B	12.64	-
	DALL-E 2 [66]	6.5B	10.39	-
	SDv1.5 [67]	0.9B	9.62	-
	GigaGAN [37]	0.9B	9.09	-
	PixArt-α [9]	0.6B	7.32	-
	Imagen [68]	34B	7.27	-
Und. and Gen.	RAPHAEL [87]	3B	6.61	-
	Emu† [75]	13B	11.66	-
	NEXT-GPT† [84]	13B	11.28	-
	SEED-X† [29]	17B	14.99	-
	Show-o [86]	1.3B	9.24	15.18
	LWM [52]	7B	12.68	17.77
	VILA-U (256) [85]	7B	-	12.81
	VILA-U (384) [85]	7B	-	7.69
	Janus (Ours)	1.3B	8.53	10.10

→ Janus가 생성하는 이미지는 고품질(High Quality)을 유지하며, 시각적 생성(Vision Generation)에서 강력한 가능성을 보여줌

4.5 Ablation Studies

- Janus의 설계 개념의 효과성을 검증

- 시각적 인코딩 (Visual Encoding) 분리 (decoupling)의 중요성과 이점을 검증하기 위한 실험을 설계
- 통합 학습 (Unified Training)이 개별 작업 (예. 멀티모달 이해 및 시각적 생성)에 미치는 영향을 분석

Table 5 | **Ablation studies.** We verify the effectiveness of decoupling visual encoding and compare unified training with task-specific training. “Und.”, “Gen.” and “SE. Tokenizer” denote “understanding”, “generation” and “semantic tokenizer”, respectively.

Exp ID	Visual Encoder	Training Task	POPE↑	MMB↑	SEED↑	MMMU↑	COCO-FID↓
A	VQ Tokenizer	Und. + Gen.	60.1	35.0	34.9	24.7	8.72
B	SE. Tokenizer	Und. + Gen.	82.4	52.7	54.9	26.6	7.11
C	SE. Tokenizer	Und.	83.9	62.1	60.8	27.5	-
D	SigLIP + VQ (Ours)	Und. + Gen.	87.0	69.4	63.7	30.5	8.53
E	SigLIP	Und.	85.9	70.6	64.8	28.8	-
F	VQ Tokenizer	Gen.	-	-	-	-	8.92

POPE, MMB, SEED, MMMU 점수 (↑ 높을수록 좋음): 멀티모달 이해 성능을 평가

COCO-FID 점수 (↓ 낮을수록 좋음): 이미지 생성 성능을 평가

SE. Tokenizer (Semantic Tokenizer)
: VQ tokenizer가 의미적 정보 추출에 weak할 수 있어서 (multimodal 이해에서

덜 효과적일 수 있어서) 이미지에서 high level 의미 정보를 추출할 수 있을 뿐만 아니라, low level 이미지를 Discrete IDs로 변환할 수 있도록 SigLIP에서 distill한 Unified Model. 통합 학습

※ 각 모델들은 동일한 학습 과정을 거침
Pure understanding: 시각적 생성 데이터 제외 학습
Pure Generation: 이해 데이터 제외 학습

→ 멀티모달 이해와 이미지 생성의 균형을 맞추기 위해 SigLIP과 VQ 토큰라이저를 조합한 Exp-D 모델이 가장 효과적

하나의 인코더로 이해와 생성을 모두 수행하면 성능이 낮아질 수 있음

시각적 인코딩을 분리하는 것이 성능 향상에 필수적

Janus는 멀티모달 이해와 생성 능력을 균형 있게 통합하여 학습할 수 있으며, 어느 한쪽의 성능이 크게 저하되지 않음을 입증

B가 D보다 visual generation이 좋은 이유에 대한 가설

- SE. tokenizer의 Discrete IDs가 의미적 일관성이 높아 LLM의 합리적 예측에 도움

- B의 visual encoder가 D의 Gen. encoder보다 파라미터 더 많음

4.6 Qualitative Results

Visualizations of Visual Generation.

- Janus 모델과 Diffusion 기반 모델(SDXL과 유사), 그리고 Autoregressive 모델(LlamaGen) 간의 정성적(qualitative) 비교
SDXL: 텍스트 기반 고품질 이미지 생성을 위한 Diffusion 모델 LlamaGen: Autoregressive 기반, 텍스트 기반 이미지 생성
- 시각적 생성 (Visual Generation)에서 뛰어난 지시 수행 능력 (Instruction-Following Capabilities)을 보여줌

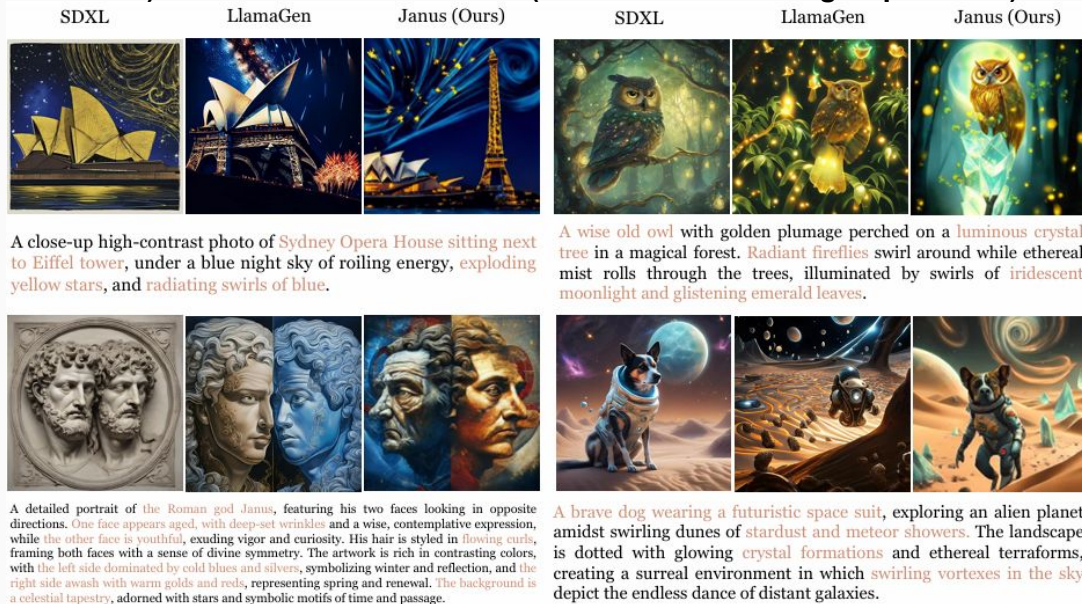


Figure 4 | Qualitative comparisons of visual generation with LlamaGen and SDXL. The images generated by Janus show better consistency with the user's prompts. The image resolutions for SDXL, LlamaGen, and ours are 1024×1024 , 512×512 , and 384×384 , respectively. Best viewed on screen.

4.6 Qualitative Results

Multimodal Understanding on MEME Images

- Janus의 멀티모달 이해 능력을 Chameleon 및 Show-o과 비교한 정성적 (qualitative) 결과

Chameleon: 멀티모달 통합 학습, Diffusion 기반 생성 가능

Show-o: Autoregressive 기반, 멀티모달 생성 및 이해

Chameleon 및 Show-o는 공유된 비전 인코더 (shared vision encoder)를 사용하기 때문에 세부적인 정보를 정확히 이해하는 데 어려움을 겪음

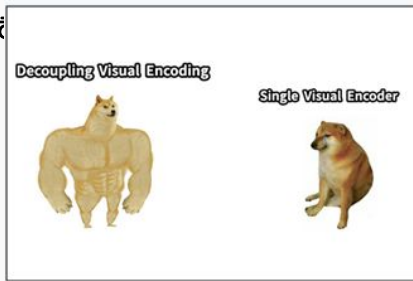


User: What is the meaning of this meme from the top panel to the bottom?

Janus (Ours): The meme is a humorous take on the concept of a Nobel Prize in Physics. The top panel shows a man in a red jacket with his hand up, seemingly confused or surprised, while the bottom panel shows the same man with a smile on his face, seemingly pleased or proud, as if he has just won the Nobel Prize. The meme is a play on the idea that the man in the top panel might be confused or surprised by the idea of winning a Nobel Prize in Physics, while the man in the bottom panel seems to be genuinely happy about it.

Chameleon: The meme is a visual representation of a conversation between two people, with the top panel featuring the text "I'm not arguing, I'm just explaining why I'm right" and the bottom panel featuring the text "No, you're not." The meme is using a technique called "irony" to convey the idea that even though one person is presenting themselves as non-argumentative, they are actually being argumentative. The use of the phrase "I'm not arguing, I'm just explaining why I'm right" is a common tactic used by people who are trying to avoid an argument but are actually engaging in one. The meme is poking fun at this tactic by showing the person's actions and words as being in direct contradiction to their claim.

Show-o: The meaning of the meme from the top panel to the bottom is that it is a humorous and lighthearted depiction of a man wearing a yellow hat and a red jacket, with the caption "I'm not a fan of the sun." This suggests that the man is not a fan of the sun, and he is using the sun as a metaphor for his feelings or emotions. The meme is likely meant to be humorous and entertaining, as it combines a serious subject with a lighthearted and playful approach.



User: Explain this meme.

Janus (Ours): The meme is comparing two different types of neural networks used in machine learning. The first type is called "Decoupling Visual Encoding," and the second type is called "Single Visual Encoder." The meme is trying to convey that the first type is more powerful and can handle more complex tasks, while the second type is simpler and less powerful.

Chameleon: This image shows a cartoon of a pig and a bear side-by-side, with the text "When you're a vegan and your friends still order sausage and bacon" above them. The image is a humorous take on the challenges of being a vegan in a non-vegan environment, where one's friends or family may not share their dietary preferences.

Show-o: The image features two dogs, one brown and one white, standing side by side. The brown dog is positioned on the left side of the image, while the white dog is on the right. Both dogs are looking at the camera, seemingly posing for a picture. The brown dog is slightly larger than the white dog. The image is captioned with the phrase "I'm not a dog, but I'm a dog." This humorous statement is accompanied by the image of the two dogs, emphasizing their ...

Figure 5 | Qualitative results of multimodal understanding on humorous memes. We compare the response with Chameleon-7B [77] and Show-o [86]. We emphasize the key-points in the response. Best viewed on screen.

5. Conclusion

- Janus, 단순하면서도 통합적이며 확장 가능한 멀티모달 이해 및 생성 모델
- 핵심 아이디어는 멀티모달 이해와 생성을 위한 시각적 인코딩을 분리(decouple)
- Janus는 멀티모달 이해 및 생성에서 더 발전할 가능성이 클 뿐만 아니라, 더 많은 입력 모달리티(input modalities)를 통합하는 것도 쉽게 확장 가능

Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling

데이터 및 모델 확장을 통한 통합 멀티모달 이해 및 생성

Abstract

- Janus-Pro는 이전 개발된 Janus의 발전된 버전임
 - (1) optimized training strategy
 - (2) expanded training data
 - (3) scaling to larger model size
- 개선 사항
 - 멀티모달 이해 및 텍스트-이미지 변환 지시 수행 능력에서 크게 향상
 - 텍스트-이미지 생성의 안정성 강화

1. Introduction

1. Introduction

- Recent advancements는

- 시각적 생성 (Visual Generation) 작업에서 **Instruction-Following Capabilities**을 향상 시켰고
- 모델의 중복성 (Model Redundancy)을 줄이는 데 효과적임이 입증

하지만 멀티모달의 이해와 생성에 동일한 **Visual encoder**를 사용하여 성능이 최적보다 낮아짐

공유된 비전 인코더는 충돌과 절충에 의해 불필요한 중복 연산 증가 및 최적화 어려움

→ 이에 **Janus**는 시각적 인코딩 (Visual Encoding) decoupling하여 두 작업 간 충돌을 완화 시켜 뛰어난 성능을 달성

- **Janus**는 10억개 파라미터 규모에서 성능을 검증 받았으나 제한된 학습데이터와 작은 모델 크기로 인해
 - suboptimal performance on short prompts image generation
 - unstable text-to-image generation quality

- **Janus-Pro**는 세 가지 측면에서 기존 버전을 개선함

- 훈련 전략 (Training Strategies)
- 데이터 (Data)
- 모델 크기 (Model Size): 1B and 7B

멀티모달 이해 성능 (MMBench
벤치마크)

- Janus-Pro-7B: 79.2
- Janus: 69.4점
- TokenFlow: 68.9점
- MetaMorph: 75.2점

텍스트-이미지 변환 지시 수행 (GenEval
벤치마크)

- Janus-Pro-7B: 0.80
- Janus: 0.61
- TokenFlow: 0.67
- MetaMorph: 0.74

2. Method

2.1 Architecture

- Janus와 동일

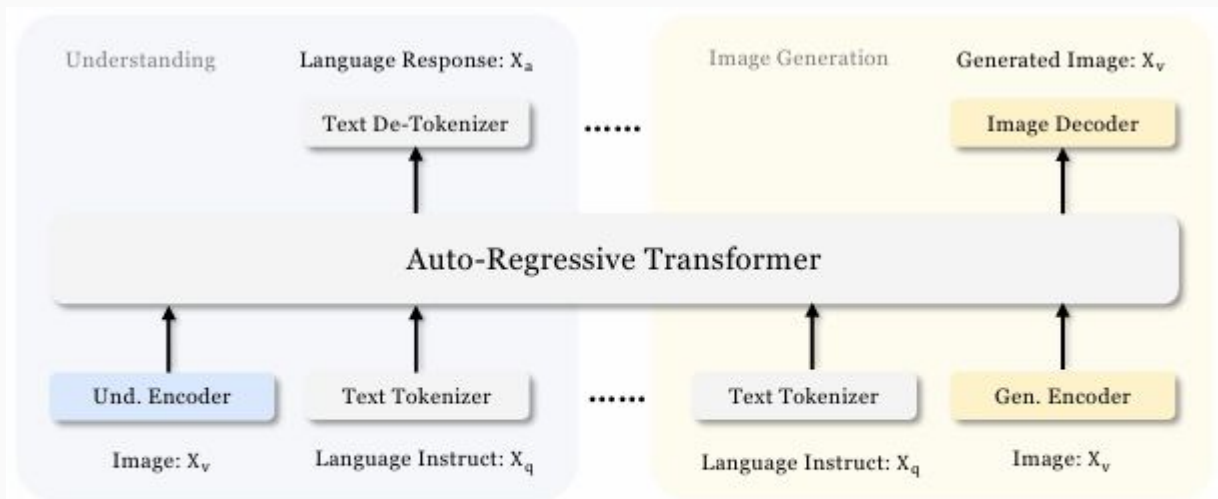


Figure 3 | **Architecture of our Janus-Pro.** We decouple visual encoding for multimodal understanding and visual generation. “Und. Encoder” and “Gen. Encoder” are abbreviations for “Understanding Encoder” and “Generation Encoder”, respectively. Best viewed on screen.

2.2 Optimized Training Strategy(1)

기존 방식 문제점

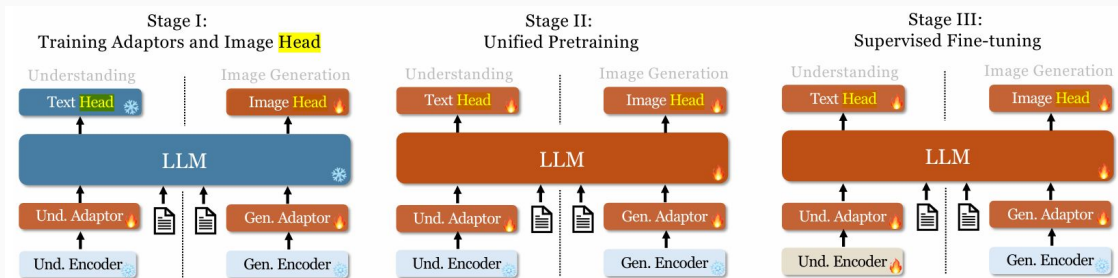


Figure 3 | Our Janus adopts a three-stage training procedure. We use flame symbols/snowflake symbols in the diagram to indicate the module updates/does not update its parameters.

- Janus는 Stage II에서 Text-to-Image 학습을 두 부분으로 나누었는데
 - ImageNet 데이터 학습: 픽셀 의존성 (Pixel Dependence)을 학습하여 이미지 생성 능력을 향상
 - 일반적인 Text-to-Image Data를 사용하여 학습

텍스트-이미지 학습 스텝 중 66.67%를 첫 번째 부분(즉, ImageNet 데이터 학습)에 할당

→ 상당한 연산 비효율성 (Computational Inefficiency)을 초래함이 확인

ImageNet 데이터 학습(픽셀 의존성 모델링)에 과도하게 많은 학습 스텝이 할당되어 전체적인 성능 최적화에 방해가 될 가능성

2.2 Optimized Training Strategy(1)

해결책

- **Stage I의 훈련 스텝(training steps)을 증가시켜 ImageNet 데이터셋에서 충분한 학습을 수행** ImageNet 데이터 활용도 ↑
 - 모델이 픽셀 의존성(pixel dependence)을 효과적으로 학습하고,
 - 카테고리 이름(category names)에 기반하여 합리적인 이미지를 생성할 수 있음이 확인됨
- **Stage II에서 집중적인 학습** ImageNet 데이터를 제거하여 효율적인 텍스트-이미지 학습을 수행
 - Stage II에서 ImageNet 데이터를 제거
 - 일반적인 텍스트-이미지 데이터(normal text-to-image data)를 직접 활용하여 학습을 진행
- **Stage III에서 지도 학습(Supervised Fine-Tuning) 과정에서, 서로 다른 유형의 데이터셋 비율을 조정** 데이터 비율을 최적화하여 이해와 생성 성능의 균형을 조정
 - **Multimodal Data:Pure Text Data:Text-to-Image Data = 5:1:4**
 - 텍스트-이미지 데이터의 비율을 약간 줄임으로써
 - 강력한 시각적 생성(Visual Generation) 능력을 유지하면서도,
 - 멀티모달 이해(Multimodal Understanding) 성능이 개선됨이 관찰

2.3 Data Scaling

- Janus의 학습 데이터를 확장하여, Multimodal Understanding와 Visual Generation 성능을 향상
 - Multimodal Understanding
 - Stage II 사전 학습(Pretraining) 데이터: DeepSeek-VL2를 참고하여 약 9천만 개(90M)의 샘플을 추가
 - 이미지 캡션 데이터셋 (예. YFCC)
 - 표(Table), 차트(Chart), 문서(Document) 이해 데이터 (예. Docmatix)
 - Stage III 지도 학습(Supervised Fine-Tuning) 데이터: DeepSeek-VL2의 추가 데이터셋을 활용하여 학습을 강화
 - MEME 이해(MEME Understanding) 데이터, 중국어 대화 데이터(Chinese Conversational Data), 대화 경험을 향상시키는 데이터(Dialogue Enhancement Datasets)
 - Visual Generation
 - Janus에서 사용된 **real-world** 데이터는 품질이 낮고 노이즈가 많아, Text-to-Image 생성에서 불안정성(Instability)이 발생하고,미적 품질(Aesthetic Quality)이 저하된 결과물이 생성
 - 이에 약 7,200만 개(72M)의 **Synthetic Aesthetic Data**를 추가
 - 통합 사전 학습(Unified Pretraining) 단계에서 실제 데이터와 합성 데이터의 비율을 1:1로 조정
 - 합성 데이터 샘플의 프롬프트(Prompts)는 공개된 소스에서 가져옴

2.4 Model Scaling

- Janus는 **1.5B(15억 개)** 파라미터를 가진 LLM에서 검증
- Janus-Pro에서는 모델을 **7B(70억 개)**로 확장
 - Multimodal Understanding와 Visual Generation에서 손실 수렴 속도(Convergence Speed of Losses)가 크게 향상
 - Janus-Pro의 접근 방식이 강력한 확장 가능성(Scalability)을 갖고 있음을 추가적으로 검증

Table 1 | **Architectural configuration for Janus-Pro.** We list the hyperparameters of the architecture.

	Janus-Pro-1B	Janus-Pro-7B
Vocabulary size	100K	100K
Embedding size	2048	4096
Context Window	4096	4096
#Attention heads	16	32
#Layers	24	30

3. Experiments

3.1 Implementation Details(1)

- 구현 세부 사항

- Base language model: 최대 4096 시퀀스 길이를 지원하는 DeepSeek-LLM (1.5B and 7B)
- vision encoder(for understanding tasks): SigLIP-Large-Patch16-384 @GOOGLE, 이미지와 텍스트를 함께 처리 할 수 있는 모델
384x384 이미지 처리, 제로샷 이미지 분류, 이미지-텍스트 검색,
- generation encoder: 16,384 크기의 codebook, 이미지를 16배 downsampling
원래 이미지 이해/연산량/정보 압축, 픽셀 공간보다 압축된 잠재 공간에서 생성
- understanding 어댑터와 generation 어댑터: two-layer MLPs
시각적 특징이나 VQ 코드북 ID를 LLM이 처리할 수 있는 텍스트 임베딩 공간으로
유리 변환
- 이미지 resize: 384x384 pixels
- multimodal understanding data에 대해 이미지의 긴 쪽을 크기 조정하고 짧은 쪽을 배경색(RGB: 127, 127, 127)
으로 패딩하여 384 크기에 맞춤
- visual generation data에 대해 짧은 쪽을 384로 크기 조정하고 긴 쪽을 384로 자름
- 학습 효율성을 높이기 위해 학습 중에 시퀀스 패킹을 사용
학습 과정에서 배치 처리의 효율성을 높이는 기술, 하나의 배치에 여러 시퀀스를 포함
- single training step에서 지정된 비율에 따라 모든 데이터 유형을 혼합
배치 1: [이미지 14와 텍스트 2의 이미지와 (384x384) 할당되어 데이터 (40%, 이미지 + 텍스트),
시각적 생성 데이터 (30%, 텍스트, “~를 그려주세요”)]
- PyTorch 기반의 경량화되고 효율적인 분산 학습 프레임워크인 HAI-LLM을 사용하여 학습되고 평가
High-flyer AI Large Language Model, 중국 환팡에서 개발, 데이터 병렬 처리, 최적화(메모리, 계산 등)를 위한 학습 프레임워크
- 전체 학습 소요 시간은 각 노드 당 8개의 Nvidia A100(40GB) GPU 사용
 - 1.5B 모델: 16개 노드에서 9일 소요
 - 7B 모델: 32개 노드에서 14일 소요

3.1 Implementation Details(2)

Table 2 | **Detailed hyperparameters for training Janus-Pro.** Data ratio refers to the ratio of multimodal understanding data, pure text data, and visual generation data.

	Janus-Pro-1B			Janus-Pro-7B		
Hyperparameters	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
Learning rate	1.0×10^{-3}	1.0×10^{-4}	4.0×10^{-5}	1.0×10^{-3}	1.0×10^{-4}	4.0×10^{-5}
LR scheduler	Constant	Constant	Constant	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0	0.0	0.0	0.0
Gradient clip	1.0	1.0	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)			AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)		
Warm-up steps	600	5000	0	600	5000	0
Training steps	20K	360K	80K	20K	360K	40K
Batch size	256	512	128	256	512	128
Data Ratio	1:0:3	2:3:5	5:1:4	1:0:3	2:3:5	5:1:4

3.2 Evaluation Setup

- Multimodal Understanding 능력 평가
 - image-based vision-language benchmarks에서 평가
 - GQA, POPE, MME, SEED, MMB, MM-Vet, MMMU

→ 모델이 이미지와 텍스트를 함께 이해하고 처리하는 능력을 다양한 벤치마크에서 평가
- Visual Generation 능력 평가
 - DPG-Bench(DensePromptGraph Benchmark)
 - 텍스트-이미지 모델의 정교한 의미 정렬 능력(Intricate Semantic Alignment Capabilities)을 평가하기 위해 설계된 포괄적인 벤치마크 데이터셋
 - 1,065개의 길고 밀도 높은 프롬프트(Lengthy, Dense Prompts)로 구성되어 있으며, 텍스트-이미지 모델의 성능을 종합적으로 평가하는 데 활용
 - GenEval
 - 이미지-텍스트 변환(Image-to-Text Generation)을 평가하는 고난이도 벤치마크
 - 모델의 종합적인 생성 능력(comprehensive generative abilities)을 평가하도록 설계
 - 개별 인스턴스 수준(instance-level)에서 모델의 구성 능력(compositional capabilities)을 분석

3.3 Comparison with State-of-the-arts

- Janus-Pro는 훨씬 더 큰 모델들과 비교해도 높은 경쟁력을 유지

Table 3 | Comparison with state-of-the-arts on multimodal understanding benchmarks. “Und.” and “Gen.” denote “understanding” and “generation”, respectively. Models using external pretrained diffusion model are marked with †.

Type	Model	# LLM Params	POPE†	MME-P†	MMB†	SEED†	GQA†	MMMU†	MM-Vet†
Und. Only	LLaVA-v1.5-Phi-1.5 [50]	1.3B	84.1	1128.0	-	-	56.5	30.7	-
	MobileVLM [6]	1.4B	84.5	1196.2	53.2	-	56.1	-	-
	MobileVLM-V2 [7]	1.4B	84.3	1302.8	57.7	-	59.3	-	-
	MobileVLM [6]	2.7B	84.9	1288.9	59.6	-	59.0	-	-
	MobileVLM-V2 [7]	2.7B	84.7	1440.5	63.2	-	61.1	-	-
	LLaVA-Phi [56]	2.7B	85.0	1335.1	59.8	-	-	-	28.9
	LLaVA [27]	7B	76.3	809.6	38.7	33.5	-	-	25.5
	LLaVA-v1.5 [26]	7B	85.9	1510.7	64.3	58.6	62.0	35.4	31.1
	InstructBLIP [8]	7B	-	-	36.0	53.4	49.2	-	26.2
	Qwen-VL-Chat [1]	7B	-	1487.5	60.6	58.2	57.5	-	-
	IDEFICS-9B [19]	8B	-	-	48.2	-	38.4	-	-
	Emu3-Chat [45]	8B	85.2	1244	58.5	68.2	60.3	31.6	37.2
	InstructBLIP [8]	13B	78.9	1212.8	-	-	49.5	-	25.6
Und. and Gen.	DreamLLM† [10]	7B	-	-	-	-	-	-	36.6
	LaVIT† [18]	7B	-	-	-	-	46.8	-	-
	MetaMorph† [42]	8B	-	-	75.2	71.8	-	-	-
	Emu† [39]	13B	-	-	-	-	-	-	-
	NEXT-GPT† [47]	13B	-	-	-	-	-	-	-
	Show-o-256 [50]	1.3B	73.8	948.4	-	-	48.7	25.1	-
	Show-o-512 [50]	1.3B	80.0	1097.2	-	-	58.0	26.7	-
	D-Dit [24]	2.0B	84.0	1124.7	-	-	59.2	-	-
	Gemini-Nano-1 [41]	1.8B	-	-	-	-	26.3	-	-
	ILLUME [44]	7B	88.5	1445.3	65.1	72.9	-	38.2	37.0
	TokenFlow-XL [34]	13B	86.8	1545.9	68.9	68.7	62.7	38.7	40.7
	LWM [28]	7B	75.2	-	-	-	44.8	-	9.6
	VILA-U [48]	7B	85.8	1401.8	-	59.0	60.8	-	33.5
	Chameleon [40]	7B	-	-	-	-	22.4	8.3	-

Table 4 | Evaluation of text-to-image generation ability on GenEval benchmark. “Und.” and “Gen.” denote “understanding” and “generation”, respectively. Models using external pretrained diffusion model are marked with †.

Type	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall†
Gen. Only	LlamaGen [38]	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [37]	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [37]	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt- α [4]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [37]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [35]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [45]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [32]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [2]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [11]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Und. and Gen.	SEED-X† [13]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o [50]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	D-Dit [24]	0.97	0.80	0.54	0.76	0.32	0.50	0.65
	LWM [28]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	Transfusion [55]	-	-	-	-	-	-	0.63
	ILLUME [44]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
	TokenFlow-XL [28]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	Chameleon [40]	-	-	-	-	-	-	0.39
	Janus [46]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Janus-Pro-1B	0.98	0.82	0.51	0.89	0.65	0.56	0.73
	Janus-Pro-7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80


Table 5 | Performances on DPG-Bench. The methods in this table are all generation-specific models except Janus and Janus-Pro.

Method	Global	Entity	Attribute	Relation	Other	Overall†
SDv1.5 [36]	74.63	74.23	75.39	73.49	67.81	63.18
PixArt- α [4]	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next [57]	82.82	88.65	86.44	80.53	81.82	74.63
SDXL [33]	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 [22]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-Dit [25]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt- Σ [5]	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen [45]	85.21	86.68	86.84	90.22	83.15	80.60
DALL-E 3 [2]	90.97	89.61	88.39	90.58	89.83	83.50
SD3-Medium [11]	87.90	91.01	88.83	80.70	88.68	84.08
Janus	82.33	87.38	87.70	85.46	86.41	79.68
Janus-Pro-1B	87.58	88.63	88.17	88.98	88.30	82.63
Janus-Pro-7B	86.90	88.90	89.40	89.32	89.48	84.19

3.4 Qualitative Results

- Janus-Pro는 멀티모달 이해뿐만 아니라, 창의적이고 논리적인 텍스트-이미지 생성에서도 뛰어난 성능을 발휘

Image Description




This image depicts a **stunning coastal landscape** featuring a series of **tall, rugged rock** formations emerging from the ocean. The rocks are layered with sedimentary rock, giving them a distinct stratified appearance. The largest formation is prominently positioned in the foreground, with smaller ones receding into the distance. The ocean waves crash against the base of these formations, creating **white foam and a dynamic sense of movement**.

The **coastline is sandy**, with a narrow strip of beach visible on the right side of the image. The beach is bordered by steep cliffs, which rise dramatically from the shoreline. The cliffs are covered in patches of green vegetation, adding a touch of color to the otherwise earthy tones of the rocks.

The sky above is **overcast**, with thick clouds that suggest an impending storm or recent rainfall. The light is diffused, casting a soft, muted glow over the entire scene. Despite the cloudy sky, there is a sense of tranquility and natural beauty in this coastal landscape.

Describe this scene in detail.


Landmark Recognition



这张图片展示的是中国杭州的西湖，特别是著名的三潭印月岛。图片中的建筑是三潭印月的亭子，周围是平静的湖水，远处是连绵的山脉。西湖是中国著名的旅游景点之一，以其美丽的自然景观和丰富的历史文化而闻名。

猜猜这是哪里？

General Knowledge



This cake is themed around the classic cartoon character **Tom and Jerry**. The cake features the characters Tom (the cat) and Jerry (the mouse) in a playful setting, with a backdrop that includes the words "Tom & Jerry". The cake itself is decorated with fondant shaped like cheese, which is a recurring theme in the cartoon series. The design also includes a small fondant house with a door, adding to the whimsical and playful theme of the cake.

Can you introduce the background story of this cake?

Text Recognition




What's written on the large black board?

The large black board has the following text:


"Serving Soul since Twenty Twelve"

Additionally, there are smaller sections with more detailed information, but the main heading is clearly visible.


Text-To-Image Generation




A golden retriever lying peacefully on a wooden porch, with autumn leaves scattered around.




A young woman with freckles wearing a straw hat, standing in a golden wheat field.




A single drop of water clinging to a green leaf, with sunlight creating a faint rainbow prism.




An ancient stone bridge arching over a crystal-clear mountain stream, surrounded by lush greenery.




A glowing crystal ball floating above a sandstone table in the middle of a desert at sunset.



A tiny galaxy contained inside a glass bottle, glowing brightly against a dark velvet cloth.



A giant whale flying through a city skyline, surrounded by floating glowing lanterns.



Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

Figure 4 | **Qualitative results of multimodal understanding and visual generation capability.** The model is Janus-Pro-7B and the image output resolution of visual generation is 384 × 384. Best viewed on screen.

4. Conclusion

- Janus의 세 가지 측면에서의 개선 사항(Training Strategy, Data, Model Size)
→ 이를 통해 Multimodal Understanding과 Text-to-Image Instruction-Following 능력 크게 향상
- Janus-Pro의 한계점
 - Input Resolution가 384×384 로 제한
OCR(Optical Character Recognition, 광학 문자 인식)과 같은 세밀한 작업(Fine-Grained Tasks)에서 성능이 저하
 - Text-to-Image Generation 한계
 - Semantic Content는 풍부하지만 Low Resolution와 Vision Tokenizer에서 발생하는 Reconstruction Losses로 인해 세부 묘사가 부족
(예. Small Facial Regions이 제한된 이미지 공간을 차지할 경우, 디테일이 부족)
→ 이미지 해상도 증가 필요