JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation

Yiyang Ma[1,2], Xingchao Liu[1,†], Xiaokang Chen[1,†], Wen Liu[1,†], Chengyue Wu[1,3], Zhiyu Wu[1], Zizheng Pan[1], Zhenda Xie[1], Haowei Zhang[1], Xingkai Yu[1], Liang Zhao[1], Yisong Wang[1,4], Jiaying Liu[2], Chong Ruan[1,‡]
[1]**DeepSeek-AI** [2]**Peking University** [3]**The University of Hong Kong** [4]**Tsinghua University**
[†]Equal Contribution [‡]Corresponding Author
Project Page: https://github.com/deepseek-ai/Janus

Abstract

*We present **JanusFlow**, a powerful framework that unifies image understanding and generation in a single model. JanusFlow introduces a minimalist architecture that integrates autoregressive language models with rectified flow, a state-of-the-art method in generative modeling. Our key finding demonstrates that rectified flow can be straightforwardly trained within the large language model framework, eliminating the need for complex architectural modifications. To further improve the performance of our unified model, we adopt two key strategies: (i) decoupling the understanding and generation encoders, and (ii) aligning their representations during unified training. Extensive experiments show that JanusFlow achieves comparable or superior performance to specialized models in their respective domains, while significantly outperforming existing unified approaches. This work represents a step toward more efficient and versatile vision-language models.*

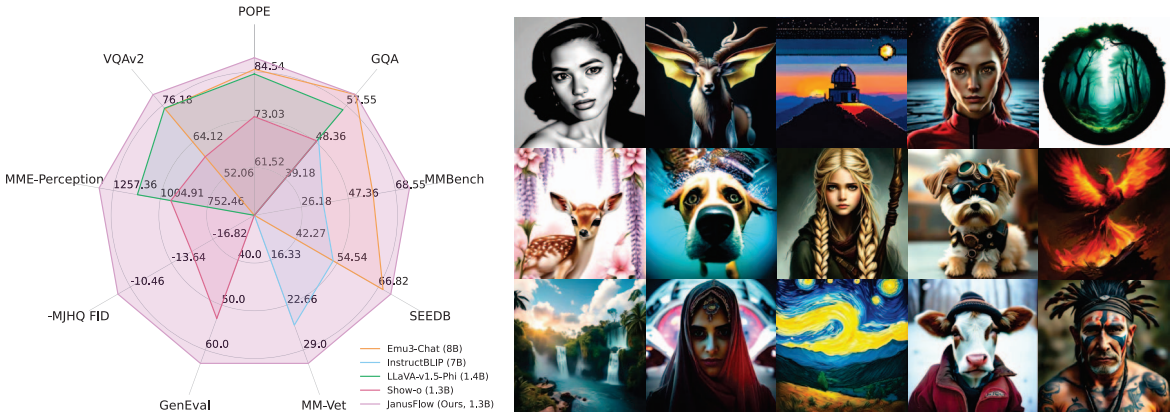
1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in learning diverse knowledge and generalizing to new scenarios [1, 7, 8, 68, 89]. Leveraging these capabilities, researchers have developed sophisticated models specialized in image comprehension [2, 15, 46, 48, 55, 57] and text-to-image generation [23, 72, 74, 77].

The field has recently shifted toward creating unified systems capable of handling both tasks simultaneously. One prominent direction involves utilizing pre-trained text-to-image models for high-quality generation while training LLMs to generate conditions for these models [19, 25–27, 85]. However, this approach introduces architectural complexity and potentially constrains the model's capabilities through maintaining separate LLM and generative components. Alternative approaches [86, 95, 97, 98, 106] propose training a single LLM for both tasks, typically incorporating either diffusion models [32, 81] or vector-quantized autoregressive models [22, 84].

Our approach builds upon recent breakthroughs in rectified flow models [3, 23, 54, 60, 61], which provide a simple framework for generative modeling while delivering exceptional empirical performance [23, 36, 44]. Building on these advances, we propose **JanusFlow**, a powerful unified multimodal model that seamlessly integrates rectified flow with LLM architecture. Following a minimalist design principle, our architecture requires only a lightweight encoder and decoder to adapt the LLM for rectified flow operations. To optimize JanusFlow's performance, we implement two key strategies: First, we maintain separate vision encoders for understanding and generation tasks, preventing task interference and thus enhancing understanding capabilities. Second, we align the intermediate features between generation and understanding modules during training, strengthening semantic coherence in the generation process.

JanusFlow shows state-of-the-art performances in both multimodal understanding and text-to-image generation compared to existing unified approaches, and even outperforms several specialized methods. Specifically, on image generation benchmarks, MJHQ FID-30k [47], GenEval [28] and DPG-Bench [34], JanusFlow achieves scores of 9.51, 0.63 and 80.09%, surpassing established text-to-image models including SDv1.5 [75] and SDXL [72]. In multimodal comprehension benchmarks, JanusFlow attains scores of 74.9, 70.5 and 60.3 on MMBench [62], SeedBench [45], and GQA [35], respectively, exceeding specialized models such as LLaVA-v1.5 [55] and Qwen-VL-Chat [4]. Notably, these results are achieved with a compact LLM architecture with only 1.3B parameters.



(a) Benchmark Performances.

(b) Visual Generation Results.

Figure 1. **Multimodal understanding and image generation with JanusFlow**. JanusFlow surpasses the state-of-the-art unified multimodal models and several task-specific understanding models on visual understanding benchmarks. It is also capable of generating high-quality images. The resolution of the images is 384×384.

2. Related Work

Visual Generation with Flow-Based Generative Models. Recent years have witnessed remarkable progress in visual generation via diffusion models [32, 81], leading to impressive models like [66, 72, 74–77]. Furthermore, flow-based models [3, 54, 60] emerged as a simplified alternative. These approaches have enabled advanced visual generation models [23, 36] that achieve superior performance with faster sampling. Our work proves that rectified flow [59–61] can be effectively integrated into LLMs, creating unified models that excel in both of understanding and generation.

Unified Models for Understanding and Generation. The development of multimodal large language models (MLLMs) has enabled effective integration of text and visual information. Building upon powerful LLMs [7, 89, 90], recent MLLMs [2, 15, 48, 55, 57, 63] have demonstrated exceptional multimodal understanding capabilities. Current research increasingly focuses on architectures that can simultaneously handle visual understanding and generation tasks. One approach extends MLLMs with pre-trained diffusion models [19, 25–27, 85, 99]. However, these systems essentially utilize diffusion models as external tools, where the MLLM generates conditions for image generation without possessing direct generative capabilities. This separation often results in suboptimal performance compared to standalone diffusion models [25, 85]. Another line of work [86, 95, 97, 98, 106] aim to train a single LLM for both tasks. Many of these methods employ vector-quantization [22, 84] to convert images into discrete tokens, enabling unified autoregressive processing [86, 95]. While straightforward to implement, these approaches are inherently limited by their image tokenization quality.

Our work focuses on developing unified models that combine autoregression with flow/diffusion models, leveraging their effectiveness in generation. Compared with similar approaches [98, 105], JanusFlow offers three key advantages: (i) a simple yet effective generation process with rectified flow, (ii) enhanced performance through decoupled vision encoders that resolve task conflicts, and (iii) improved generation quality through representation alignment regularization, enabled by our decoupled encoder design.

3. JanusFlow

3.1. Background

Multimodal LLMs. Given a dataset $\mathcal{D}$ containing discrete token sequences, each of which can be formulated as $x = (x_1, \cdots, x_\ell)$, large language models (LLMs) are trained to model the distribution autoregressively,

$$\log \mathrm{P}_{\theta_{LLM}}(x) = \sum_{i=0}^{\ell-1} \log \mathrm{P}_{\theta_{LLM}}(x_{i+1}|x_1, \ldots, x_i), \quad (1)$$

where θ_{LLM} denotes the parameters of the LLM and ℓ is the sequence length. After being trained on large-scale datasets, LLMs exhibit the ability to generalize across various tasks and follow diverse instructions [1, 8, 68]. To extend these models to handle visual inputs, LLMs are augmented with vision encoders [2, 55, 57]. For instance, LLaVA [57] integrates an LLM with a pre-trained CLIP [73] image encoder via a projection layer, transforming the extracted image features into a joint embedding space. By leveraging large-scale multimodal datasets and increasingly powerful LLMs, this architecture has facilitated the development of advanced multimodal models capable of addressing a wide range of vision-language tasks [4, 46, 55, 63].

Rectified Flow. For a dataset $\mathcal{D}$ consisting of continuous d-dimensional data points $x = (x_1, \cdots, x_d)$ drawn from an

unknown data distribution π_1, rectified flow [54, 60] models the data distribution by learning an ordinary differential equation (ODE) defined over time $t \in [0, 1]$:

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = v_{\theta_{NN}}(z_t, t), \quad z_0 \sim \pi_0, \tag{2}$$

where θ_{NN} represents the parameters of the neural network and π_0 is a simple distribution, typically standard Gaussian $\mathcal{N}(0, I)$. The network is trained to minimize the Euclidean distance between the neural velocity and the directions of linear paths connecting random points from π_0 and π_1,

$$\min_{\theta} \mathbb{E}_{t\sim \mathrm{P}(t), z_0\sim\pi_0, x\sim\pi_1} \left[||v_{\theta_{NN}}(z_t, t) - (x - z_0)||^2 \right], \\ \text{where } \; z_t = tx + (1-t)z_0, \tag{3}$$

where $\mathrm{P}(t)$ is a distribution over time $t \in [0, 1]$. When the network has sufficient capacity and the objective is perfectly minimized, the optimal velocity field $v_{\theta^*_{NN}}$ maps the elementary distribution π_0 to the true data distribution π_1. More precisely, the distribution of $z_1 = \int_0^1 v_{\theta^*_{NN}}(z_t, t)\mathrm{d}t$, with $z_0 \sim \pi_0$, follows π_1. Despite its conceptual simplicity, rectified flow has shown superior performance in various generative modeling tasks, including text-to-image generation [23] and audio generation [39].

3.2. A Unified Framework for Multimodal Understanding and Generation

JanusFlow presents a unified framework designed to address both vision understanding and image generation tasks.

Multimodal Understanding. In multimodal understanding tasks, the LLM processes an input sequence consisting of interleaved text and image data. The text is tokenized into discrete tokens, each of which is transformed into an embedding of dimension D_{emb}. For the images, an image encoder f_{enc} encodes each image x_{im} into a feature map of shape $H_{im} \times W_{im} \times D_{enc}$. This feature map is flattened and projected through a linear transformation layer into a sequence of embeddings with shape $H_{im}W_{im} \times D_{emb}$. H_{im} and W_{im} are determined by the image encoder. The text and image embeddings are concatenated to form the input sequence to the LLM, which then autoregressively predicts the next tokens based on the input sequence of embeddings. According to common practice [86, 95, 98], we add special token `|BOI|` and `|EOI|` before and after the image to help the model locate the image embeddings in the sequence.

Image Generation. For image generation, our LLM takes a text sequence x^{con} as condition and generates a corresponding image using rectified flow. To improve computational efficiency, generation occurs in the latent space of the pre-trained SDXL-VAE [72].

The generation process begins by sampling Gaussian noise z_0 of shape $H_{latent} \times W_{latent} \times D_{latent}$ in the latent space, which is then processed by a generation encoder g_{enc} into a sequence of embeddings $H_{gen}W_{gen} \times D_{emb}$. This sequence is concatenated with a time embedding representing the current time step t ($t = 0$ at the beginning), resulting in a sequence of length $H_{gen}W_{gen} + 1$. Unlike previous approaches that employ various attention masking strategies [98, 106], we found that causal attention suffices, as our preliminary experiments showed no performance benefits from alternative masking schemes. The LLM's output corresponding to z_0 is transformed back into the latent space by a generation decoder g_{dec}, producing a velocity vector of shape $H_{latent} \times W_{latent} \times D_{latent}$. The state is updated by a standard Euler solver,

$$z_{t+\mathrm{d}t} = z_t + v(z_t, t)\mathrm{d}t, \tag{4}$$

where $\mathrm{d}t$ is a user-defined step size. We replace z_0 with $z_{\mathrm{d}t}$ on the input and iterate the process until we get z_1, which is then decoded into the final image by the VAE decoder. To enhance generation quality, we employ classifier-free guidance (CFG) when computing the velocity:

$$v(z_t, t) = wv(z_t, t \mid x^{con}) + (1 - w)v(z_t, t \mid \varnothing), \tag{5}$$

where $v(z_t, t \mid \varnothing)$ denotes the velocity inferred without text conditioning and $w \geq 1$ controls the magnitute of CFG. Empirically, increasing w yields higher semantic alignment [23, 61, 72, 75]. Analogous to multimodal understanding, we prepend the special token `|BOI|` to indicate the start of image generation in the sequence.

Decoupling Encoders for the Two Tasks. Previous approaches that unify autoregressive generation and diffusion models within a joint LLM training framework [98, 106] employ identical encoders (f_{enc} and g_{enc}) for both of the tasks. For instance, Zhou et al. [106] performs both tasks in the same VAE latent space using a shared U-Net or linear encoder, while Xie et al. [98] leverages MAGVIT-v2 [100] to encode image patches into discrete tokens for both tasks.

However, recent work on unified autoregressive models has shown this shared encoder design to be suboptimal [95], particularly in models that generate images through autoregression on vector-quantized tokens. Drawing from these insights, JanusFlow adopts a decoupled encoder design. Specifically, we employ a pre-trained SigLIP-Large-Patch/16 [104] model as f_{enc} to extract semantic continuous features for multimodal understanding, while using separate ConvNeXt blocks [94] initialized from scratch as g_{enc} and g_{dec} for generation, chosen for its effectiveness. Following established practices [5, 14, 91], we incorporate a long skip connection between g_{enc} and g_{dec}. Ablation studies in Sec. 4.5 demonstrate that this decoupled encoder design significantly improves the performance of our model. The complete architecture of JanusFlow is illustrated in Fig. 2.

3.3. Training Schemes

As illustrated in Fig. 3, we train our model in three sequential stages, detailed below.

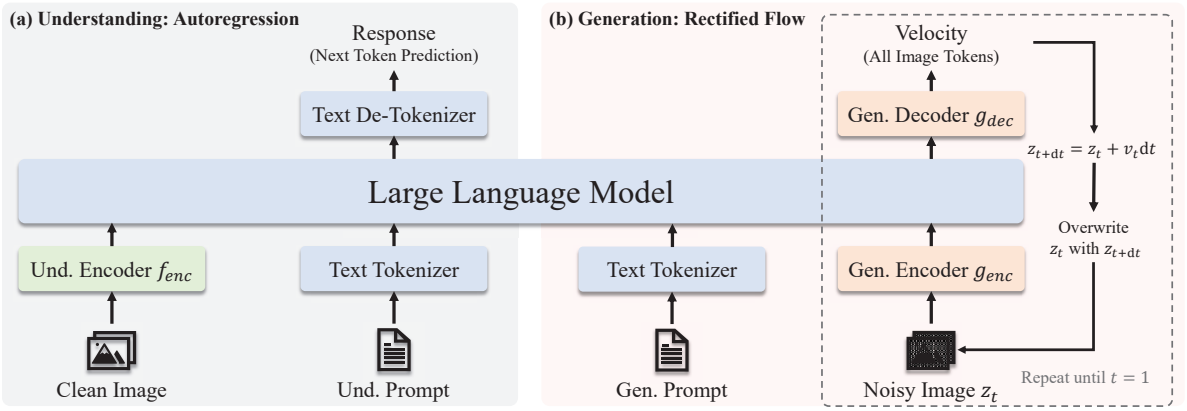


Figure 2. **Architecture of the proposed JanusFlow.** For visual understanding, the LLM performs autoregressive next-token prediction to generate responses. For image generation, the LLM employs images with rectified flow. Starting from Gaussian noise at $t = 0$, the LLM iteratively updates z_t by predicting velocity vectors until reaching $t = 1$. We omit the VAE encoder, the skip connection leveraged in generation and the linear layer after f_{enc} for simplicity.

Stage 1: Adaptation of Randomly Initialized Components. In the first stage, we focus on training only the randomly initialized components: the linear layers, generation encoder, and generation decoder. This stage serves to adapt these new modules to work effectively with the pre-trained LLM and SigLIP encoder, essentially functioning as an initialization phase for the newly introduced components.

Stage 2: Unified Pre-Training. Following the adaptation stage, we train the entire model except for the visual encoder, consistent with previous approaches [57, 63]. The training incorporates three data types: multimodal understanding, image generation, and text-only data. We initially allocate a higher proportion of multimodal understanding data to establish the model's understanding capabilities. Subsequently, we increase the ratio of image generation data to accommodate the convergence requirements of diffusion-based models [18, 71].

Stage 3: Supervised Fine-Tuning (SFT). In the final stage, we fine-tune the pre-trained model using instruction tuning data, which comprises dialogues, task-specific conversations, and high-quality image generation examples. During this stage, we also unfreeze the SigLIP encoder parameters [63, 88, 95]. This fine-tuning process enables the model to effectively respond to user instructions for both multimodal understanding and image generation tasks.

3.4. Training Objective

Training JanusFlow involves two types of data, multimodal understanding data and image generation data. Both types of data contain two parts: "condition" and "response". "Condition" refers to the prompting of the tasks (e.g. text prompts in the task of generation and images in the task of understanding) while "response" refers to the corresponding responses of the two tasks. The data can be formatted as $x = (x^{con}, x^{res})$, where the superscript con denotes "condition" and res denotes "response". We denote the length of the whole sequence x as ℓ, the length of x^{con} as ℓ_{con} and the length of x^{res} as ℓ_{res}. We use θ to represent the collection of all the trainable parameters in JanusFlow, including the LLM, f_{enc}, g_{enc}, g_{dec} and the linear transformation layers.

Autoregression Objective. For mutimodal understanding tasks, x^{res} contains only text tokens. JanusFlow is trained using the maximum likelihood principle,

$$\mathcal{L}_{AR}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_{und}} \left[\sum_{i=\ell_{con}}^{\ell-1} \log \mathrm{P}_\theta \left(x_{i+1} | x_1, \ldots, x_i\right) \right], \tag{6}$$

where the expectation is taken over all (x^{con}, x^{res}) pairs in our multimodal understanding dataset $\mathcal{D}_{und}$, computing loss only over tokens in x^{res}.

Rectified Flow Objective. For image generation tasks, x^{con} consists of text tokens and x^{res} is the corresponding image. JanusFlow is trained with the RF objective,

$$\mathcal{L}_{RF}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{gen}, t\sim\mathrm{P}(t), z_0\sim\mathcal{N}(0,I)} \left[||v_\theta(z_t, t \mid x^{con}) - (x^{res} - z_0)||^2 \right], \tag{7}$$

where $z_t = t x^{res} + (1-t) z_0$. Following Stable Diffusion 3 [23], we set the time distribution $\mathrm{P}(t)$ to the logit-normal distribution. To enable CFG inference, we randomly drop 10% of the text prompts in training.

Representation Alignment Regularization. Recent work [101] has shown that aligning intermediate representations between diffusion transformers and semantic vision encoders enhances diffusion model generalization. Our decoupled vision encoder design enables efficient implementation of this alignment. Specifically, for generation tasks,

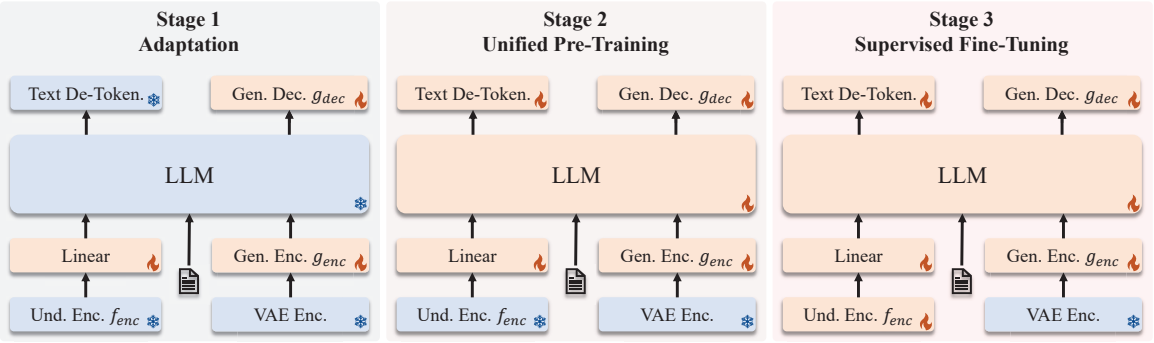


Figure 3. **Three training stages of JanusFlow.** The trainable modules are with flame and the frozen modules are with snowflakes.

we align features from the understanding encoder f_{enc} with the LLM's intermediate features,

$$\mathcal{L}_{REPA}(\theta, \phi) = -\mathbb{E}_{x \sim \mathcal{D}_{gen}} \left[\text{sim}\left(\text{stop\_grad}(f_{enc}(x^{res})), h_\phi(q_\theta(z_t)) \right) \right], \quad (8)$$

where $q_\theta(z_t)$ denotes an intermediate LLM representation given input z_t, and h_ϕ is a small trainable MLP that projects $q_\theta(z_t)$ to dimension D_{enc}. The function $\text{sim}(\cdot, \cdot)$ computes the mean of element-wise cosine similarity between embeddings. Before computing the loss, we reshape $h_\phi(q_\theta(z_t))$ to $H_{gen} \times W_{gen} \times D_{enc}$. To simplify the implementation, we intentionally adjust the configuration of g_{enc} and g_{dec} to ensure $H_{gen} = H_{im}$ and $W_{gen} = W_{im}$. The gradient of $\mathcal{L}_{REPA}$ is not back-propagated through the understanding encoder. This alignment loss helps the LLM's internal feature space (given noisy input z_t) align with the understanding encoder's semantic feature space, thereby improving generation quality when producing images from new random noise and text conditions during inference.

Summary. All three objectives are applied across all training stages. Multimodal understanding tasks use $\mathcal{L}_{AR}$, while image generation tasks employ the loss of $\mathcal{L}_{RF} + \mathcal{L}_{REPA}$. The losses are directly summed with ratio $1 : 1 : 1$.

4. Experiments

4.1. Experiment Setup and Implementation Details

JanusFlow builds upon the LLM of DeepSeek-LLM (1.3B) [7, 31]. The LLM consists of 24 transformer blocks and supports a sequence length of $4,096$. In our model, both understanding and generation exploits images of resolution 384. For multimodal understanding, we use SigLIP-Large-Patch/16 [104] as f_{enc}. For image generation, we utilize the SDXL-VAE [72] for its latent space. Table 1 details the hyper-parameters for each training stage. In the alignment regularization, we use the LLM features after the 6-th block as $q_\theta(z_t)$ and a 3-layer MLP as h_φ. More implementation details refer to the appendix.

Table 1. **Hyper-parameters of the training stages of JanusFlow**. Data ratio denotes the proportion of multimodal understanding data, image generation data and text-only data. In the initial $10,000$ steps of Stage 2, we apply a data ratio of $30 : 50 : 20$ to boost the understanding ability.

	Stage 1	Stage 2	Stage 3
Learning Rate	1.0×10^{-4}	1×10^{-4}	2.0×10^{-5}
LR Scheduler	Constant	Constant	Constant
Weight Decay	0.0	0.0	0.0
Gradient Clip	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)		
Warm-up Steps	$2,000$	$2,000$	$1,000$
Training Steps	$10,000$	$380,000$	$26,000$
Batch Size	512	512	256
Data Ratio	$50 : 50 : 0$	$14 : 80 : 6$	$21 : 70 : 9$

4.2. Training Data Settings

Data for Stage 1 and Stage 2. The first two stages of our framework uses three types of data: multimodal understanding data, image generation data and text-only data.

1. **Multimodal Understanding Data.** This type of data contains several sub-categories: (a) Image caption data. We incorporate caption datasets from [20, 40, 49, 50, 52, 80] and generate additional captions for images from [16, 42] using open-source multimodal understanding models. The names of the datasets are provided in the supplementary materials. The data follows template formats, e.g. "`<image>Generate the caption of this picture. <caption>`". (b) Charts and tables. We directly adopt the chart and table data from the training data of DeepSeek-VL [63]. (c) Task data. ShareGPT4V [11] data is utilized to facilitate basic question-answering capabilities during pre-training, structured as "`<image><question><answer>`". (d) Interleaved text-image data. This sub-category is sourced from [41, 82].
2. **Image Generation Data.** Our image generation dataset combines high-quality images from [16, 21, 40, 42, 67, 70, 80, 83] and 2 million in-house data. We enhance them with machine-generated captions using multimodal

understanding models. We filter the images in [16, 80] with aspect ratios and aesthetic scores, retaining approximately 20% of the original datasets. 25% of the data contains single-sentence captions. These kind of data assist the model to be able to process short prompts. All the data points are formatted as "<prompt><image>".

3. **Text-Only Data.** We directly use the text corpus of DeepSeek-LLM [7] without modification.

Data for Stage 3. The SFT stage also uses three types:

1. **Multimodal Instruction Data.** We leverage the instruction tuning datasets from [29, 33, 35, 46, 64, 78].
2. **Image Generation Data.** We reformat the high-quality text-image pairs from [16, 80, 83] into an instruction format: "User:<user prompt>\n\n Assistant:<image>".
3. **Text-Only Data.** We directly incorporate the text-only data from [46] without modification.

4.3. Evaluation Settings

Image Generation. We evaluate the generated images using both visual quality and semantic accuracy metrics. For quality assessment, we employ the Fréchet Inception Distance [30] (FID) metric and compute FID between $30,000$ generated images and their corresponding reference images from the MJHQ dataset [47]. The FID computation follows the implementation from GigaGAN [38]. To evaluate semantic accuracy, we utilize two benchmarks: GenEval [28] and DPG-Bench [34]. These frameworks are designed to assess whether the generated images accurately contain the objects and relationships specified in the input prompts, providing a broad evaluation of the generation capabilities.

Multimodal Understanding. We evaluate JanusFlow's multimodal understanding abilities across a diverse set of vision-language benchmarks including POPE [51], MME [24], MMBench [62], SEEDBench [45], VQAv2 [29], GQA [35], MM-Vet [102], MMMU [103], ChartQA[69] and TextVQA[79].

4.4. Quantitative Results

Image Generation Performances. We report the performances on GenEval, DPG-Bench and MJHQ FID-30k. In Tab. 2, we give comparisons on the overall scores of GenEval and DPG-Bench. The detailed scores of all the sub-tasks refer to the appendix. JanusFlow achieves an overall score of 0.63 on GenEval, surpassing the previous unified framework and several generation specific models including SDXL [72] and DALL-E 2 [74]. The results on GenEval and DPG-Bench demonstrate the ability of instruction following of our model. We give the comparisons on MJHQ FID-30k in Tab. 3. The images which are sampled to calculate FID are generated with a CFG factor $w = 2$ and a number of sampling steps 30. We sweep the CFG factor and the sampling steps and provide the results in the appendix.

Table 2. **Performances on GenEval and DPG-Bench.** "Gen." denotes "generation" while "Unified" denotes unified understanding and generation models. Models using external pre-trained generative models are signed with †

Type	Method	Params	GenEval↑	DPG↑
Gen. Only	LlamaGen [84]	0.8B	0.38	-
	LDM [75]	1.4B	0.37	-
	SDv1.5 [75]	0.9B	0.43	63.18
	PixArt-α [9]	0.6B	0.48	71.11
	SDv2.1 [75]	0.9B	0.50	-
	DALL-E 2 [74]	6.5B	0.52	-
	Emu3-GEN [93]	8B	0.54	80.60
	SDXL [72]	2.6B	0.55	74.65
	IF-XL [17]	4.3B	0.61	-
	DALL-E 3 [6]	-	0.67	83.50
	Lumina-Next [108]	2B	-	74.63
	Playground v2.5 [47]	2.6B	-	75.47
	Hunyuan-DiT [53]	1.5B	-	78.87
	PixArt-Σ [10]	0.6B	-	80.54
Unified	Chameleon [86]	34B	0.39	-
	LWM [58]	7B	0.47	-
	SEED-X† [27]	17B	0.49	-
	Show-o [98]	1.3B	0.53	-
	Janus [95]	1.3B	0.61	-
	Transfusion [106]	7.3B	0.63	-
	JanusFlow	1.3B	0.63	80.09

Table 3. **Results of MJHQ FID-30k.** The models which have similar scales to our model are marked with blue background. JanusFlow achieves the best FID among 1.3B models.

Method	Params	FID↓
LWM [58]	7B	17.77
VILA-U 256 [97]	7B	12.81
VILA-U 384 [97]	7B	7.69
Show-o [98]	1.3B	15.18
Janus [95]	1.3B	10.10
JanusFlow (Ours)	1.3B	9.51

Our method achieves the best performance among all the models with 1.3B LLM. The results prove that the rectified flow is able to improve the quality of generated images over autoregressive models such as Janus [95].

Multimodal Understanding Performances. We show comparisons of our method and other methods including understanding-specific models and unified models in Tab. 4. Our results demonstrate that our method harmonizes autoregressive LLM and rectified flow, achieving satisfying performance in both understanding and generation.

4.5. Ablation Studies

We conduct comprehensive ablation studies to validate the effectiveness of our key design choices. For computational efficiency, all ablation experiments are performed on 256×256 resolution images[1]. All models are trained on our unified pre-training dataset for $50,000$ iterations, except for the understanding-only and generation-only variants, which are trained for proportionally fewer iterations based on their respective data ratios in the pre-training phase. The quantitative results of are presented in Tab. 5.

[1]The understanding encoders in the 256×256-based ablation studies is also SigLIP-Large-Patch/16 which is pre-trained on 256×256 images.

Table 4. **Comparison with other methods on multimodal understanding benchmarks**. "Und." denotes "understanding" and "Unified" denotes unified understanding and generation models. The models employing external pre-trained generative models are marked with †. The models with LLMs which have similar number of parameters to us are marked with blue background under the line of dashes.

Type	Model	LLM Param	POPE	MME-P	MMB_{dev}	SEED	$\text{VQAv2}_{\text{test}}$	GQA	MMMU	MM-Vet	ChartQA	TextVQA
Und. Only	MobileVLM [12]	2.7B	84.9	1288.9	59.6	-	-	59.0	-	-	-	47.5
	MobileVLM-V2 [13]	2.7B	84.7	1440.5	63.2	-	-	61.1	-	-	-	57.5
	LLaVA-Phi [107]	2.7B	85.0	1335.1	59.8	-	71.4	-	-	28.9	-	48.6
	LLaVA [57]	7B	76.3	809.6	38.7	33.5	-	-	-	25.5	-	-
	LLaVA-v1.5 [55]	7B	85.9	1510.7	64.3	58.6	78.5	62.0	35.4	31.1	-	58.2
	InstructBLIP [15]	7B	-	-	36.0	53.4	-	49.2	-	26.2	-	50.1
	Qwen-VL-Chat [4]	7B	-	1487.5	60.6	58.2	78.2	57.5	-	-	66.3	61.5
	LLaVA-NeXT [56]	7B	-	1519.3	-	-	-	-	35.1	-	54.8	-
	Qwen2-VL [92]	7B	-	-	-	-	-	-	54.1	62.0	83.0	84.3
	IDEFICS-9B [43]	8B	-	-	48.2	-	50.9	38.4	-	-	-	25.9
	Emu3-Chat [93]	8B	85.2	-	58.5	68.2	75.1	60.3	31.6	-	68.6	64.7
	InstructBLIP [15]	13B	78.9	1212.8	-	-	-	49.5	-	25.6	-	50.7
	LLaVA-v1.5-Phi-1.5 [98]	1.3B	84.1	1128.0	-	-	75.3	56.5	30.7	-	-	-
	MobileVLM [12]	1.4B	84.5	1196.2	53.2	-	-	56.1	-	-	-	41.5
	MobileVLM-V2 [13]	1.4B	84.3	1302.8	57.7	-	-	59.3	-	-	-	52.1
Unified	Gemini-Nano-1 [87]	1.8B	-	-	-	-	62.7	-	26.3	-	53.6	62.5
	LWM [58]	7B	75.2	-	-	-	55.8	44.8	-	9.6	-	-
	VILA-U [97]	7B	85.8	1401.8	-	59.0	79.4	60.8	-	33.5	-	60.8
	Chameleon [86]	7B	-	-	-	-	-	-	22.4	8.3	-	-
	DreamLLM† [19]	7B	-	-	-	-	72.9	-	-	36.6	-	41.8
	LaVIT† [37]	7B	-	-	-	-	66.0	46.8	-	-	-	-
	Emu† [85]	13B	-	-	-	-	52.0	-	-	-	-	-
	NExT-GPT† [96]	13B	-	-	-	-	66.7	-	-	-	-	-
	Show-o [98]	1.3B	73.8	948.4	-	-	59.3	48.7	25.1	-	-	-
	Janus [95]	1.3B	87.0	1338.0	69.4	63.7	77.3	59.1	30.5	34.3	-	-
	JanusFlow (Ours)	1.3B	88.0	1333.1	74.9	70.5	79.8	60.3	29.3	30.9	64.6	55.5

Table 5. **Ablation studies.** The weights of the modules with † are frozen during training. "Exp." denotes "experiment". "FID" in this table is MJHQ FID-10k with CFG factor $w = 7.5$ and 30 steps. "CLIP" denotes CLIP similarity with the backbone of CLIP-ViT-Large-Patch/14. Exp. F is the final configuration for training JanusFlow.

Exp. ID	Model Setting				Train. Iter.	Evaluation Benchmarks				
	REPA	Und. Modules	Gen. Modules	Type		POPE↑	VQAv2_{val} ↑	GQA↑	FID↓	CLIP ↑
A	×	SigLIP	VAE†+ConvNeXt	Unified	50,000	82.40	69.62	54.43	19.84	24.94
B	✓	Shared VAE†+ConvNeXt		Unified	50,000	78.13	53.94	44.04	18.05	26.38
C	✓	VAE+ConvNeXt	VAE†+ConvNeXt	Unified	50,000	75.30	55.41	44.44	17.53	26.32
D	✓	SigLIP	-	Und. Only	13,000	85.03	69.10	54.23	-	-
E	✓	-	VAE†+ConvNeXt	Gen. Only	37,000	-	-	-	16.69	26.89
F	✓	SigLIP	VAE†+ConvNeXt	Unified	50,000	84.73	69.20	54.83	17.61	26.40

Impact of Representation Alignment. The comparison between Exp. A and F demonstrates the significant benefits of incorporating representation alignment regularization [101] during training. Specifically, models trained with representation alignment show notably lower MJHQ FID and higher CLIP scores, indicating simultaneous improvements in both image quality and semantic alignment. Importantly, our architecture differs from previous studies [65, 71] examined in [101] due to our incorporation of LLM and an additional skip connection between g_{enc} and g_{dec}. The effectiveness of representation alignment in our modified architecture suggests its broad generalization capability across different network structures.

Impact of Decoupling Visual Encoders. e efficacy of using powerful pre-trained visual encoders in multimodal understanding. The comparison among Exp. B, C, and F demonstrates the advantages of using separate visual encoders for understanding and generation tasks. In Exp. B, following a design similar to Transfusion [106], we implement shared ConvNeXt blocks in the SDXL-VAE latent space for both understanding and generation encoders. Exp. C employs separate encoders with identical architectures and initialization parameters, but trained independently. The performance differences between these configurations validate the necessity of decoupled visual encoders in improving our unified model's capabilities. Moreover, the superior results in Exp. C and F highlight the benefits of leveraging pre-trained semantic visual encoders for multimodal understanding tasks.

Fair Comparison with Understanding / Generation-Only Models. To establish meaningful benchmarks, we evaluate task-specific models trained under identical conditions - using the same pre-training data and infrastructure. Exp. D and E represent the specialized models, trained with data volumes matching the unified model in Tab. 5. The minimal performance gap between Exp. F and these task-

A corgi's head depicted as an explosion of a nebula, with vibrant cosmic colors like deep purples, blues, and pinks swirling around. The corgi's fur blends seamlessly into the nebula, with stars and galaxies forming the texture of its fur. Bright bursts of light emanate from its eyes, and faint constellations can be seen in the background, giving the image a surreal, otherworldly feel.

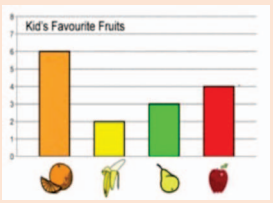
Beautiful surreal symbolism the mesmerizing vision of a Cleopatra Queen of Egypt, mesmerizing brown eyes, black hair and ethereal features, radiating celestial aura, super high definition, true lifelike color, perfect exposure, razor sharp focus, golden ratio, soft reflections, bokeh effect, fine art photography, cinematic compositing, authentic, professional.

A lone figure in dark robes ascends worn stone steps toward a glowing light in an ancient temple entrance. Ornate arches, lush greenery, and intricate carvings adorn the scene, evoking a mystical, high-fantasy atmosphere reminiscent of works by artists like Randy Vargas, with cinematic lighting and epic storytelling.

Figure 4. **Image generation results of JanusFlow,** which can generate high-quality images that are semantically aligned to text prompts.

User: What are the kinds of fruits in this picture?

JaunsFlow (Ours): The fruits in the picture are banana, strawberry, mango, persimmon, blueberry, and lime.



User: Describe this image.

JaunsFlow (Ours): Kid's Favourite Fruits is a bar graph. The x-axis shows the fruits. The y-axis shows the number of kids who like the fruit. The orange is the most popular fruit. The banana is the least popular fruit.

User: How many penguins in the picture?

JaunsFlow (Ours): There are 11 penguins in the picture.

Figure 5. **Visual Understanding with JanusFlow.** Our model effectively handles various visual understanding tasks, such as question answering, plot interpretation, and object counting.

specific baselines demonstrates that our framework successfully integrates understanding and generation without significant compromise in either task's performance.

4.6. Qualitative Results

We present qualitative evaluations of our method for both image generation and understanding tasks. Fig. 1b and Fig. 4 showcases the image generation capabilities of JanusFlow. These results demonstrate both the high visual quality of our generated images and our framework's ability to faithfully follow diverse instructions. For multimodal understanding, Fig. 5 presents example conversations that show our model's understanding capabilities across various scenarios. These interactions demonstrate the model's ability to understand and reason about visual content in natural language dialogues. Additional qualitative examples of JanusFlow are provided in the appendix.

5. Conclusion

We present JanusFlow, a unified framework that successfully harmonizes autoregressive and rectified flow models for multimodal understanding and generation tasks. Our extensive experiments demonstrate that this unification achieves comparable performance to task-specific models. The successful integration of these fundamentally different model architectures not only addresses current challenges in multimodal learning but also opens new possibilities for future research in training unified models.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 1, 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2022. 1, 2

[3] Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In Proc. Int'l Conf. Learning Representations, 2023. 1, 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023. 1, 2, 7

[5] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2023. 3

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. Computer Science, 2023. 6

[7] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. DeepSeek LLM: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954, 2024. 1, 2, 5, 6

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712, 2023. 1, 2

[9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426, 2023. 6

[10] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Sigma: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. arXiv preprint arXiv:2403.04692, 2024. 6

[11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023. 5

[12] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. MobileVLM: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886, 2023. 7

[13] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. MobileVLM V2: Faster and stronger baseline for vision language model. arXiv preprint arXiv:2402.03766, 2024. 7

[14] Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In Proc. Int'l Conf. Machine Learning, 2024. 3

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2023. 1, 2, 7

[16] dclure. LAION-Aesthetics-UMAP, 2022. 5, 6

[17] DeepFloyd. DeepFloyd IF, 2023. 6

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2021. 4

[19] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. DreamLLM: Synergistic multimodal comprehension and creation. In Proc. Int'l Conf. Learning Representations, 2024. 1, 2, 7

[20] echo840. Detailed caption, 2023. 5

[21] Ben Egan, Alex Redden, XWAVE, and SilentAntagonist. DALLE-3 1 million+ high quality captions, 2024. 5

[22] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2021. 1, 2

[23] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Proc. Int'l Conf. Machine Learning, 2024. 1, 2, 3, 4

[24] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2024. 6

[25] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a SEED of vision in large language model. arXiv preprint arXiv:2307.08041, 2023. 1, 2

[26] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making LLaMA SEE and draw with SEED tokenizer. arXiv preprint arXiv:2310.01218, 2023.

[27] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396, 2024. 1, 2, 6

[28] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 1, 6

[29] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2017. 6

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. Proc. Annu. Conf. Neural Inf. Process. Systems, 2017. 6

[31] High-flyer. HAI-LLM: Efficient and lightweight training tool for large models, 2023. 5

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2020. 1, 2

[33] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. ScreenQA: Large-scale question-answer pairs over mobile app screenshots. arXiv preprint arXiv:2209.08199, 2022. 6

[34] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135, 2024. 1, 6

[35] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2019. 1, 6

[36] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954, 2024. 1, 2

[37] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, CHEN Bin, Chengru Song, Di ZHANG, Wenwu Ou, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In Proc. Int'l Conf. Learning Representations, 2024. 7

[38] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2023. 6

[39] Sungwon Kim, Kevin Shih, Joao Felipe Santos, Evelina Bakhturina, Mikyas Desta, Rafael Valle, Sungroh Yoon, Bryan Catanzaro, et al. P-Flow: a fast and data-efficient zero-shot tts through speech prompting. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 3

[40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proc. IEEE Int. Conf. Comput. Vision, 2023. 5

[41] Mahnaz Koupaee and William Yang Wang. WikiHow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305, 2018. 5

[42] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. Int'l Journal of Computer Vision, 2020. 5

[43] H Laurençon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, et al. Introducing IDEFICS: An open reproduction of state-of-the-art visual language model, 2023, 2023. 7

[44] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. VoiceBox: Text-guided multilingual universal speech generation at scale. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 1

[45] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023. 1, 6

[46] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024. 1, 2, 6

[47] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024. 1, 6

[48] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proc. Int'l Conf. Machine Learning, 2023. 1, 2

[49] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arXiv: A dataset for improving scientific comprehension of large vision-language models. In Annual Meeting of the Association for Computational Linguistics, 2024. 5

[50] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. DenseFusion-1M: Merging vision experts for comprehensive multimodal perception. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 5

[51] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Proc. Conf. on Empirical Methods in Natural Language Process., 2023. 6

[52] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. MMSci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In AI for Accelerated Materials Design, 2024. 5

[53] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer

with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748, 2024. 6

[54] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In Proc. Int'l Conf. Learning Representations, 2023. 1, 2, 3

[55] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2024. 1, 2, 7

[56] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 7

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 1, 2, 4, 7

[58] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. arXiv preprint arXiv:2402.08268, 2024. 6, 7

[59] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. arXiv preprint arXiv:2209.14577, 2022. 2

[60] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In Proc. Int'l Conf. Learning Representations, 2023. 1, 2, 3

[61] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In Proc. Int'l Conf. Learning Representations, 2024. 1, 2, 3

[62] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In Proc. European Conf. Computer Vision, 2024. 1, 6

[63] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. DeepSeek-VL: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525, 2024. 2, 4, 5

[64] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2021. 6

[65] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. arXiv preprint arXiv:2401.08740, 2024. 7

[66] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. arXiv preprint arXiv:2303.09319, 2023. 2

[67] madebyollin. Megalith-10M, 2024. 5

[68] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020. 1, 2

[69] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Annual Meeting of the Association for Computational Linguistics, 2022. 6

[70] mehdidc. YFCC-15M, 2024. 5

[71] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proc. IEEE Int. Conf. Comput. Vision, 2023. 4, 7

[72] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In Proc. Int'l Conf. Learning Representations, 2024. 1, 2, 3, 5, 6

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In Proc. Int'l Conf. Machine Learning, 2021. 2

[74] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022. 1, 2, 6

[75] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2022. 1, 3, 6

[76] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2022.

[77] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2022. 1, 2

[78] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: Knowledge-aware visual question answering. In Proc. AAAI Conf. on Artificial Intelligence, 2019. 6

[79] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2019. 6

[80] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. arXiv preprint arXiv:2406.10328, 2024. 5, 6

[81] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In Proc. Int'l Conf. Learning Representations, 2021. 1, 2

[82] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proc. ACM SIGIR Conf. Research and Develop. in Info. Retrieval, 2021. 5

[83] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. JourneyDB: A benchmark for generative image understanding. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 5, 6

[84] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: LLaMA for scalable image generation. arXiv preprint arXiv:2406.06525, 2024. 1, 2, 6

[85] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. In Proc. Int'l Conf. Learning Representations, 2024. 1, 2, 7

[86] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024. 1, 2, 3, 6, 7

[87] Gemini Team. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023. 7

[88] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. arXiv preprint arXiv:2406.16860, 2024. 4

[89] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 1, 2

[90] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. 2

[91] Cristina Nader Vasconcelos, Abdullah Rashwan, Austin Waters, Trevor Walker, Keyang Xu, Jimmy Yan, Rui Qian, Yeqing Li, SHIXIN LUO, Yasumasa Onoe, et al. Greedy growing enables high-resolution pixel-based diffusion models. Transactions on Machine Learning Research, 2024. 3

[92] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024. 7

[93] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869, 2024. 6, 7

[94] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2023. 3

[95] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848, 2024. 1, 2, 3, 4, 6, 7

[96] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In Proc. Int'l Conf. Machine Learning, 2024. 7

[97] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. VILA-U: A unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429, 2024. 1, 2, 6, 7

[98] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528, 2024. 1, 2, 3, 6, 7

[99] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-VILA: Cross-modality alignment for large language model. arXiv preprint arXiv:2405.19335, 2024. 2

[100] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In Proc. Int'l Conf. Learning Representations, 2024. 3

[101] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940, 2024. 4, 7

[102] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In Proc. Int'l Conf. Machine Learning, 2024. 6

[103] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2024. 6

[104] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proc. IEEE Int. Conf. Comput. Vision, 2023. 3, 5

[105] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong

Wang. MonoFormer: One transformer for both diffusion and autoregression. arXiv preprint arXiv:2409.16280, 2024. 2

[106] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039, 2024. 1, 2, 3, 6, 7

[107] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. LLaVA-Phi: Efficient multi-modal assistant with small language model. arXiv preprint arXiv:2401.02330, 2024. 7

[108] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. arXiv preprint arXiv:2406.18583, 2024. 6