Emu3: Next-Token Prediction is All You Need

Emu3 Team[*] **BAAI**
`https://emu.baai.ac.cn`

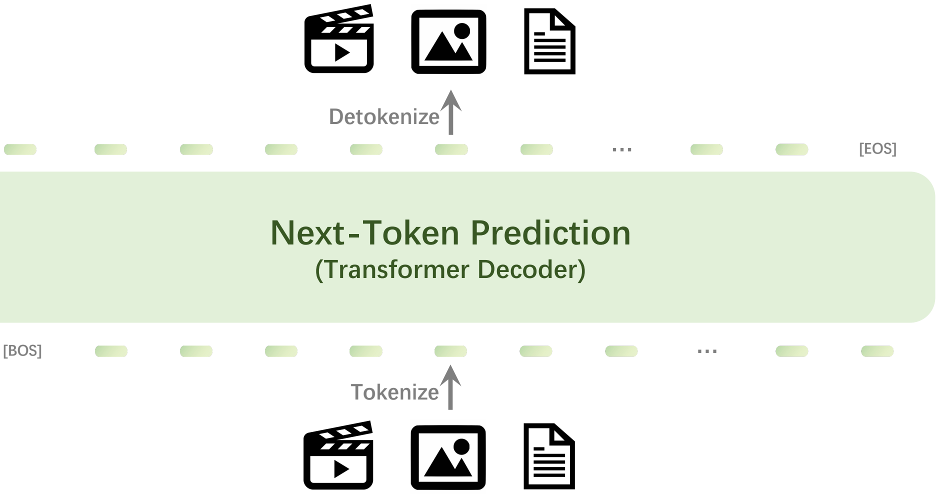


Figure 1: **Emu3** is trained to predict the next token with a single Transformer on a mix of video, image, and text tokens. **Emu3** achieves state-of-the-art performance compared to well-established task-specific models in generation and perception tasks.

Abstract

While next-token prediction is considered a promising path towards artificial general intelligence, it has struggled to excel in multimodal tasks, which are still dominated by diffusion models (*e.g.*, Stable Diffusion) and compositional approaches (*e.g.*, CLIP combined with LLMs). In this paper, we introduce **Emu3**, a new suite of state-of-the-art multimodal models trained solely with next-token prediction. By tokenizing images, text, and videos into a discrete space, we train a single transformer from scratch on a mixture of multimodal sequences. **Emu3** outperforms several well-established task-specific models in both generation and perception tasks, surpassing flagship models such as SDXL and LLaVA-1.6, while eliminating the need for diffusion or compositional architectures. **Emu3** is also capable of generating high-fidelity video via predicting the next token in a video sequence. We simplify complex multimodal model designs by converging on a singular focus: tokens, unlocking great potential for scaling both during training and inference. Our results demonstrate that next-token prediction is a promising path towards building general multimodal intelligence beyond language. We open-source key techniques and models to support further research in this direction.

---

[*] See Contributions section for full author list. September 27, 2024

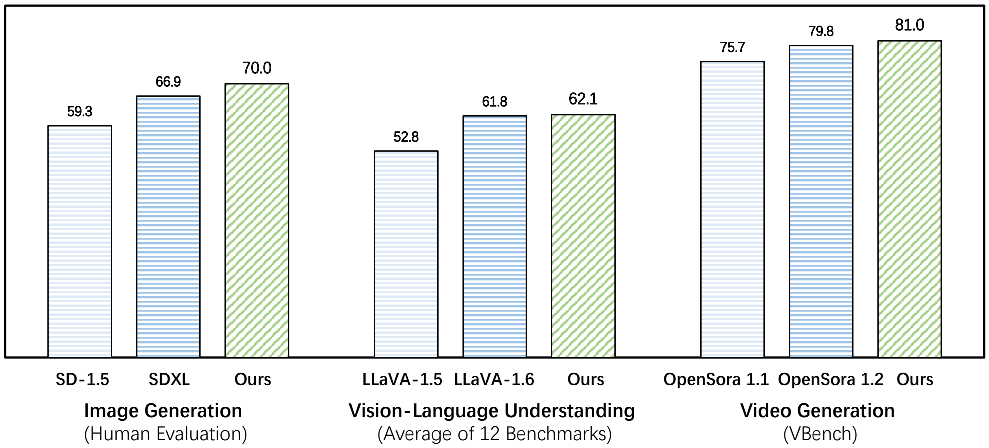


Figure 2: Comparison with open-source flagship models in vision generation and perception. Based solely on next-token prediction, **Emu3** beats SDXL [66], LLaVA-1.6-7B [56], OpenSora-1.2 [107] respectively, dispensing with diffusion and CLIP entirely. For the image generation task, we present comparison results of human evaluation scores based on English prompts. For the vision-language understanding task, we assess the average scores across twelve benchmarks: SEEDBench-Img [45], OCRBench [59](with normalized results), MMVet [98], POPE [51], VQAv2 [27], GQA [34], TextVQA [78], ChartQA [61], AI2D [36], RealWorldQA [91], MMMU [99], and MMbench [58]. For the video generation task, we present comparison results of VBench.

1 Introduction

Next-token prediction has revolutionized the field of language models [86, 69, 9], enabling breakthroughs like ChatGPT [64] and sparking discussions about the early signs of artificial general intelligence (AGI) [10]. However, the applicability of this paradigm to multimodal models remains unclear, with limited evidence of its efficacy in achieving competitive performance across different tasks.

In the realm of multimodal models, vision generation has been dominated by complex diffusion models (*e.g.*, Stable Diffusion [73]), while vision-language perception has been led by compositional approaches such as CLIP [67] with LLMs (*e.g.*, LLaVA [57]). Despite early attempts at unifying generation and perception, such as Emu [82] and Chameleon [83], these efforts either resort to connecting LLMs with diffusion models or fail to match the performance of task-specific methods tailored for generation and perception.

In this work, we present **Emu3**, a new set of state-of-the-art multimodal models based solely on next-token prediction, eliminating the need for diffusion or compositional approaches entirely. We tokenize images, text, and videos into a discrete space, and jointly train a single transformer from scratch on a mix of multimodal sequences.

Emu3 achieves state-of-the-art performance compared to well-established task-specific models in generation and perception tasks. **Emu3** outperforms the flagship Stable Diffusion model, *i.e.*, SDXL [66], in both the human evaluation and the public text-to-image benchmarks such as MSCOCO-30K [15], GenEval [26], T2I-CompBench [32], and DPG-Bench [31]. For vision-language understanding, **Emu3** competes with the popular vision-language model, *i.e.*, LLaVA-1.6 [56], on a series of public vision-language benchmarks, including SEED-Bench [45], RealWorldQA [91], OCRBench [59], *etc*.

Emu3 is capable of generating videos. Unlike Sora [8] that employs the video diffusion model to generate a video from noise, **Emu3** simply generates a video causally by predicting the next token in a video sequence. The model can simulate some aspects of environments, people and animals in the physical world. With a video in context, **Emu3** extends the video and predicts what will happen next. Given the user's prompt, the model can generate high-fidelity videos following the text description. **Emu3** stands out and competes with other video diffusion models on the VBench benchmark [33] for text-to-video generation.

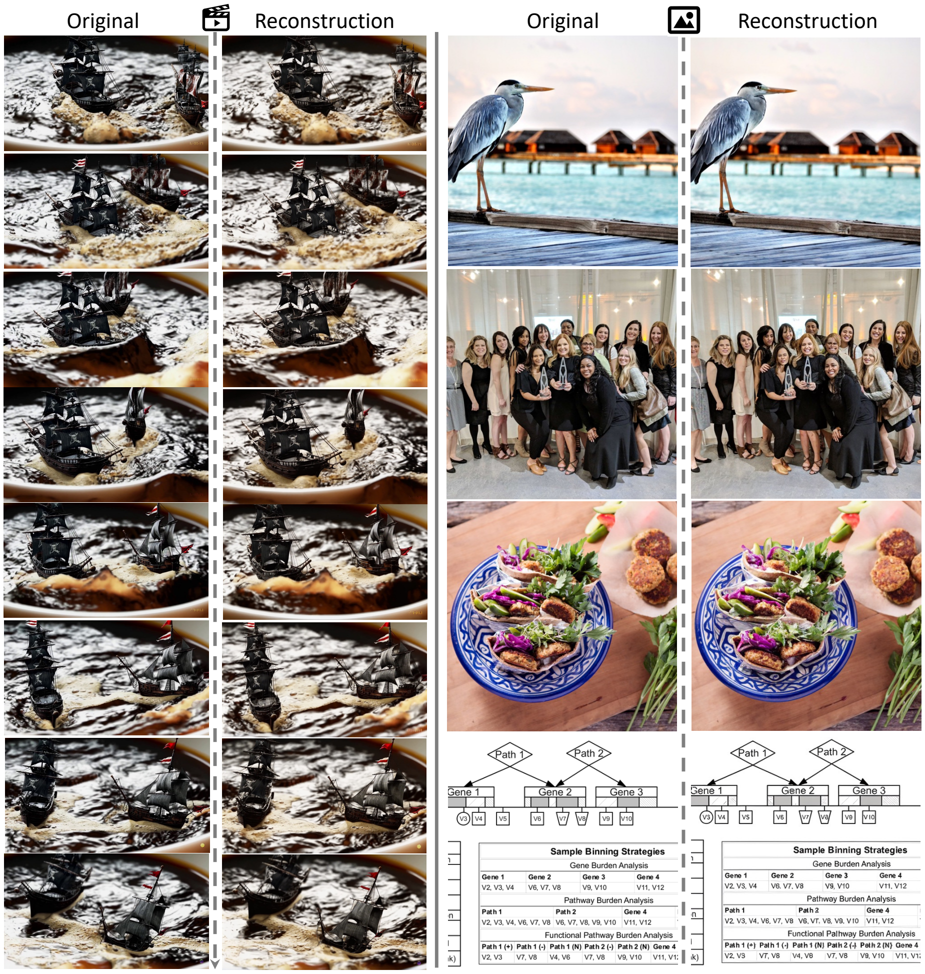


Figure 3: Reconstruction samples. Left: Original and reconstructed videos at 540 × 960 resolution, showcasing a sampling of 8 frames at 30 FPS. Right: original and reconstructed 512 × 512 resolution images. Zoom in to see the details.

We open-source key techniques and models to facilitate future research in this direction. Notably, we provide a robust vision tokenizer, enabling the transformation of videos and images into discrete tokens, which was previously publicly unavailable. We also demonstrate the versatility of the next-token prediction framework, showing that direct preference optimization (DPO) [68] can be seamlessly applied to autoregressive vision generation, aligning the model with human preferences.

Our results provide strong evidence that next-token prediction can serve as a powerful paradigm for multimodal models, scaling beyond language models and delivering state-of-the-art performance across multimodal tasks. By simplifying complex model designs and focusing solely on *tokens*, it unlocks significant potential for *scaling* both during training and inference. We believe that next-token prediction offers a promising path towards building general multimodal intelligence.

2 Approach

2.1 Data

Emu3 is trained from scratch on a mix of language, image, and video data.

Language Data. We use the same language data as in Aquila [101], which is a high-quality corpus consisting of both Chinese and English data.

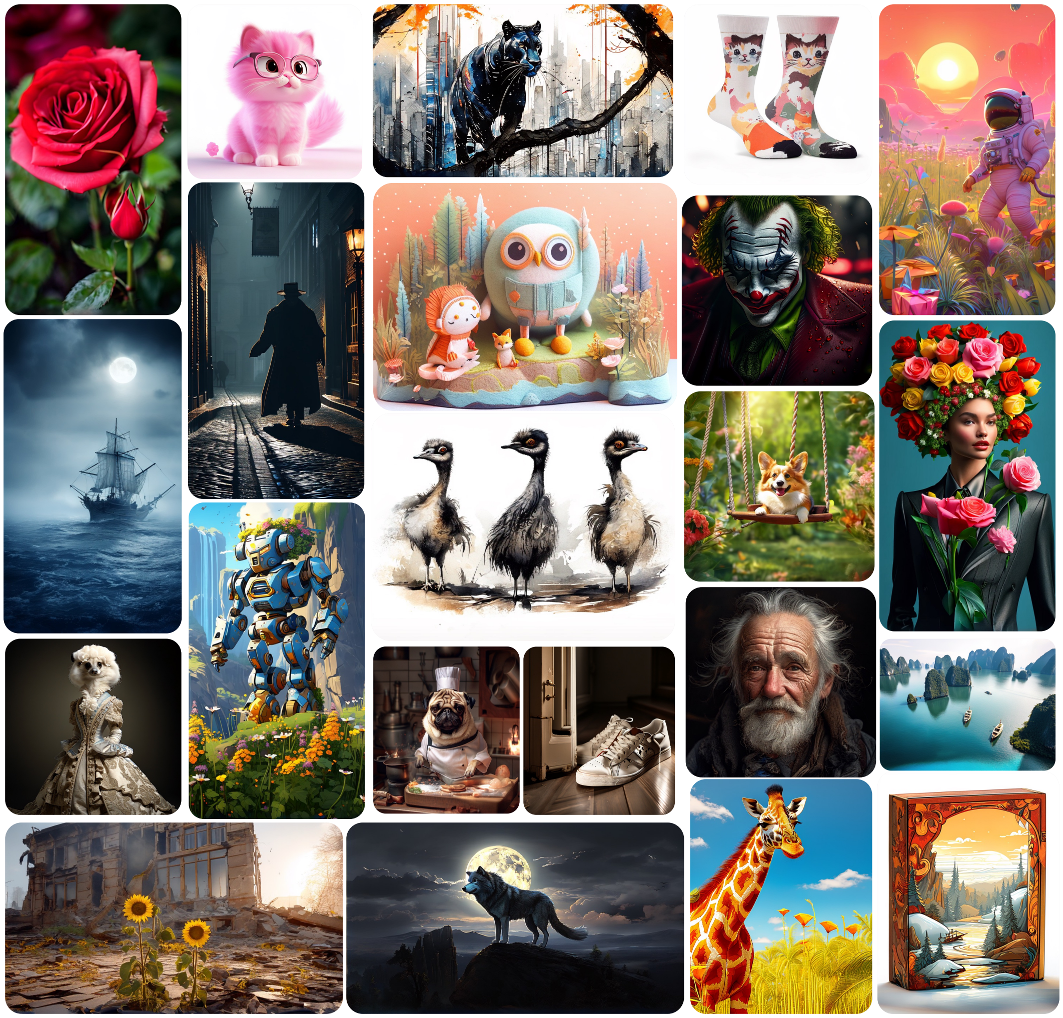


Figure 4: Qualitative results of **Emu3** text-to-image generation.

Image Data. We curate a large-scale image-text dataset comprising open-source web data, AI-generated data, and high-quality in-house data. The filtering process involves several key steps: **1**) We apply a resolution filter, discarding samples with a resolution below 512 × 512 pixels. **2**) We assess the aesthetic quality of each image using the LAION-AI aesthetic predictor[2], excluding images with scores below 5.5 to ensure the overall aesthetic quality. **3**) For images that did not pass the aesthetic filter, we employ text detection[3] and color filtering to retain non-monochromatic images and those with minimal text, improving the filtering recall of open-world images. **4**) Additionally, we prepare supplementary data for image understanding. By following the data processing pipeline in DenseFusion [50], we extract millions of representative images that encompass a wide range of categories, including charts, table, text-rich content, and more, sourced from diverse open-source web data.

To annotate the filtered dataset, we develop an image captioning model based on Emu2 [81] to construct dense synthetic captions. We leverage GPT-4V [64] with detailed prompts to generate approximately 1 million image-caption pairs. This annotated dataset is then used to fine-tune the Emu2-17B [81] model as our image captioner. Additionally, we utilize the open-source vLLM library [40] to accelerate the labeling process.

---

[2] `https://github.com/LAION-AI/aesthetic-predictor`

[3] `https://github.com/PaddlePaddle/PaddleOCR`

Configurations	VisionTokenizer
Pretrained Weights	SBER-MoVQGAN-270M[5]
Codebook Size	32768
Latent Size	4
Compression	4 × 8 × 8

Table 1: **Emu3** vision tokenizer configurations.

Video Resolution	LPIPS↓	PSNR↑	SSIM↑
128 × 128	0.099	21.71	0.630
256 × 256	0.109	21.59	0.622
512 × 512	0.112	22.69	0.690
720 × 720	0.110	24.30	0.771

Table 2: Video compression metrics.

Video Data. We collect videos covering a wide range of categories, such as landscapes, animals, plants, games, and actions. These videos are prepossessed with a sophisticated pipeline [6] with the following four stages: **1**) We split the videos to scenes with PySceneDectect[4], employing both ContentDetector and ThresholdDetector to identify content changes and fade-in/out events, respectively. **2**) Text detection are performed using PaddleOCR[3] and clips with excessive text coverage were removed. To reduce computational costs, we sample video frames at 2 FPS and resize the shorter edge to 256. **3**) We further calculate the optical flow [84] to eliminate clips with minimal or extreme motion. As with the previous step, we sample and resize video frames for efficiency. The flow score is defined as the ratio between the average flow magnitude of all pixels and the shorter edge. We exclude clips with flow scores outside the acceptable range. **4**) Finally, we assess the aesthetic quality of each clip using the LAION-AI aesthetic predictor[1]. We sample three frames and get three scores for each clip, and clips whose lowest score is smaller than 5 are discarded.

We caption the filtered video clips using a video captioner trained based on our image captioner. The training data is initially labeled by GPT-4V [64]. For each video clip, we sample eight frames and create a detailed prompt for GPT-4V to describe both the content and motion within these frames. Some of the labeled data undergoes manual revision. We then fine-tune our image captioner on this labeled data to develop our video captioner. For large-scale deployment, we accelerate captioning with vLLM [40]. Clips shorter than 20 seconds are captioned using 12 evenly sampled frames, while longer clips are split into 10-20 second sub-clips, each captioned independently.

2.2 Vision Tokenizer

We train the vision tokenizer based on SBER-MoVQGAN[5], which can encode a 4 × 512 × 512 video clip or a 512 × 512 image into 4096 discrete tokens from a codebook of size 32,768. Our tokenizer achieves 4× compression in the temporal dimension and 8×8 compression in the spatial dimension, applicable to any temporal and spatial resolution. Building on the MoVQGAN architecture [106], we incorporate two temporal residual layers with 3D convolution kernels into both the encoder and decoder modules to enhance video tokenization capabilities. The tokenizer is trained end-to-end on the LAION-High-Resolution[6] image dataset and the InternVid [89] video dataset using combined objective functions of L2 loss, LPIPS perceptual loss [104], GAN loss, and commitment loss [23].

Qualitative results are presented in Fig. 3. We report LPIPS (computed by the AlexNet features), PSNR, and SSIM scores in Tab. 2 using an evaluation dataset of 3,172 videos from Pexels[7]. The videos were reconstructed over 5 seconds while maintaining the aspect ratio. During evaluation, original and reconstructed videos were resized and cropped based on the shorter side and uniformly sampled with 8 frames at 12 FPS.

2.3 Architecture

The **Emu3** model retains the architectural framework of established large language models (LLMs) such as Llama-2 [85], with the primary modification being the expansion of the embedding layer to accommodate discrete vision tokens. We use RMSNorm [100] for normalization and GQA [1] for attention mechanisms, while employing the SwiGLU [76] activation function and rotary positional embeddings (RoPE) [79]. Biases

Configurations	**Emu3**
Parameters	8B
Layers	32
Hidden Size	4096
Intermediate Size	14336
Heads	32
KV Heads	8
Vocabulary Size	184622
RoPE Base	1000000
Context Length	131072

Table 3: Model configurations.

---

[4] `https://github.com/Breakthrough/PySceneDetect`

[5] `https://github.com/ai-forever/MoVQGAN`

[6] `https://huggingface.co/datasets/laion/laion-high-resolution`

[7] `https://www.pexels.com/search/videos/videos`

in the qkv and linear projection layers are removed. Additionally, a dropout rate of 0.1 is implemented to improve training stability. We use the QwenTokenizer[8] to tokenize multilingual texts. Detailed configurations are provided in Tab. 3.

2.4 Pre-training

Data Preparation. During pre-training we first define the multimodal data format. Unlike diffusion models that rely on an external text encoder, **Emu3** natively integrates textual conditional information for image/video generation. We rescale images/videos while preserving their aspect ratio to a size with an area close to 512×512, and then generate vision tokens using our vision tokenizer. Then, we incorporate five special tokens to merge text and vision data, creating document-like inputs for the training process. The resulting training data is structured as follows:

[BOS] {caption text} [SOV] {meta text} [SOT] {vision tokens} [EOV] [EOS].

Where [BOS] and [EOS] are the original special tokens in the text tokenizer, [SOV] marking the start of the vision input, [SOT] marking the start of vision tokens, and [EOV] indicating the end of the vision input. Additionally, [EOL] and [EOF] are inserted into the vision tokens to denote line breaks and frame breaks, respectively. The "meta text" contains information about the resolution for images, and for videos, it includes resolution, frame rate, and duration, all presented in plain text format. We also move the "caption text" field in a portion of the dataset to follow the [EOV] token, thereby constructing data aimed at vision understanding tasks.

Training Objective. Since vision signals in **Emu3** are fully converted into discrete tokens, we only need to train with the next-token prediction task using the standard cross-entropy loss. To prevent vision tokens from dominating the learning process, we apply a weight of 0.5 to the loss associated with vision tokens.

Training Details. The **Emu3** model utilizes an extensive context length during pre-training to handle video data. To facilitate training, we employ a combination of tensor parallelism (TP), context parallelism (CP), and data parallelism (DP). We simultaneously pack text-image data into the maximum context length to fully utilize computational resources, while ensuring that complete images are not segmented during the packing process. The pre-training process is conducted in two stages. In the first stage, which does not utilize video data, training begins from scratch with a context length of 5120 for text and image data. In the second stage, video data is introduced, and a context length of 131072 is employed. Both stages use a learning rate of 5×10^{-5}, with a cosine annealing of the learning rate to zero.

2.5 Post-training

2.5.1 Vision Generation

Quality Fine-Tuning. Following the pre-training phase, we conduct post-training for vision generation tasks to enhance the quality of generated outputs. We apply quality fine-tuning (QFT) using high-quality data. The model continues training with the next token prediction task using standard cross-entropy loss; however, supervision is applied exclusively to the vision tokens. For the image data in QFT, we select diverse high-quality sources and filtered them based on the average of three popular preference scores: HPSv2.1 [90], MPS [105], and the LAION Aesthetics score [43]. During QFT, we increase the training data resolution from 512 pixels to 720 pixels to improve generation quality. For the video data, we sample from high-quality sources and apply stringent resolution and optical flow filters to ensure quality. Additionally, at the end of training, we use an annealing strategy to linearly decay the learning rate to zero.

Direct Preference Optimization. Direct Preference Optimization (DPO) [68], an effective approach for better aligning models with human preferences. We adopt DPO techniques for autoregressive multimodal generation tasks, leveraging human preference data to enhance model performance. We divide the dataset construction into three steps: **1**) We perform 8-10 inferences for each user-collected prompt (p) using the quality fine-tuned model, creating an initial data pool (x). **2**) Each prompt

---

[8] `https://huggingface.co/Qwen/Qwen-7B/blob/main/tokenization_qwen.py`

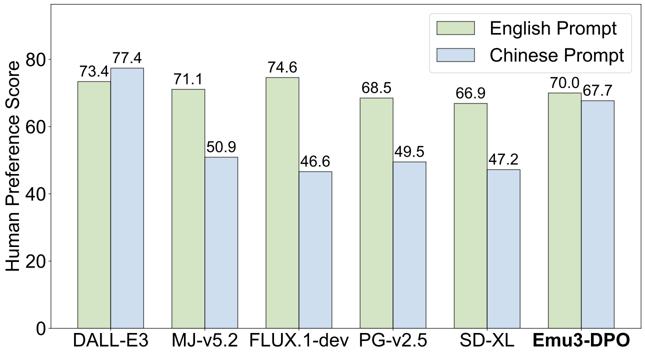


Figure 5: Human evaluation overall score comparison of closed and open generative image models under English and Chinese prompts.

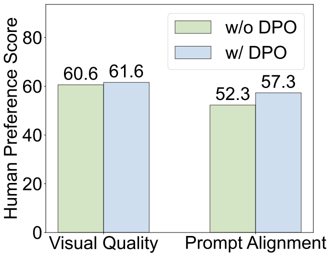


Figure 6: DPO improves visual quality and prompt alignment.

is evaluated by three voters, focusing on vision appeal and prompt alignment. **3**) Based on the scores, the highest scoring sample is chosen, and the lowest is rejected to form a triplet (p_i, x_i^{chosen}, $x_i^{rejected}$) with the prompt for further training. Specifically, the tokens from the data construction process are stored for direct use in future training phases. This strategy eliminates reconstruction differences caused by re-tokenization. **Emu3-DPO** minimizes the DPO loss and the next token prediction cross-entropy loss to fine-tune the QFT model.

2.5.2 Vision-Language Understanding

The pretrained model undergoes a two-stage post-training process for vision-language understanding: 1) image-to-text training, and 2) instruction tuning. During the first stage, our approach integrates image understanding data with pure-language data, while losses associated with vision tokens are disregarded for text-only prediction. Each image is resized to a resolution of about 512 × 512 while preserving the original aspect ratio. In the second stage, we sample a subset of question-answer pairs from [44] to enhance the vision instruction following ability. Images below 512 × 512 or above 1024 × 1024 will be resized to the lower or upper resolution limit while keeping the aspect ratio accordingly, while others maintain their original resolution.

3 Main Results

3.1 Image Generation

3.1.1 Automated Metric Evaluation

We present the performance of **Emu3** through automated metric evaluation on popular text-to-image benchmarks: MSCOCO-30K [15], GenEval [26], T2I-CompBench [32], and DPG-Bench [31]. The comparison results of **Emu3** against diffusion methods, autoregressive diffusion methods, and autoregressive-based methods across these four benchmarks are shown in Tab. 4. Our method outperforms autoregressive diffusion methods in image-text alignment evaluation and is comparable to state-of-the-art diffusion-based models, despite not utilizing any pre-trained language models.

We report the results of GenEval and T2I-CompBench after employing a rewriter to expand short prompts. Due to **Emu3** utilizing a significant proportion of synthetic labels during training, it exhibits superior performance in dense captioning compared to shorter prompts. However, the evaluation prompts in GenEval and T2I-CompBench are too brief to accurately reflect the model's true performance. Following DALL-E 3, we also report our evaluation results using GPT-4V as the rewriter. The GenEval overall score results indicate that **Emu3** significantly outperforms Chameleon, a multi-modal autoregressive model, as well as the latest autoregressive diffusion methods, Show-O and Transfusion. Additionally, **Emu3** surpasses SDXL and matches the performance of state-of-the-art diffusion models, including DALL-E 3. Detailed comparisons across all dimensions, including results from the original prompts, are provided in Appendix B.1.

		MSCOCO			GenEval	T2I-CompBench			DPG-Bench
Method	Text Pretrain	CLIP-I	CLIP-T	FID	Overall	Color	Shape	Texture	Average
Diffusion-based									
SDv1.5 [73]	CLIP ViT-L/14	0.667	0.302	9.93	0.43	0.3730	0.3646	0.4219	63.18
DALL-E 2 [70]	CLIP ViT-H/16	-	0.314	10.93	0.52	0.5750	0.5464	0.6374	-
SDv2.1 [73]	CLIP ViT-H/14	-	-	-	0.50	0.5694	0.4495	0.4982	-
SDXL [66]	CLIP ViT-bigG	0.674	0.310	-	0.55	0.6369	0.5408	0.5637	74.65
PixArt-alpha [13]	Flan-T5-XXL	-	-	7.32	0.48	0.6886	0.5582	0.7044	71.11
DALL-E 3 [5]	Flan-T5-XXL	-	0.320	-	0.67†	0.8110	0.6750	0.8070	83.50
SD3 [22]	Flan-T5-XXL	-	-	-	0.74	-	-	-	-
Autoregressive meets diffusion									
Emu [82]	LLaMA-7B	0.656	0.286	11.6	-	-	-	-	-
Show-o [92]	Phi-1.5	-	-	9.24	0.53	-	-	-	-
Transfusion [108]	-	-	-	6.78	0.63	-	-	-	-
Autoregressive-based									
Chameleon [83]	-	-	-	26.74	0.39	-	-	-	-
LlamaGen [80]	FLAN-T5 XL	-	-	-	0.32	-	-	-	-
Emu3	-	0.689	0.313	12.8	0.66†	0.7913†	0.5846†	0.7422†	80.60
Emu3-DPO	-	0.680	0.312	19.3	0.64†	0.7544†	0.5706†	0.7164†	81.60

Table 4: **Comparison with state-of-the-art models on text-to-image benchmarks.** We evaluate on MSCOCO-30K [15]; GenEval [26]; T2I-CompBench [32] and DPG-Bench [31]. † result is with rewriting.

To further assess state-of-the-art text-to-image methods, particularly diffusion models, we evaluate the alignment between generated images and text conditions using T2I-CompBench. **Emu3** demonstrates competitive performance compared to SoTA diffusion-based models. Additionally, we compare our models with state-of-the-art (SoTA) models on the DPG-Bench, which features longer prompts with more detailed information for evaluation. Our **Emu3-DPO** achieves an overall score of 81.6, surpassing SDXL and PixArt-alpha, and is comparable to DALL-E 3, providing further evidence of the model's ability to follow long prompts. When comparing **Emu3** with **Emu3-DPO**, we observe a slight decline in the evaluation results after applying DPO, which may be attributed to preferences in our DPO datasets that emphasize overall aesthetic quality–a focus that differs from the domains of the automated evaluation models, complicating conclusions drawn solely through automated evaluation. We therefore introduced human evaluation in Sec.3.1.2.

3.1.2 Human Evaluation

We conduct a human evaluation comparing the text-to-image generation capabilities of different models. A set of 100 diverse user prompts is collected, and each is evaluated by three independent voters. The evaluation focuses on two main aspects: visual quality and prompt following, with a weighted score reflecting the overall performance. As shown in Fig.5, we present a comparison of human preferences for current closed and open generative image models. The results indicate that **Emu3** outperforms SDXL and is on par with DALL-E 3 and MJ-v5.2 in terms of overall score. Furthermore, Fig. 6 demonstrates the impact of alignment through DPO fine-tuning, which effectively improves visual quality and prompt following.

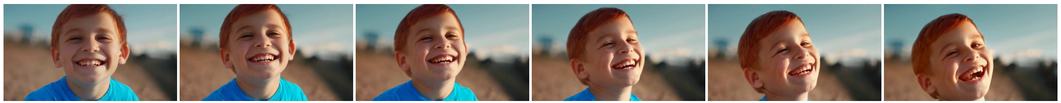
3.1.3 Qualitative Results

Fig. 4 shows 25 images generated by **Emu3** to showcase its capabilities. **Emu3** supports flexible resolutions, aspect ratios, and is capable of handling various styles.

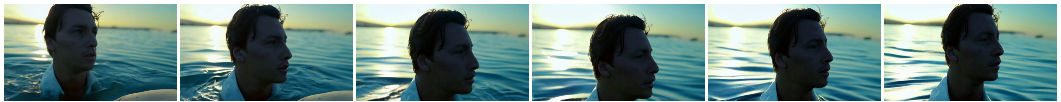
3.2 Video Generation

Consistent with training stage, **Emu3** natively supports the generation of 5-second videos at 24 FPS and can be infinitely extended through an autoregressive approach. Fig. 7 presents qualitative examples of video generation, with 6 frames extracted from the first 3 seconds for showcase.

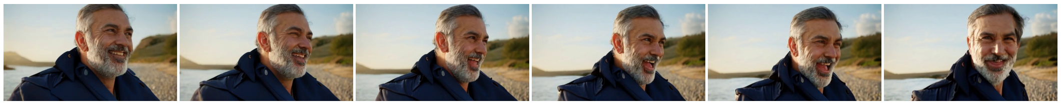
We conducted a quantitative comparison between **Emu3** and the 13 best-performing open-source and proprietary text-to-video models. The used benchmark is VBench [33], a comprehensive toolkit for evaluating video generation performance, which assesses the quality and semantic capabilities of each



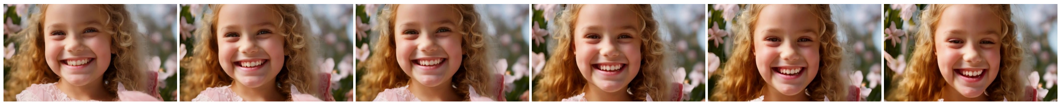
"A red-haired child smiles brightly at the camera, wearing a blue t-shirt. The soft background highlights his joyful expression."



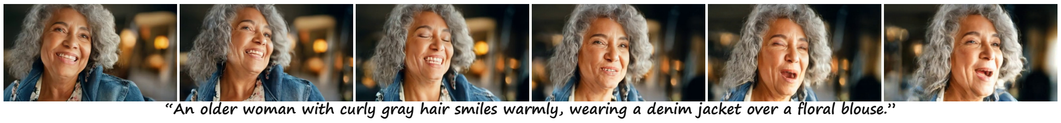
"A man in his late twenties stands in calm water at sunset, wearing a light shirt. Warm colors reflect on the water, with a steady view."



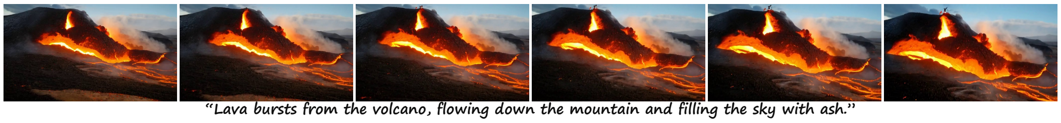
"A middle-aged man with a salt-and-pepper beard smiles as he turns. He wears a navy coat, and sunlight softly highlights his calm face."



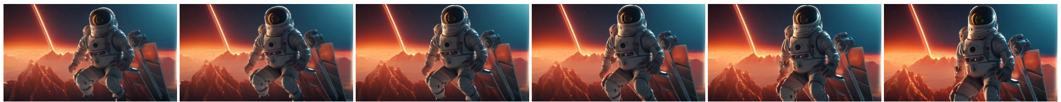
"A curly-haired blonde girl smiles softly, her bright eyes joyful. She wears a pink lace dress, with rosy cheeks against a flowered background."



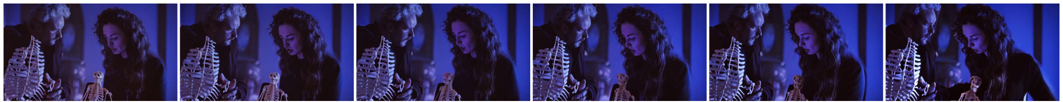
"An older woman with curly gray hair smiles warmly, wearing a denim jacket over a floral blouse."



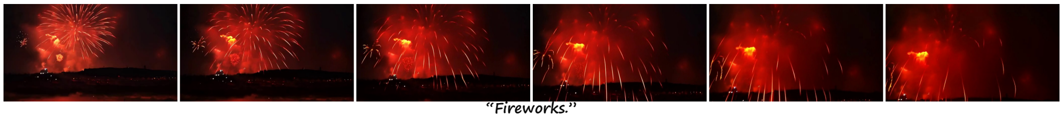
"Lava bursts from the volcano, flowing down the mountain and filling the sky with ash."



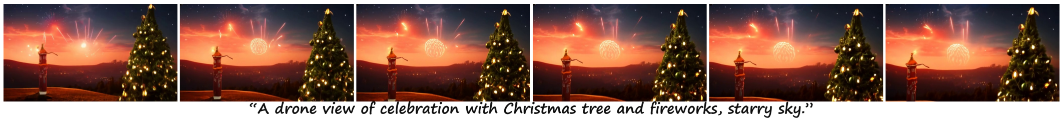
"An astronaut flying in space, in cyberpunk style."



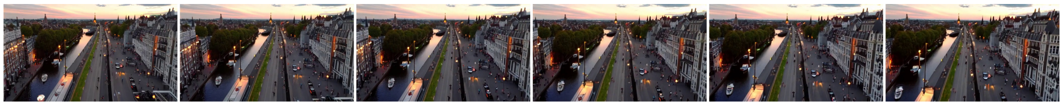
"A man and woman in a blue-lit room examine a small skeleton. He has curly hair, she has long hair. They focus on the skeleton."



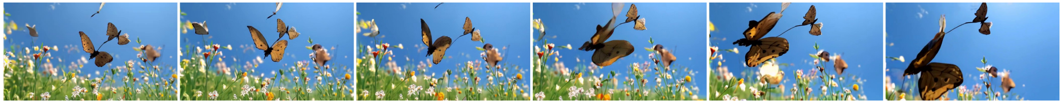
"Fireworks."



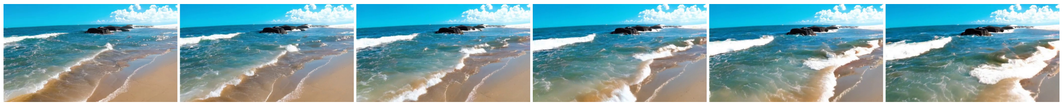
"A drone view of celebration with Christmas tree and fireworks, starry sky."



"An aerial view of a city at dusk, with an orange-pink sky. A canal runs through gabled buildings, boats docked nearby."



"Two butterflies flutter among colorful flowers under a blue sky, their wings shimmering in the sunlight."



"A beautiful coastal beach in spring, waves lapping on sand, with an intense shaking effect."

Figure 7: Qualitative results of **Emu3** text-to-video generation.

Models	Type	Total score	Motion smoothness	Dynamic degree	Aesthetic quality	Object class	Multiple objects	Human action	Spatial relationship	Scene	Appearance style	Subject consistency	Background consistency
ModelScope [87]	Diff	75.75	95.79	66.39	56.39	82.25	38.98	92.4	33.68	39.26	25.67	89.87	95.29
LaVie [88]	Diff	77.08	96.38	49.72	54.94	91.82	33.32	96.8	34.09	52.69	23.56	91.41	97.47
OpenSoraPlan V1.1 [41]	Diff	78.00	98.28	47.72	56.85	76.3	40.35	86.80	53.11	27.17	22.90	95.73	96.73
Show-1 [102]	Diff	78.93	98.24	44.44	57.35	93.07	45.47	95.60	53.50	47.03	23.06	95.53	98.02
OpenSora V1.2 [107]	Diff	79.76	98.50	42.39	56.85	82.22	51.83	91.20	68.56	42.44	23.95	96.75	97.61
AnimateDiff-V2 [28]	Diff	80.27	97.76	40.83	67.16	90.90	36.88	92.60	34.60	50.19	22.42	95.30	97.68
Gen-2 [74]	Diff	80.58	99.58	18.89	66.96	90.92	55.47	89.20	66.91	48.91	19.34	97.61	97.61
Pika [42]	Diff	80.69	99.50	47.50	62.04	88.72	43.08	86.20	61.03	49.83	22.26	96.94	97.36
VideoCrafter-2.0 [11]	Diff	80.44	97.73	42.50	63.13	92.55	40.66	95.00	35.86	55.29	25.13	96.85	98.22
T2V-Turbo (VC2) [47]	Diff	81.01	97.34	49.17	63.04	93.96	54.65	95.20	38.67	55.58	24.42	96.28	97.02
CogVideoX-5B [94]	Diff	81.61	96.92	70.97	61.98	85.23	62.11	99.40	66.35	53.20	24.91	96.23	96.52
Kling (2024-07) [39]	Diff	81.85	99.40	46.94	61.21	87.24	68.05	93.40	73.03	50.86	19.62	98.33	97.60
Gen-3 [75]	Diff	82.32	99.23	60.14	63.34	87.81	53.64	96.4	65.09	54.57	24.31	97.10	96.62
Emu3 (Ours)	AR	80.96	98.93	79.27	59.64	86.17	44.64	77.71	68.73	37.11	20.92	95.32	97.69

Table 5: **Comparison with state-of-the-art text-to-video models on VBench [33] benchmark.**. We selected 11 out of the 16 evaluation dimensions from VBench, along with the final score, for presentation. Except for Emu3, which is an autoregressive (AR) model, *all other publicly comparable method are diffusion (Diff) models*. The higher metrics indicate the better results.

model across 16 dimensions. Aside from **Emu3**, which is an autoregressive model, all other publicly comparable methods are diffusion models. Nevertheless, as shown in Tab. 5., **Emu3** demonstrates highly competitive results compared to other state-of-the-art models in the overall score. Specifically, while it falls short of the most advanced proprietary models such as Kling [39] and Gen-3 [75], it outperforms the majority of open-source text-to-video models. These results highlight the strong video generation capabilities of **Emu3**.

3.3 Future Prediction

Emu3 can extend videos by predicting future frames. In Fig. 8, we illustrate qualitative examples of video extension, where 2-second videos at 24 FPS are tokenized into discrete vision tokens as context. **Emu3** predicts the subsequent 2 seconds of content in the same form of discrete vision tokens, which can be detokenized to generate future predicted videos. These examples demonstrate that utilizing only next-token prediction facilitates the temporal extension of videos, including the prediction of human and animal actions, interactions with the real world, and variations in three-dimensional animations. Furthermore, by extending the video duration in this manner, our approach is capable of iteratively generating videos that surpass its contextual length. We have observed that successfully expanding future video frames by 8 seconds using 2 seconds of video data as context is feasible.

Method	Pretrained-LLM	SEEDB	OCRB	MMV	POPE	VQAv2	GQA	SQA	TQA	CQA	DVQA	IVQA	AI2D	RWQA	MMMU	MMB
Encoder-based																
InstructBLIP [18]	Vicuna-7B	53.4	276	26.2	–	–	49.2	60.5	50.1	12.5	13.9	–	33.8	37.4	30.6	36.0
IDEFICS-9B [35]	LLaMA-7B	–	252	–	–	50.9	38.4	–	25.9	–	–	–	42.2	42.1	18.4	48.2
QwenVL-Chat [3]	Qwen-7B	58.2	488	–	–	78.2*	57.5*	68.2	61.5	49.8	66.3	–	45.9	49.3	35.9	60.6
LLaVA-1.5 [55]	Vicuna-7B	64.3	318	30.5	85.9	78.5*	62.0*	66.8	46.1	18.2	28.1	25.8	54.8	54.8	35.3	64.3
InternVL-Chat [16]	Vicuna-7B	–	–	–	86.4	79.3*	62.9*	–	57.0	–	–	–	–	–	–	–
mPLUG-Owl2 [95]	LLaMA2-7B	57.8	255	36.5	86.2	79.4*	56.1*	68.7	58.2	22.8	–	–	55.7	50.3	32.7	64.5
ShareGPT4V [14]	Vicuna-7B	–	371	37.6	–	80.6*	63.3*	68.4	60.4	21.3	–	–	58.0	54.9	37.2	68.8
LLaVA-1.6(HD) [56]	Vicuna-7B	64.7	532	43.9	86.5	81.8*	64.2*	70.2	64.9	54.8*	74.4*	37.1	66.6*	57.8	35.1	67.4
VILA [53]	LLaMA2-7B	61.1	–	34.9	85.5	80.8*	63.3*	73.7	66.6	–	–	–	–	–	–	68.9
Encoder-free																
Fuyu-8B(HD) [4]	Persimmon-8B	–	–	21.4	74.1	74.2	–	–	–	–	–	–	64.5	–	27.9	10.7
Chameleon-MT-34B [83]	–	–	–	–	–	69.6	–	–	–	–	–	–	–	–	–	–
Show-o [92]	Phi-1.5-1.3B	–	–	–	73.8	59.3*	48.7*	–	–	–	–	–	–	–	25.1	–
EVE-7B(HD) [19]	Vicuna-7B	56.8	–	25.7	85.0	78.6*	62.6*	64.9	56.8	–	–	–	–	–	–	52.3
Emu3	–	68.2	687	37.2	85.2	75.1*	60.3*	89.2*	64.7	68.6*	76.3*	43.8*	70.0*	57.4	31.6	58.5

Table 6: **Comparison on vision-language benchmarks.** We collect evaluations including: SEEDB: SEEDBench-Img [45]; OCRB: OCRBench [59]; MMV: MMVet [98]; POPE [51]; VQAv2 [27]; GQA [34]; SQA: ScienceQA-Img [60]; TVQA: TextVQA [78]; CQA: ChartQA [61]; DVQA: DocVQA [63]; IVQA: InfoVQA [62]; AI2D [36]; RWQA: RealWorldQA [91]; MMMU [99]; MMB: MMBench [58]. * The images of related training datasets are observed during training.

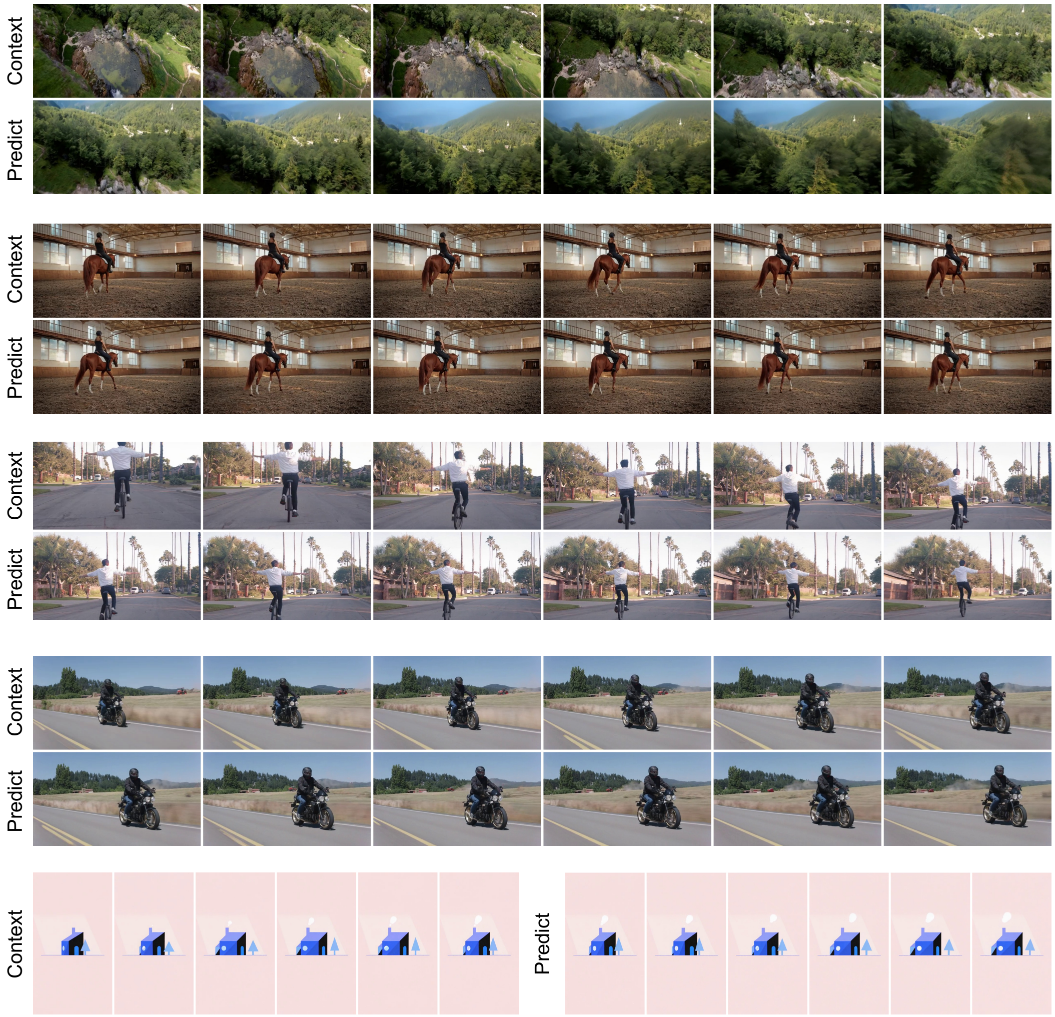


Figure 8: Qualitative results of **Emu3** on video extension. We sample 3 frames per second for display.

3.4 Vision-Language Understanding

To evaluate the vision-language understanding capabilities of **Emu3** fine-tuned in Sec. 2.5.2, we test our model across various public vision-language benchmarks. The primary results, detailed in Tab. 6, compare two categories of methods: **1**) encoder-based approaches that utilize pretrained CLIP vision encoders, and **2**) encoder-free methodologies that operate without pretrained encoders. **Emu3** stands out as a pure encoder-free method, notably surpassing its counterparts across several benchmarks. This achievement is made without depending on a specialized pretrained LLM and CLIP, underscoring intrinsic capabilities and promising potential of **Emu3** in multimodal understanding.

4 Related Work

Vision-Language Understanding. CLIP [67] learns generalizable vision representations through contrastive learning on massive image-text pairs, achieving impressive zero-shot results in image classification tasks. Flamingo [2], by connecting pretrained language models and vision encoders akin to CLIP, initially showcases promising few-shot multimodal understanding capabilities. The increasing availability and progress of LLMs have popularized the fusion of pretrained vision encoders with LLMs, forming a common approach to train extensive vision-language models (VLMs). The BLIP series [49, 48], MiniGPT4 [109], and LLaVA [57] exhibit encouraging results by linking vision encoders with LLMs and training on image-text pairs and vision instruction tuning data. Further improvements in performance are seen in LLaVA series [55, 56] and other impressive works [3, 17] through curated datasets and improved training strategies. While models like Fuyu [4] and EVE [19] introduce encoder-free vision-language architectures that feed image patches into LLMs, they still face challenges in competing with state-of-the-art VLMs. For the first time, **Emu3** demonstrates that a decoder-only model trained solely on next-token prediction can achieve comparable or even superior performance compared to encoder-based VLMs. This paves the way for further improvement of such architecture.

Vision Generation. Recent advancements in vision generation have been largely dominated by diffusion models [73, 70, 66, 65, 5]. These models demonstrate impressive capabilities in generating high-resolution images via the diffusion process. The open-source release of the Stable Diffusion series has led to widespread research and development in this direction. Another research line is to train autoregressive models to generate images via predicting the next token in a sequence, such as DALL-E [71], CogView [20], and Parti [96]. VideoGPT [93] and VideoPoet [38] also leverage autoregressive approaches in the video domain. However, they either fail to match the performance with diffusion models or rely on cascade/compositioinal approaches, *e.g.*, VideoPoet uses a two-stage generate-and-refine framework and an extra text encoder. In this work, **Emu3** demonstrates state-of-the-art image and video generation capabilities with a single Transformer decoder. Notably, we open source to support further research and development in this direction.

Unified Understanding and Generation. There have been early efforts to unify vision understanding and generation [82, 97, 25, 21], exploring various generative objectives on image and text data. Emu and Emu2 [82, 81] introduce a unified autoregressive objective: predicting the next multimodal element, by regressing visual embeddings or classifying textual tokens. CM3Leon [97] and Chameleon [83] trained token-based autoregressive models on mixed image and text data. More recent methods like TransFusion [108] and Show-o [92] attempt to combine diffusion and autoregressive approaches to boost performance. However, these models still fall behind task-specific architectures like SDXL [66] and LLaVA-1.6 [56] in terms of vision generation and understanding. **Emu3** for the first time demonstrates that next-token prediction across images, video, and text can surpass these well-established models, without relying on compositional methods.

5 Conclusion

In this paper, we introduced **Emu3**, a new series of multimodal models that excel at multimodal generation and perception through next-token prediction. By tokenizing images, text, and videos into a discrete space and training a single transformer from scratch, **Emu3** not only eliminates the reliance on diffusion and compositional methods but also surpasses the performance of established task-specific models such as SDXL and LLaVA-1.6. Our results provide compelling evidence that next-token prediction can serve as a powerful paradigm for multimodal models, scaling beyond language models and delivering state-of-the-art performance across diverse tasks, including challenging video generation. We believe that next-token prediction is not only viable but also advantageous in the quest for general multimodal intelligence, bringing us closer to the realization of artificial general intelligence.

Contributors and Acknowledgements

Project Lead

Xinlong Wang

Contributors

* indicates core contributors with equal contributions.

Xiaosong Zhang*, Zhengxiong Luo*, Quan Sun*, Yufeng Cui*, Jinsheng Wang*, Fan Zhang*, Yueze Wang*, Zhen Li*, Qiying Yu*, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang

Senior Leads

Zhongyuan Wang, Yonghua Lin, Tiejun Huang

Acknowledgements
We thank Xuan Liu, Shaokai Nie, Hua Zhou, Yance Jiao, Liao Zhang, Mengyu Lu, Yiwen Shao, Yaohui Chen, Donglin Hao, Mengsi Lv, Teng Dai, Jiakang Liu for their support for the Emu3 project.

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

[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. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models. `https://www.adept.ai/blog/fuyu-8b`, 2023.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023.

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. `https://openai.com/index/sora/`, 2024.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[10] 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.

[11] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024.

[12] 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.

[13] 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.

[14] Lin Chen, Jisong 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.

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[16] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[17] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[18] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267, 2023.

[19] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

[20] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021.

[21] 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. *arXiv preprint arXiv:2309.11499*, 2023.

[22] 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 *Forty-first International Conference on Machine Learning*, 2024.

[23] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *arXiv preprint arXiv:2012.09841*, 2021.

[24] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4744–4753, 2024.

[25] 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.

[26] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[27] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[28] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[29] 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. *Advances in neural information processing systems*, 30, 2017.

[30] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[31] 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.

[32] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[33] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[34] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[35] IDEFICS Research Team. Introducing idefics: An open reproduction of state-of-the-art visual language model. `https://huggingface.co/blog/idefics`, 2023.

[36] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[37] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[38] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[39] Kuaishou. Kling ai. `https://klingai.com/`, 2024.

[40] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[41] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024.

[42] Pika Labs. Pika. `https://pika.art/home/`, 2023.

[43] LAION. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022.

[44] 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.

[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.

[46] 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.

[47] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024.

[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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[49] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[50] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024.

[51] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[52] 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.

[53] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024.

[54] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024.

[55] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[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. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2024.

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[58] 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? *arXiv preprint arXiv:2307.06281*, 2023.

[59] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[60] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[61] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[62] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[63] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021.

[64] OpenAI. Chatgpt. `https://chat.openai.com/`, 2023.

[65] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[66] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[67] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[68] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[69] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

[70] 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.

[71] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[72] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[73] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[74] Runway. Gen-2: Generate novel videos with text, images or video clips. `https://runwayml.com/research/gen-2/`, 2023.

[75] Runway. Gen-3 alpha: A new frontier for video generation. `https://runwayml.com/research/introducing-gen-3-alpha/`, 2024.

[76] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[77] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[78] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[79] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[80] 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.

[81] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[82] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023.

[83] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[84] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[85] 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.

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[87] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[88] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[89] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2024.

[90] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[91] XAI. Realworldqa, 2024.

[92] 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.

[93] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[94] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[95] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024.

[96] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[97] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023.

[98] 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. *arXiv preprint arXiv:2308.02490*, 2023.

[99] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[100] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[101] Bo-Wen Zhang, Liangdong Wang, Jijie Li, Shuhao Gu, Xinya Wu, Zhengduo Zhang, Boyan Gao, Yulong Ao, and Guang Liu. Aquila2 technical report. *arXiv preprint arXiv:2408.07410*, 2024.

[102] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.

[103] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

[104] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*, 2018.

[105] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024.

[106] Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *arXiv preprint arXiv:2209.09002*, 2022.

[107] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024.

[108] 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.

[109] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[110] 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.

A Dataset Details

A.1 Video Dataset

We analyze the distribution of the remaining clips. The duration distribution of the remaining clips is shown in Fig. 9. The flow score distribution of filtered clips is shown in Fig. 10.

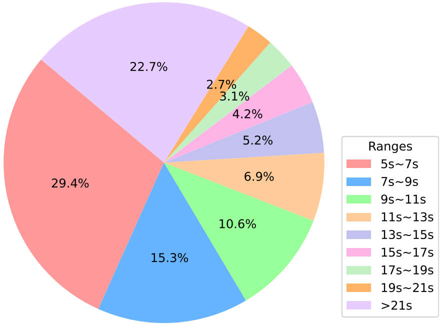


Figure 9: Duration distribution.

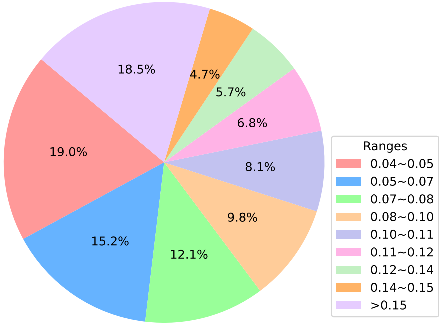


Figure 10: Flow score distribution.

B Evaluation Details

B.1 Image Generation

For all T2I evaluations, we set Top-k to 16,384 and Top-p to 1.0 for image generation. The output resolution for **Emu3** is 512 x 512. The output resolution for **Emu3-DPO** is 720 x 720.

Results on MSCOCO 30K. We present zero-shot CLIP score and FID of **Emu3** and **Emu3-DPO** on MSCOCO 30K in Tab. 4. Following [82], we randomly sample 30k prompts from the validation set and calculate the zero-shot FID [29]. We employ CLIP-ViT-B [67] to calculate the CLIP-T score to assess prompt-following ability. Additionally, we utilize CLIP-ViT-L [67] to compute the CLIP-I score for measuring image similarity. For the DALL-E3 and DALL-E2, CLIP-T score is calculated on 4,096 samples. We adopt classifier-free guidance [30] for better generation quality. The guidance scale is set to 5.0. The results of other methods in the MSCOCO 30K are sourced from [82, 92, 108]

Results on GenEval. Following SD3 [22], we evaluate text-to-image generation capability of **Emu3** on the GenEval benchmark [26]. We present the scores for the GenEval benchmark in Tab. 7 across six dimensions including "Single Object", "Two Objects", "Counting", "Colors", "Position", "Color Attribute". We generate 4 images for each prompt with a guidance scale of 5.5. Following with Dalle-3, we also report our evaluation results utilizing GPT4-V as a rewriter. The results of other methods in the GenEval are sourced from [26, 92, 108, 22].

Results on T2I CompBench. Folloing the Dalle-3 [5], we report the scores of color binding, shape binding and texture binding in Tab. 7. We use the BLIP-VQA model to evaluate these results. We generate 10 images for each prompt with a guidance scale of 5.0. The results of other methods in the T2I CompBench are sourced from [5, 24, 13]

Results on DPG-bench. To assess the ability to follow dense text, we compared our models with state-of-the-art (SoTA) diffusion models on the DPG-Bench, which provides longer prompts containing more detailed information for evaluation. We measured DPG-bench follows [31] shown in the Tab. 8 , and our model achieved an overall score of 81.60, which is higher than SDXL and PixArt-alpha, and is comparable to the results of Dalle-3. We utilized mPLUG-large model to evaluate the generated images according to the designated questions. The results of other methods in the DPG-Benchmark are sourced from [31, 54]. We generate 4 images for each prompt with guidance scale is 5.0.

	GenEval							T2I-CompBench		
Method	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Color	Shape	Texture
Diffusion-based										
DALL-E 2 [70]	0.52	0.94	0.66	0.49	0.77	0.10	0.19	0.5750	0.5464	0.6374
SDv1.5 [73]	0.43	0.97	0.38	0.35	0.76	0.04	0.06	0.3730	0.3646	0.4219
SDv2.1 [73]	0.50	0.98	0.51	0.44	0.85	0.07	0.17	0.5694	0.4495	0.4982
SDXL [66]	0.55	0.98	0.74	0.39	0.85	0.15	0.23	0.6369	0.5408	0.5637
PixArt-alpha [13]	0.48	0.98	0.50	0.44	0.80	0.08	0.07	0.6886	0.5582	0.7044
DALL-E 3 [5]	0.67	0.96	0.87	0.47	0.83	0.43	0.45	0.8110	0.6750	0.8070
SD3 [22]	0.74	0.99	0.94	0.72	0.89	0.33	0.60	-	-	-
Autoregressive meets diffusion										
Show-o [92]	0.53	0.95	0.52	0.49	0.82	0.11	0.28	-	-	-
Transfusion [108]	0.63	-	-	-	-	-	-	-	-	-
Autoregressive-based										
Chameleon [83]	0.39	-	-	-	-	-	-	-	-	-
LlamaGen [80]	0.32	0.71	0.34	0.21	0.58	0.07	0.04	-	-	-
Emu3	0.54	0.98	0.71	0.34	0.81	0.17	0.21	0.6107	0.4734	0.6178
+ Rewriter	0.66	0.99	0.81	0.42	0.80	0.49	0.45	0.7913	0.5846	0.7422
Emu3-DPO	0.52	0.98	0.69	0.33	0.78	0.15	0.16	0.5514	0.4641	0.5476
+ Rewriter	0.64	0.99	0.76	0.38	0.85	0.45	0.40	0.7544	0.5706	0.7164

Table 7: **Comparison with state-of-the-art models on GenEval and T2I CompBench.** Obj.: Object. Attri.: Attribute.

Method	Overvall	Global	Entity	Attribute	Relation	Other
Diffusion-based						
SDv1.5 [73]	63.18	74.63	74.23	75.39	73.49	67.81
SDXL [66]	74.65	83.27	82.43	80.91	86.76	80.41
PixArt-alpha [13]	71.11	74.97	79.32	78.60	82.57	76.96
Playground v2.5 [46]	75.47	83.06	82.59	81.20	84.08	83.50
Lumina-Next [110]	74.63	82.82	88.65	86.44	80.53	81.82
Hunyuan-DiT [52]	78.87	84.59	80.59	88.01	74.36	86.41
PixArt-Sigma [12]	80.54	86.89	82.89	88.94	86.59	87.68
DALLE 3 [5]	83.50	90.97	89.61	88.39	90.58	89.83
SD3-Medium [22]	84.08	87.90	91.01	88.83	80.70	88.68
Playground v3 [54]	87.04	91.94	85.71	90.90	90.00	92.72
Autoregressive-based						
Emu3	80.60	85.21	86.68	86.84	90.22	83.15
Emu3-DPO	81.60	87.54	87.17	86.33	90.61	89.75

Table 8: **Comparison with state-of-the-art models on DPG-bench**

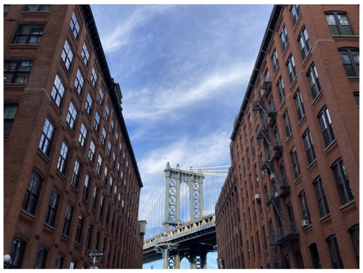
B.2 Post Processing

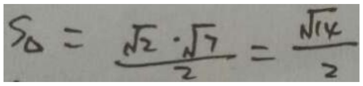
To further improve the temporal consistency and visual quality, we apply stabilization and super resolution techniques to the generated videos. Video evaluation is also conducted on the processed videos. Specifically, we train specialized models for these two tasks.

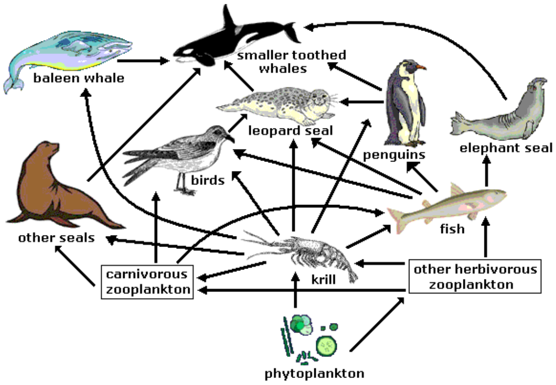
Video Stabilization. We train the video stabilization model based on the temporal VAE of stable video diffusion [7]. The model is trained on our curated video data with a combined objective comprising L1 loss, LPIPS perceptual loss [104], GAN loss, and KL penalty [37, 72]. A training data pair consists of an autoencoded video clip output from our tokenizer and the groundtruth video clip, both having dimensions of $16 \times 256 \times 256$.

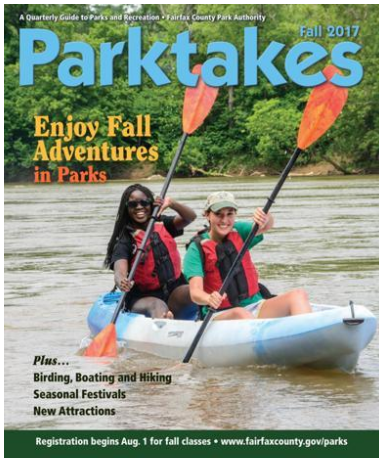
Super-Resolution. We implement a spatial-temporal unet model for super-resolution task, capable of upsampling any image or video clip by a factor of 4. We adopt the BlurPool [103] for downsample operations and sub-pixel [77] for upsample operations. The model is trained on random crops of $8 \times 256 \times 256$ from part of our curated videos, which have a resolution greater than 1024×1024, with a combined loss of L2 loss, LPIPS perceptual loss [104], and GAN loss.

C Qualitative Examples for Multimodal Understanding

Example	
	
User	Describe the landmark.
Emu3	The image showcases the Brooklyn Bridge, a prominent suspension bridge in New York City, recognized by its distinctive towers and the web of cables that support it. This iconic structure connects the boroughs of Brooklyn and Manhattan.

Example	
	 $S_\Delta = \frac{\sqrt{2}\cdot\sqrt{7}}{2} = \frac{\sqrt{14}}{2}$
User	Please write out the expression of the formula in the image using LaTeX format.
Emu3	$S_{\Delta} = \frac{\sqrt{2}\cdot\sqrt{7}}{2} = \frac{\sqrt{14}}{2}$

Example	
	 smaller toothed whales baleen whale leopard seal penguins elephant seal birds other seals fish carnivorous zooplankton krill other herbivorous zooplankton phytoplankton
User	According to the given food chain what would happen if phytoplankton decreases? A. Seal population will become extinct B. Fish population would decrease. C. Whale population would decrease. D. Penguin population would increase. Answer with the option's letter from the given choices directly.
Emu3	B

Example	
	 A Quarterly Guide to Parks and Recreation • Fairfax County Park Authority Fall 2017 Parktakes Enjoy Fall Adventures in Parks Plus... Birding, Boating and Hiking Seasonal Festivals New Attractions Registration begins Aug. 1 for fall classes • www.fairfaxcounty.gov/parks
User	Please describe the image in detail
Emu3	The image is the cover of a magazine titled "Parktakes," which is the Fall 2017 issue. The cover features two individuals in a white kayak, paddling on a body of water. The kayak is positioned diagonally, with the two individuals seated towards the front. Both individuals are wearing life jackets, which are red with black straps, and they are holding paddles with orange blades. The person on the left is a woman with long dark hair, wearing sunglasses and a black top. The person on the right is a man with short hair, wearing a green shirt and a light-colored cap. They both appear to be smiling and enjoying their time on the water. The background shows a natural setting with trees and greenery, indicating that they are likely in a park or natural reserve. The water is calm, and the sky is not visible, suggesting a serene and peaceful environment. The magazine cover includes several text elements: 1. **Title**: "Parktakes" is prominently displayed at the top in large, bold, blue letters. 2. **Fall 2017 Issue**: The issue is marked in the top right corner in a smaller, white font. 3. **Main Feature**: "Enjoy Fall Adventures in Parks" is written in a large, bold, orange font, emphasizing the theme of outdoor activities. 4. **Additional Features**: The cover lists several topics covered in the magazine: - "Birding, Boating, and Hiking" - "Seasonal Festivals" - "New Attractions" 5. **Registration Information**: At the bottom, there is a call to action to "Registration begins Aug. 1 for full classes" and a website link: "www.fairfaxcounty.gov/parks." Analysis and Description The cover of "Parktakes" effectively conveys the theme of outdoor recreational activities, particularly in parks. The image of the two individuals in a kayak suggests that the magazine features content related to kayaking, canoeing, and other water sports. The inclusion of life jackets indicates a focus on safety and preparedness for water activities. The mention of seasonal festivals and new attractions implies that the magazine includes information on various events and attractions within the parks. The cover also highlights the magazine's commitment to providing a comprehensive guide to outdoor activities, with a clear emphasis on safety and accessibility. The registration information suggests that the magazine is a resource for planning and participating in various outdoor activities, likely targeting a wide audience interested in parks and recreational activities. In summary, the magazine cover is designed to attract readers interested in outdoor adventures, safety, and seasonal events within parks. The use of a kayak and life jackets emphasizes the magazine's focus on water sports and safety, while the inclusion of various features and registration information suggests a well-rounded resource for park visitors.