Emu3: Next-Token Prediction is All You Need

Emu3 Team,* **BAAI**
https://emu.baai.ac.cn

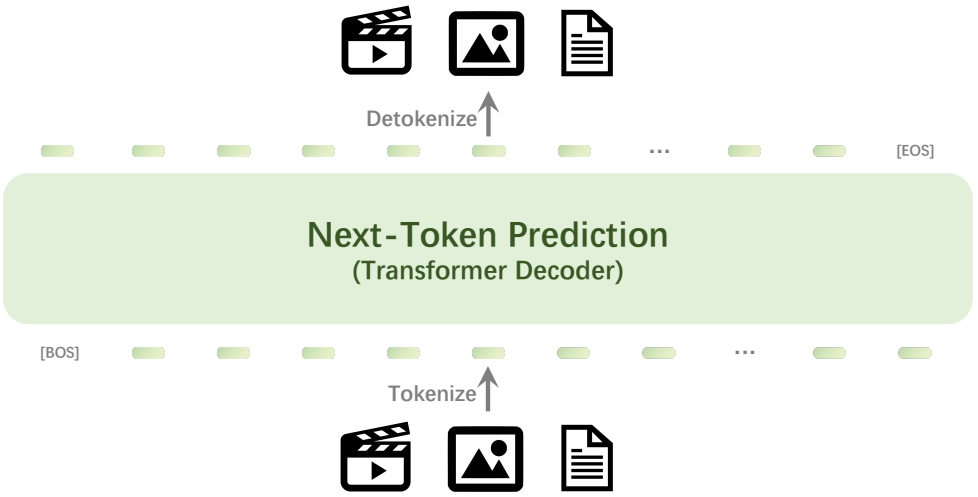

Figure 1: **Emu3** is trained to predict the next token with a single Transformer on a mix of video, image, and text tokens. **Emu3** achieves state-of-the-art performance compared to well-established task-specific models in generation and perception tasks.

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.

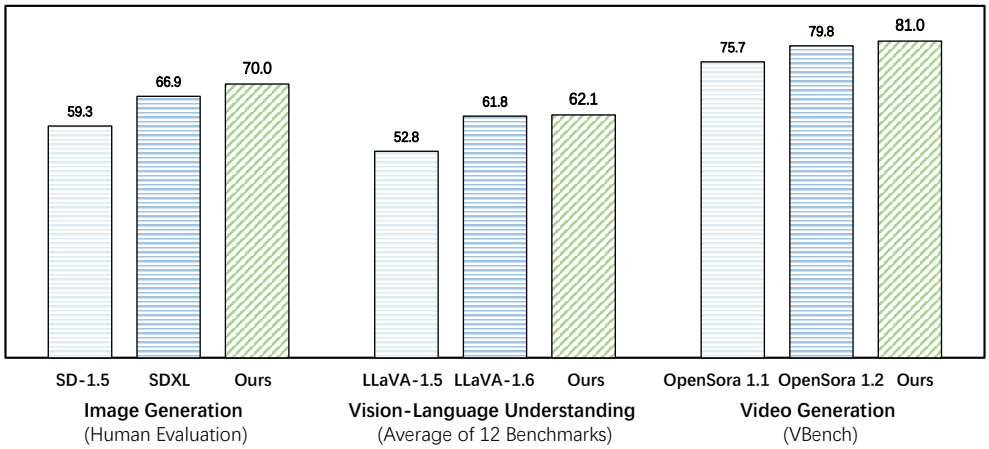

Figure 2: Comparison with open-source flagship models in vision generation and perception. Based solely on next-token prediction, **Emu3** beats SDXL [66], LLaVA-1.6-7B [56], OpenSora-1.2 [107] respectively, dispensing with diffusion and CLIP entirely. For the image generation task, we present comparison 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.