Fan Zhang

135-5235-3162 zhangfan.ryan@gmail.com Google Scholar ryanzhangfan.github.io

Education

Insititute of Software, Chinese Academy of Sciences **Sep. 2017 – Jun. 2020**
Master of Software Engineering *Beijing, China*

Beijing University of Technology **Sep. 2013 – Jun. 2017**
Bachelor of Science in Computer Science (Honors Class) *Beijing, China*

Work Experience

Beijing Academy of Artificial Intelligence **March 2023 – Present**
Researcher *Large Multimodal Model Research Center* *Beijing, China*

- **Emu**: A series of native generative multimodal models capable of seamlessly generating both visual and textual content within a multimodal context.
  - **Emu-1**: We are the first to integrate multimodal visual and textual generation within a single MLLM. **Emu-1** is trained with a unified objective of predicting the next text token and regressing the next visual embedding on large scale interleaved image/video-text data. We employ EVA-CLIP and a causal transformer to extract and compress visual embeddings, which are then processed alongside textual features within the MLLM. We use a diffusion model to decode regressed visual features to high-quality images. **Emu-1** excels in various multimodal understanding and image generation tasks such as VQAv2, MS-COCO while demonstrating novel capabilities like image blending and in-context generation.
  - **Emu-2**: Building on **Emu-1**, we scale the model size to 37B and remove the causal transformer in **Emu-2**. We further train the diffusion decoder directly on features from EVA-CLIP, enhancing both semantic and spatial consistency. Trained on text, interleaved (grounded) image/video-text data, **Emu-2** not only surpasses **Emu-1** but also shows strong in-context abilities and excels in tasks like visual grounding(RefCOCO) and subject-driven generation(DreamBench). It it also shows novel skills like generating from any sequence and chain of editing.
  - **Emu-3**: To unify multimodal tasks in a simple and unified architecture and mitigate error accumulation in continuous visual feature regression as in **Emu-1/2**, we propose **Emu-3**, a new suite of multimodal models trained from scratch solely with next-token prediction. It consists of a MLLM and a video VQGAN which encodes both image and video data into discrete tokens. Trained on large-scale image/video-text pairs and text data, **Emu-3** surpasses flagship models like LLaVA1.6, SDXL and OpenSora in multimodal understanding and image/video generation. In addition to pretrain and supervised finetuning, we also apply direct preference optimization(DPO) to further enhance the generation results.
  - **Emu-3.5**: Extending the spirit of **Emu-3**, we rethink multimodal pretraining from a data scaling perspective, exploring large-scale video data as a fundamental source to advance multimodal model learning. This design strengthens world modeling, enabling the model to represent and reason about complex environments while scaling seamlessly across tasks such as image generation, image–text interleaved generation, and world modeling. Covering the full pipeline from pretraining to reinforcement learning, **Emu-3.5** achieves performance comparable to Nano Banana (Gemini 2.5). In addition, we introduce Discrete Diffusion Adaptation (DiDA), which accelerates image generation by up to 20× without compromising quality.
- **DenseFusion-1M**: We emphasize the importance of hyper-detailed caption dataset for complex visual understanding. To address the scarcity of such data, we propose integrating diverse low budgets visual experts as image priors to provide explicit information on visual elements and adopts an efficient MLLM as a caption engine to emulate the perception abilities of advanced MLLMs. We curated a dataset of 1 million samples and validated the effectiveness of our approach on LLaVA-style models with various base LLMs such as Vicuna, LLaMA-3, Qwen-2, etc.
- **ETT**: We identify several drawbacks of using VQ tokenizers in native multimodal models. 1) The low-level reconstruction objective misaligns with downstream requirements. 2) The use of token indices results in significant information loss, making learning more challenging. 3) The isolation of VQ training from downstream tasks is inherently suboptimal. Hence, we propose using both indices and quantized embeddings as input to downstreams, enabling the end-to-end tuning for both VQ and LLM, boosting the performance compared to the indices counterpart. Our 1.5B model achieves comparable performance compared to Emu3-8B.
- **URSA**: Existing discrete generation paradigms exhibit limited scalability and lack iterative refinement once tokens are generated, constraining global coherence and leaving them behind the continuous diffusion counterparts. We propose a unified discrete diffusion framework that performs global iterative refinement over discrete tokens. Specifically, we introduce a metric path formulation and a resolution-dependent timestep shifting strategy, enabling scalable discrete diffusion training for high-resolution image synthesis and long-duration video generation. The

proposed method achieves comparable performance to SOTA continuous diffusion models (Wan2.1, Bagel) on GenEval, DPG-Bench, and VBench, with no more than 50 inference steps.

Inspir.AI **August 2022 – December 2022**

Senior Algorithm Engineer *Platform* *Beijing, China*

- **Game Art Assets Generation via Diffusion Models**
  - **2D Assets**: To address the scarcity of game art assets, we propose training diffusion models on a large-scale web-crawled anime and game dataset, followed by finetuning on a small set of high quality target source images. I developed an algorithm to extract large volumes of anime data from panel comics books, enabling a more diverse and representative dataset. Additionally, I explored the style injection through textual inversion training on a pre-trained diffusion model to generate the target domain assets.
  - **3D Assets**: We propose to generate 3D assets through a 2D-to-Multiview-to-3D roadmap. For 2D-to-Multiview generation, we use a clip image encoder to extract the image feature of the source-view, which are then paired with a text prompt describing the target view and fed as the condition to train a diffusion model to generate images of the target view.

Bytedance **July 2020 – July 2022**

Computer Vision Engineer *Vision Technology* *Beijing, China*

- **Video Thumbnail Selection**
  - **Content-Based Selection**: Designed a series of video thumbnail selection algorithms tailored to diverse needs of different scenarios. Integrating algorithms such as image quality assessment, face detection, text detection, etc to select the optimal thumbnail while balancing image-text consistency and visual appeal.
  - **User-Preference-Based Selection**: Developed a user-behavior oriented thumbnail selection algorithm by leveraging a Wide&Deep model, incorporating user, video, and image features to train a CTR prediction model. The optimal thumbnail is selected by identifying the image with the highest predicted user engagement.
- **Image/Video Padded Border Detection**
  - **Image-level Algorithm**: Implemented padded border detection via YOLOv3. The model is directly used for image intelligent tailor. For video border detection, we extract multiple frames and applied the model to each frame, the final detection results are refined through a series of post-processing strategies.
  - **Video-level Algorithm**: To incorporate temporal information into image-based models, I propose a new video-level border detection model, which processes multiple frames simultaneously. Besides, the problem is reformulated from multi-object detection to single object regression. These modifications improve the detection P@0.9 by 0.17, has-border accuracy by 0.08 and border type classification accuracy by 0.12.
- **Service Platform**: Designed and developed a service platform which utilize the factory method to integrate basic algorithms to address complex business requirements. The platform is highly extensible, supporting new feature integration through a modular plugin system and enabling flexible, configuration-based composition of functionalities to quickly adapt to evolving business needs. It has been deployed to support multiple business requirements, including video thumbnail selection, video thumbnail generation, video intelligent tailor, image intelligent tailor, etc.

Publication (* for equal contribution, † for corresponding author)

- **Zhang F.**, Chen Y., Li Z., Hong Z., Liu J., et al. Acfnet: Attentional class feature network for semantic segmentation[C]. Proceedings of the IEEE/CVF international conference on computer vision (**ICCV**), 2019.
- Sun Q.*, Yu Q.*, Cui Y.*, **Zhang, F.***, Zhang X.*, Wang Y., et al. Emu: Generative pretraining in multimodality[C]. The Twelfth International Conference on Learning Representations (**ICLR**), 2024.
- Sun Q.*, Cui Y.*, Zhang X.*, **Zhang F.***, Yu Q.*, et al. Generative multimodal models are in-context learners[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024.
- Wang X.*, Zhang X.*, Luo Z.*, Sun Q.*, Cui Y.*, Wang J.*, **Zhang F.***, Wang Y.*, Li Z.*, et al. Emu3: Next-token prediction is all you need[J]. arXiv preprint arXiv:2409.18869, 2024.
- Cui Y.*, Chen H.*, Deng H.*, Huang X.*, Li X.*, Liu J.*, Liu Y.*, Luo Z.*, Wang J.*, Wang W.*, Wang Y.*, Wang C.*, **Zhang F.***, Zhao Y.*, et al. Emu3.5: Native Multimodal Models are World Learners[J]. arXiv preprint arXiv:2510.26583, 2025.
- Li X.*, **Zhang F.***, Diao H.*, Wang Y., Wang X., et al. Densefusion-1m: Merging vision experts for comprehensive multimodal perception[C]. The Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (**NeurIPS D&B**), 2024.

- Wang W.*, **Zhang F.***, Cui Y.*, Diao H.*, Luo, Z., et al. End-to-End Vision Tokenizer Tuning[C]. The Thirty-Ninth Conference on Neural Information Processing Systems (**NeurIPS**), 2025.
- Deng H.*, Pan T.*, **Zhang F.***, Liu Y.*, et al. Uniform Discrete Diffusion with Metric Path for Video Generation[J]. arXiv preprint arXiv:2510.24717, 2025.
- Yu Q.*, Sun Q.*, Zhang X., Cui Y., **Zhang F.**, et al. Capsfusion: Rethinking image-text data at scale[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024.
- Wang W.*, Sun Q.*, **Zhang F.**, Tang Y., Liu J., et al. Diffusion feedback helps clip see better[J]. The Thirteenth International Conference on Learning Representations (**ICLR**), 2025
- Sun Q.*, Wang J.*, Yu Q.*, Cui Y., **Zhang F.**, et al. Eva-clip-18b: Scaling clip to 18 billion parameters[J]. arXiv preprint arXiv:2402.04252, 2024.
- Li T.†, **Zhang F.**, Duan N. Automatic User Preferences Elicitation: A Data-Driven Approach[C]. Requirements Engineering: Foundation for Software Quality: 24th International Working Conference, (**REFSQ**), 2018.
- Liu M. **Zhang F.**, Huang P., Niu S., Ma F., et al. Learning the satisfiability of pseudo-Boolean problem with graph neural networks[C]. Principles and Practice of Constraint Programming: 26th International Conference (**CP**), 2020.
- Ge C., Ma F., Ma X., **Zhang F.**, Huang P., et al. Approximating Integer Solution Counting via Space Quantification for Linear Constraints[C]. International Joint Conferences on Artificial Intelligence (**IJCAI**), 2019.
- Liu M., Huang P., Jia F., **Zhang F.**, Sun Y., et al. Can graph neural networks learn to solve the MaxSAT problem[C] Proceedings of the AAAI Conference on Artificial Intelligence (**AAAI Student Abstract**), 2023.
- Yan R., Zhu D. **Zhang F.**, Lv Y., Yang J., et al. Resource-aware design for reliable autonomous applications with multiple periods. In International Symposium on Formal Methods[C]. International Symposium on Formal Methods(**FM**), 2018.
- Lu Q., Wu T., Yan J., Yan J., Ma F., **Zhang F.** Lightweight method-level energy consumption estimation for android applications[C]. 2016 10th International Symposium on Theoretical Aspects of Software Engineering (**TASE**), 2016.

Awards and Certifications

2018	CVPR WAD Video Segmentation Challenge	4/145
2017	"Star of Tomorrow" Internship Award of Excellence, MSRA	
2016	Chinese Collegiate Programming Contest (CCPC) Hefei Site	Bronze Medal
2016	IEEE XtremeProgramming Competition	56/2144
2016	China Collegiate Computing Contest Group Programming Ladder Tournament	Group Second Prize
2015	National English Competition for College Students	Third prize
2015	Certified Software Professional (CSP)	Top 1.4%