ICLR 2026 Workshop on

Multimodal Intelligence: Next Token Prediction and Beyond

Organizers: Ivona Najdenkoska [1], Mohammad Mahdi Derakhshani [1], Marzieh Fadaee [3], Kai Han [4], Saining Xie [5,6], Yuki M. Asano [2], Cees G. M. Snoek [1]
Affiliations: [1]University of Amsterdam, [2]Technical University of Nuremberg, [3]Cohere Lab, [4]University of Hong Kong, [5]New York University, [6]NYU Center for Data Science

Content

1 Workshop Summary

Foundation models have unlocked impressive capabilities in multimodal intelligence—from open-ended dialogue grounded in visual and auditory inputs to cross-modal reasoning and instruction following. A growing body of work now frames this progress under the umbrella of next-X prediction, where X can represent tokens, frames, or scales across discrete or continuous spaces. Discrete formulations—such as Chameleon [1], which autoregressively predicts mixed-modal tokens—extend next-token prediction beyond language. In contrast, continuous formulations such as VAR [2], MAR [3], TransFusion [4], BAGEL [5], and Fluid [6], are modeling next-scale or next-frame dynamics directly in latent or continuous spaces. Recent models like the Cosmos family [7] and CausVid [8] further generalize this paradigm across modalities and temporal hierarchies (video generation), illustrating that "next-X prediction" has become the central modeling principle unifying modern multimodal foundation models.

In contrast to emitting the next token, predictive encoders learn by forecasting future or missing representations (often without discrete tokens) [9]. The V-JEPA 2 family [10] demonstrates that large-scale joint-embedding prediction on internet video (plus a small amount of interaction data) can yield strong understanding, prediction, and planning performance. After aligning its visual encoder with a large language model, V-JEPA 2 even matched top systems on video QA benchmarks – all *without* generating words token-by-token. The key idea is to capture predictable aspects of the world while ignoring high-entropy details; unlike generative sequence models that spend capacity modeling every pixel or token, a predictive encoder focuses on salient structure (e.g., object trajectories) that underlies intelligent behavior.

Beyond autoregressive and continuous next-X paradigms, discrete diffusion models have recently emerged as a powerful alternative for both language and multimodal generation. Unlike autoregressive models that generate tokens sequentially from left to right, discrete diffusion large language models (dLLMs) treat generation as an iterative denoising process over discrete token sequences. This formulation enables parallel, bidirectional generation with improved structural controllability and global coherence. Building on early diffusion language models such as Diffusion-LM [11], D3PM [12], and DiffuSeq [13], recent large-scale systems like LLaDA [14] and Dream [15] extend this approach to scalable text generation. Following a similar trajectory, discrete diffusion multimodal models including Dimple [16], LLaDA-V [16], and LaViDa [17] integrate visual and textual modalities through multimodal alignment, instruction tuning, and preference optimization. Empirical studies show that these diffusion-based models can match or approach the performance of autoregressive counterparts trained at comparable scales, while offering significant inference speed-ups due to their inherently parallel decoding process.

This workshop provides a timely venue to compare the different paradigms – next-token generation, representation prediction, diffusion-based generators, or maybe even a fourth one. Central questions include:

- Which paradigm yields better representations for downstream tasks?
- How do their scaling behaviors and data efficiency differ?
- Can they be combined, and what can they learn from each other?

By bringing together researchers working on these diverse yet related foundations, this workshop aims to chart a unifying perspective on the next generation of multimodal foundation models—beyond token prediction alone, toward models that truly predict, perceive, and reason about the world.

Topics of Interest

We invite discussions and contributions on a broad range of topics, including but not limited to:

- **Objective comparisons (next-token prediction vs. predictive encoding vs. diffusion)**: Head-to-head studies under matched data/compute, with clear win conditions and ablation of design choices.
- **Hybrid training recipes**: Joint or two-stage objectives (e.g., AR + latent forecasting; diffusion as a refiner), cross-paradigm distillation, and criteria for when hybrids help.
- **Tokenization & representation interfaces**: Discrete tokens vs. continuous latents, visual/audio tokenizers, quantizers, and their impact on grounding, fidelity, and controllability.
- **Scaling & data efficiency**: Scaling behavior across objectives, data mixtures (images/video/audio/text), synthetic/weak supervision, and long-video or streaming data curricula.
- **Evaluation protocols**: Standardized suites for multimodal reasoning, temporal/spatial grounding, planning/control, editing/controllability, and reproducible reporting.
- **Applications & embodied agents**: Robotics and long-horizon control, video QA, document understanding; linking objective choice to downstream performance with rigorous ablations.

2 Workshop Structure / Tentative Schedule

In the following section, we outline the planned schedule and structure of the workshop, including paper presentations, poster sessions, and a panel discussion. The workshop will be organized into morning and afternoon sessions, featuring six invited talks, four short oral presentations, two poster sessions, and a panel discussion. The structure is designed to balance the research overviews from leading scientists with interactive discussions and presentations of emerging work from the community. The tentative schedule is provided in Table 2.

Submission tracks

- **Main track.** The main track welcomes submissions of up to eight pages, excluding references and supplementary materials. We invite high-quality research that presents original, unpublished work or work currently under submission elsewhere. We will also consider recently published papers (e.g., from late 2024 or 2025). Submissions in this track are expected to make technical or conceptual contributions related to multimodal foundation models, next-token prediction paradigms, and efficient or unified modeling frameworks.
- **Tiny papers track.** In addition to the main track, we will feature a Tiny Papers Track for shorter contributions of up to four pages. This track is intended for works-in-progress, exploratory studies, and intermediate research milestones, providing a space for authors to receive feedback and refine early ideas. We encourage submissions from students, early-career researchers, and participants from underrepresented or under-resourced backgrounds to share preliminary findings, exchange ideas, and build collaborations within the community.

Reviewing & Selection Process

All submissions will undergo a double-blind review process on **OpenReview**, ensuring fairness, transparency, and constructive feedback for authors.

- **Peer Review:** Each submission will be evaluated by at least **two independent reviewers** based on novelty, clarity, technical soundness, and relevance to the workshop topics. Reviewers will be encouraged to provide detailed, constructive feedback to support authors in improving their work.
- **Selection Criteria:** Accepted papers will be chosen for **oral or poster presentations** based on overall quality, originality, and potential to spark discussion at the workshop. We will aim for a balanced program representing diverse perspectives and topics.

- **Camera-Ready Submissions:** Authors of accepted papers will be invited to submit a final version incorporating reviewer feedback. Accepted papers and abstracts will be made publicly available on **OpenReview** before the workshop.
- **Best Paper Award:** We will present a Best Paper Award recognizing outstanding technical quality and originality. The selection committee—composed of organizers and invited senior researchers—will evaluate shortlisted papers based on reviewer feedback, presentation quality, and contribution significance.

Workshop Structure

- **Invited talks**: The workshop will open with two invited talks that set the stage for the day, providing a broad perspective on multimodal foundation models and the paradigms shaping their development. Subsequent invited talks will explore complementary viewpoints across next-token modeling, predictive encoders, and diffusion-based language models.
- **Oral presentations**: We will select four contributed papers for short oral presentations, emphasizing works that present innovative ideas, comparisons, or analyses relevant to the workshop's topic. These talks will include studies that investigate trade-offs between next-token prediction and predictive modeling or hybrid architectures combining generative and predictive components. The oral presentations will offer early-career researchers and contributors an opportunity to present their work to a diverse audience and receive expert feedback.
- **Poster sessions**: All accepted papers, including those from the Tiny Papers track, will be presented during two dedicated poster sessions—one in the morning and one in the afternoon. These sessions will foster interactive discussion and knowledge exchange, giving participants an informal setting to engage with presenters and explore diverse research ideas.
- **Panel**: The workshop will conclude with a panel discussion featuring invited speakers and selected organizers. The panel will address key open questions around the future of multimodal foundation models: Which learning paradigm—autoregressive, predictive, or diffusion-based—best scales to world understanding? Can these approaches be unified into a single modeling framework? What trade-offs arise between interpretability, efficiency, and generalization? The panel will be interactive, inviting questions from the audience to foster a lively and forward-looking discussion.

Time	Activity
09:00–09:15	**Opening Remarks**
09:15–09:50	**Invited Talk 1**: *Kaiming He (MIT/Google DeepMind)*
09:50–10:25	**Invited Talk 2**: *Mike Z. Shou (National University of Singapore)*
10:25–10:40	*Coffee Break & Networking*
10:40–11:10	**Oral Presentations** — Selected papers (10 + 5 min each).
11:10–12:30	**Poster Session 1**
12:00–12:30	*Lunch & Poster Session*
12:30–13:05	**Invited Talk 3**: *Hannaneh Hajishirzi (AI2, University of Washington)*
13:05–13:40	**Invited Talk 4**: *Juan-Carlos Niebles (Salesforce Research)*
13:40–14:10	**Oral Presentations** — Selected papers (10 + 5 min each).
14:10–14:25	*Afternoon Coffee Break & Networking*
14:25–15:00	**Invited Talk 5**: *Lucas Beyer (Meta AI)*
15:00–15:35	**Invited Talk 6**: *Chelsea Fin (Stanford)*
15:35–16:35	**Panel Discussion** *with selected speakers*
16:35–17:35	**Poster Session 2**
17:35–17:45	**Closing Remarks & Best paper award**

Table 1: Tentative full-day schedule for the *ICLR 2026 Workshop on Multimodal Intelligence: Next-Token Prediction and Beyond.*

Anticipated audience size

We expect an audience of approximately 150–200 participants, as the topics of multimodal foundation models attract strong interest from both academic and industry communities. This estimation is informed by attendance patterns observed at similar workshops in recent years, mentioned in section 7.

3 Invited Speakers

We have invited the following speakers as leading researchers whose work has made high-impact contributions across computer vision, natural language processing, and multimodal foundation models, both in industry and academia. We plan to select four of these speakers to participate in the panel discussion. The specific names are not listed here, as participation will depend on individual availability and scheduling constraints during ICLR.

Kaiming He (MIT/Google DeepMind) [google scholar] is an Associate Professor of Electrical Engineering and Computer Science at MIT and a Distinguished Scientist at Google DeepMind. He is best known for developing Deep Residual Networks (ResNets), one of the most influential architectures in deep learning, whose residual connections underpin modern models such as Transformers and AlphaFold. His research spans computer vision, deep learning, and self-supervised representation learning, with landmark contributions including Faster R-CNN, Mask R-CNN, MoCo, and MAE. Before joining MIT, he was a Research Scientist at Meta AI and previously worked at Microsoft Research Asia (confirmed).

Mike Z. Shou (National University of Singapore) [google scholar] is an Assistant Professor in the Department of Electrical and Computer Engineering at the National University of Singapore, where he leads research on computer vision and deep learning for video understanding and generation. He received his Ph.D. from Columbia University and was previously a Research Scientist at Facebook AI in Menlo Park. His research focuses on developing next-generation video intelligence systems for understanding complex human actions and events (confirmed).

Hannaneh Hajishirzi (University of Washington/Allen Institute for AI) [google scholar] is an Associate Professor of Computer Science at the University of Washington and a Senior Research Manager at the Allen Institute for AI (AI2). Her research focuses on natural language processing and machine learning, with an emphasis on reasoning, question answering, and multimodal understanding. She co-leads the open language and reasoning projects OLMo and Tulu, demonstrating how open and transparent research can drive scientific progress in generative AI (confirmed).

Juan Carlos Niebles (Salesforce/Stanford University) [google scholar] is Research Director at Salesforce and an Adjunct Professor of Computer Science at Stanford University, where he co-directs the Stanford Vision and Learning Lab. He received his Ph.D. from Princeton University and has previously held research and faculty positions at Stanford and Universidad del Norte (Colombia). He has served as Area Chair for CVPR, ICCV, and ECCV, and as Associate Editor for IEEE TPAMI. His research focuses on computer vision, machine learning, and multimodal AI for video understanding and autonomous agents (confirmed).

Lucas Beyer (Meta AI) [google scholar] is a Research Scientist at Meta AI in Zurich, where he works on vision-language and multimodal modeling. He earned his Ph.D. on deep learning for robotic perception and computer vision at RWTH Aachen University. Prior to his current role, Lucas co-led multimodal research at Google Brain/DeepMind and established the OpenAI office in Zürich. His research portfolio includes work on VLMs (vision-language models), captioning, representation learning, and model efficiency (e.g., SigLIP, CapPa, NoFilter). Lucas has published widely at top-tier conferences such as NeurIPS, CVPR, and ICCV, and maintains a strong open-source orientation (confirmed).

Chelsea Finn (Stanford University) [google scholar] is an Assistant Professor in Computer Science and Electrical Engineering at Stanford University, the William George and Ida Mary Hoover Faculty Fellow, and a co-founder of Physical Intelligence (Pi). Her research interests lie in the capability of robots and other agents to develop broadly intelligent behavior through learning and interaction. To this end, her work has pioneered end-to-end deep learning methods for vision-based robotic manipulation, meta-learning algorithms for few-shot learning, and approaches for scaling robot learning to broad datasets. Her research has been recognized by awards such as the Presidential Early Career Award for Scientists and Engineers, the Sloan Fellowship, and the ACM doctoral dissertation award (tentative).

4 Diversity & Inclusivity

Diversity and inclusion are fundamental to the mission of our workshop. We have made a conscious effort to assemble an organizing committee and invited speakers that reflect different races, genders, geographies, academic/industry affiliations, levels of seniority, and backgrounds across the global AI community.

a) Diversity of the organizing team & invited speakers: Our organizers and invited speakers represent both academia and industry, spanning institutions in Europe, the Middle East, North America, and Asia. We include researchers at various career stages - from PhD students and early-career scientists to established faculty and senior industry researchers. The team is gender-balanced and ethnically diverse, bringing together individuals from different regions and cultures around the world. We also prioritize the inclusion of first-time organizers (Ivona Najdenkoska and Mohammad Mahdi Derakhsani), ensuring that the workshop not only highlights established experts but also nurtures the next generation of AI researchers. We believe that this diverse selection not only strengthens the scientific scope of the workshop but also ensures that participants are exposed to a broad set of ideas, experiences, and perspectives.

b) Early-career and Student Researchers: We are strongly committed to fostering the participation and visibility of early-career researchers and students. Recognizing that conducting machine learning research often requires significant resources and institutional support, we aim to create an inclusive environment where emerging scholars can meaningfully contribute and engage with the community. To this end, our workshop will include a **Tiny Papers track** designed to encourage the submission of late-breaking or small-scale research that would benefit from feedback and discussion at ICLR. In addition, we will offer a **Best Paper Award** to recognize outstanding contributions. We are also seeking sponsorship for travel support to ensure equitable access and engagement for all participants.

5 Advertising the Workshop

To build visibility and engagement, we will establish a comprehensive outreach strategy that includes a **workshop website** and an active **Twitter/X** and **Bluesky** accounts. The workshop call for papers will be circulated through established mailing lists, while organizers will use their professional networks to extend personal invitations to both academic and industry researchers. We will leverage the **ELLIS Network** to advertise the workshop, as several members of the organizing team are active ELLIS Fellows and Members, providing strong connections within the European AI research community. We will also highlight the event within open and inclusive research groups such as **ML Collective** and **Cohere Lab** to attract participation from students and early-career scholars. Because the organizing committee spans multiple institutions, each member will promote the workshop locally to reach a broader audience. Invited speakers will likewise be encouraged to share information about the event through their own institutional channels and communities. A dedicated publicity chair from the organizing team will oversee all communications and ensure consistent activity across platforms commonly used by the machine learning community.

6 Workshop Access

All talks and panel discussions will be recorded and made publicly available after the event to ensure long-term accessibility and visibility of the presented work. During each session, participants will be encouraged to ask questions using in-room microphones, fostering direct and engaging exchanges with speakers. The workshop website will act as a lasting repository—featuring the full program schedule, accepted papers, poster galleries, video recordings, presentation materials, and summaries of discussions and insights from breakout sessions. To support communication and coordination, *Whova* will serve as the primary platform for sharing updates, session details, and logistical information during the workshop days. Finally, we plan to explore a follow-up special issue or curated collection featuring selected contributions, and we will invite participants to provide anonymous feedback on the workshop's content and organization to guide future editions and assess its broader impact.

7 Previous Related Workshops

While there have been several workshops over the past few years exploring topics adjacent to multimodal and generative modeling, none have directly targeted the unified study of **next-token prediction** as a canonical interface across modalities and tasks. The following workshops at ICLR, ICML, NeurIPS, ECCV, and CVPR have touched on complementary directions:

- ICLR 2024: Generative Models for Decision Making (bridging generative modeling with reinforcement learning and control).
- ICML 2024: Multi-Modal Foundation Models meet Embodied AI (MFM-EAI) (linking large multimodal models with perception, action, and robotics), Trustworthy Multi-Modal Foundation Models and AI Agents (TiFA) (focused on robustness, safety, and reliability in multimodal systems).
- NeurIPS 2024: Responsibly Building Next Generation of Multimodal Foundation Models (emphasizing ethical and responsible design), Multimodal Algorithmic Reasoning (MAR) (investigating reasoning and compositionality across modalities).
- ECCV 2024: Audio-Visual Generation and Learning (AVGenL) (joint modeling of sight and sound for scene understanding and synthesis), OmniLabel: Vision and Language Foundations for Complex Perception (exploring fine-grained vision-language grounding and open-world perception).
- ICLR 2025: World Models: Understanding, Modeling, and Scaling (scaling multimodal generative models for predictive simulation), Reasoning and Planning for Large Language Models (bridging reasoning and multimodal decision-making in large models).
- CVPR 2025: Emergent Visual Abilities and Limits of Foundation Models (EVAL-FoMo 2) (focusing on analysis and evaluation of emergent visual capabilities and failure modes in visual foundation models), Efficient Large Vision Models (eLVM) (addressing efficiency in large vision models, including architectures, inference strategies, and generative acceleration for vision-centric systems), PixFoundation: Workshop on Pixel-level Vision Foundation Models (focusing on pixel-level foundation models within vision-centric foundational modeling).

Recent workshops have made significant progress across related domains—from generative decision-making and embodied intelligence to multimodal reasoning, efficiency, and trustworthy alignment—but each has largely focused on a single aspect of multimodal modeling. A unified perspective that **integrates next-token generation, predictive encoding, and diffusion** as core paradigms across modalities remains largely unexplored. The proposed **ICLR 2026 Workshop on Multimodal Intelligence: Next Token Prediction and Beyond** seeks to bridge this gap by convening researchers in generative modeling, embodied AI, and multimodal reasoning to establish shared principles, training objectives, and evaluation standards for advancing unified multimodal intelligence.

8 Organizers & Biographies

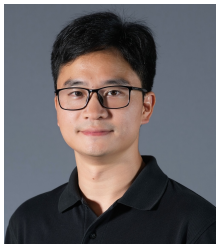
Our organizing team is composed of 7 members - a mixture of PhD students, postdoctoral researchers, professors, and industry researchers with diverse demographics, backgrounds, seniority, expertise, and organizational experiences.

Point of contact: Ivona Najdenkoska and Mohammad Mahdi Derakhshani.

Ivona Najdenkoska [website; google scholar; email: `i.najdenkoska@uva.nl`] is a Postdoctoral Researcher at the University of Amsterdam, working on multimodal foundation models, where she is part of the MultiX Amsterdam research group. Her research centers on designing efficient approaches for multimodal understanding and generative tasks. Ivona was a Research Scientist Intern at Meta GenAI in 2023, where she worked on image generation and in-context learning. She is a member of the ELLIS Society and has served as a reviewer for leading machine learning and computer vision conferences, including CVPR, ICCV, NeurIPS, ICLR, and ICML.

Mohammad Mahdi Derakhshani [website; google scholar; email: `m.m.derakhshani@uva.nl`] is a Ph.D. student at the University of Amsterdam, where he is part of the Video & Image Sense Lab (VISLab) focusing on multimodal foundation models. His research explores scalable generative methods for multimodal understanding and generation. Previously, he was a Research Scientist Intern at Samsung AI Research (2022) and Microsoft Research (2023), working on few-shot learning, image generation, and federated learning. He is a Cohere Lab Scholar, a member of the ELLIS Society, and has served as a reviewer for top-tier conferences including CVPR, ICCV, NeurIPS, ICLR, and ICML.

Marzieh Fadaee [website; google scholar; email: marzieh@cohere.com] is the Head of Cohere Labs, where she leads research on fundamental problems in artificial intelligence. Her work spans multilingual language models, data-efficient learning, model evaluation, and trustworthy AI, with a focus on building systems that are robust, inclusive, and globally impactful. She co-leads the Aya initiative that brought together over 3,000 collaborators worldwide to create the world's largest multilingual instruction dataset and develop state-of-the-art multilingual language and vision models. Before joining Cohere Labs, Marzieh was the Research Lead at Zeta Alpha Vector, where she pioneered innovative approaches to knowledge discovery and organization. She holds a Ph.D. from the University of Amsterdam, where she conducted foundational research on neural machine translation. Her work has appeared in NeurIPS, ACL, EMNLP, and ICLR.



Kai Han [website; google scholar; email: kaihanx@hku.hk] is an Assistant Professor in the Department of Computer Science at The University of Hong Kong, where he directs the Visual AI Lab. His research focuses on computer vision, machine learning, and artificial intelligence, with a particular interest in open-world learning, 3D vision, generative AI, and foundation models. The goal of his work is to advance comprehensive visual understanding and develop reliable AI systems for open-world applications. Previously, he was a Visiting Faculty Researcher at Google Research, an Assistant Professor at the University of Bristol, and a Postdoctoral Researcher in the Visual Geometry Group (VGG) at the University of Oxford. He received his Ph.D. in Computer Science from the University of Hong Kong.

Saining Xie [website; google scholar; email: saining.xie@nye.edu] is an Assistant Professor of Computer Science at the Courant Institute of Mathematical Sciences and a member of the CILVR group at New York University. He is also affiliated with the NYU Center for Data Science. Before joining NYU, he was a Research Scientist at Facebook AI Research (FAIR) in Menlo Park. Saining received his Ph.D. and M.S. in Computer Science from the University of California, San Diego, advised by Zhuowen Tu, and his B.Eng. from Shanghai Jiao Tong University. His research focuses on computer vision and machine learning, with the goal of developing scalable and robust visual intelligence systems capable of interpreting visual events, reasoning about them, and forming a common-sense understanding of the world.

Yuki M. Asano [website; google scholar; email: yuki.asano@utn.de] is a Full Professor of Computer Science and Head of the Fundamental AI (FunAI) Lab at the Technical University of Nuremberg. His research focuses on representation learning, multimodal learning, and efficient foundation models for vision and language. Before joining TU Nuremberg, he led the QUVA Lab at the University of Amsterdam in collaboration with Qualcomm AI Research. Yuki completed his Ph.D. at the Visual Geometry Group (VGG) at the University of Oxford, where he worked with Andrea Vedaldi and Christian Rupprecht. He has extensive experience in community engagement and academic organization, having previously co-organized multiple workshops, tutorials, and research events at major conferences such as ICLR, NeurIPS, ECCV, and CVPR.

Cees G. M. Snoek [website; google scholar; email: c.g.m.snoek@uva.nl] is a Full Professor of Computer Science at the University of Amsterdam, where he heads the Video & Image Sense Lab and the interdisciplinary Human-Aligned Video AI Lab. He is also Director of several public–private AI research labs, including QUVA Lab with Qualcomm, Atlas Lab with TomTom, and OpenBots Lab with Delft University of Technology, TNO, and the Ministry of Defence. At university spin-off Kepler Vision Technologies, he serves as Chief Scientific Officer. His research focuses on making sense of video and images through deep learning and foundation models. He has published over 300 refereed papers and frequently serves as Area Chair at major computer vision and machine learning conferences. He is an ELLIS Fellow and has been active in organizing multiple workshops at major conferences, such as CVPR and events, such as the ELLIS Winter School on Foundation Models (2025, 2024).

References

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

[2] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *Advances in neural information processing systems*, 2024.

[3] T. Li, Y. Tian, H. Li, M. Deng, and K. He, "Autoregressive image generation without vector quantization," *Advances in Neural Information Processing Systems*, 2024.

[4] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, "Transfusion: Predict the next token and diffuse images with one multi-modal model," *arXiv preprint arXiv:2408.11039*, 2024.

[5] C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song, *et al.*, "Emerging properties in unified multimodal pretraining," *arXiv preprint arXiv:2505.14683*, 2025.

[6] L. Fan, T. Li, S. Qin, Y. Li, C. Sun, M. Rubinstein, D. Sun, K. He, and Y. Tian, "Fluid: Scaling autoregressive text-to-image generative models with continuous tokens," *arXiv preprint arXiv:2410.13863*, 2024.

[7] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, *et al.*, "Cosmos world foundation model platform for physical ai," *arXiv preprint arXiv:2501.03575*, 2025.

[8] T. Yin, Q. Zhang, R. Zhang, W. T. Freeman, F. Durand, E. Shechtman, and X. Huang, "From slow bidirectional to fast autoregressive video diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[9] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

[10] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, *et al.*, "V-JEPA 2: Self-supervised video models enable understanding, prediction and planning," *arXiv preprint arXiv:2506.09985*, 2025.

[11] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *Advances in neural information processing systems*, vol. 35, pp. 4328–4343, 2022.

[12] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," *Advances in neural information processing systems*, vol. 34, pp. 17981–17993, 2021.

[13] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," *arXiv preprint arXiv:2210.08933*, 2022.

[14] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," *arXiv preprint arXiv:2502.09992*, 2025.

[15] J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, and L. Kong, "Dream 7b." https://hkunlp.github.io/blog/2025/dream, 2025.

[16] R. Yu, X. Ma, and X. Wang, "Dimple: Discrete diffusion multimodal large language model with parallel decoding," *arXiv preprint arXiv:2505.16990*, 2025.

[17] Z. You, S. Nie, X. Zhang, J. Hu, J. Zhou, Z. Lu, J.-R. Wen, and C. Li, "Llada-v: Large language diffusion models with visual instruction tuning," *arXiv preprint arXiv:2505.16933*, 2025.