Audiovisual demos

avdit2024.github.io

A Versatile Diffusion Transformer with Mixture of Noise Levels for Audiovisual Generation

Gwanghyun Kim[1,†,*] Alonso Martinez[*] Yu-Chuan Su[*] Brendan Jou[2] Jose Lezama[2] Agrim Gupta[2]
Lijun Yu[2] Lu Jiang[*] Aren Jansen[2] Jacob Walker[2] Krishna Somandepalli[2,†]

[1]Seoul National University [2]Google DeepMind

([*]Work done while at Google. [†]Equal contribution.)

Motivation

- Recent years have witnessed a remarkable surge in the development and exploration of multimodal diffusion models. Prominent examples include text-to-image (T2I), text-to-video (T2V).

Imagen 3[1]

"Photographic portrait of a real life dragon resting peacefully in a zoo, curled in a zoo, curled up next to its pet sheep. Cinematic movie still, high quality DSLR photo"

Veo[2]

"A lone cowboy rides his horse across an open plain at beautiful sunset, soft light, warm colors"

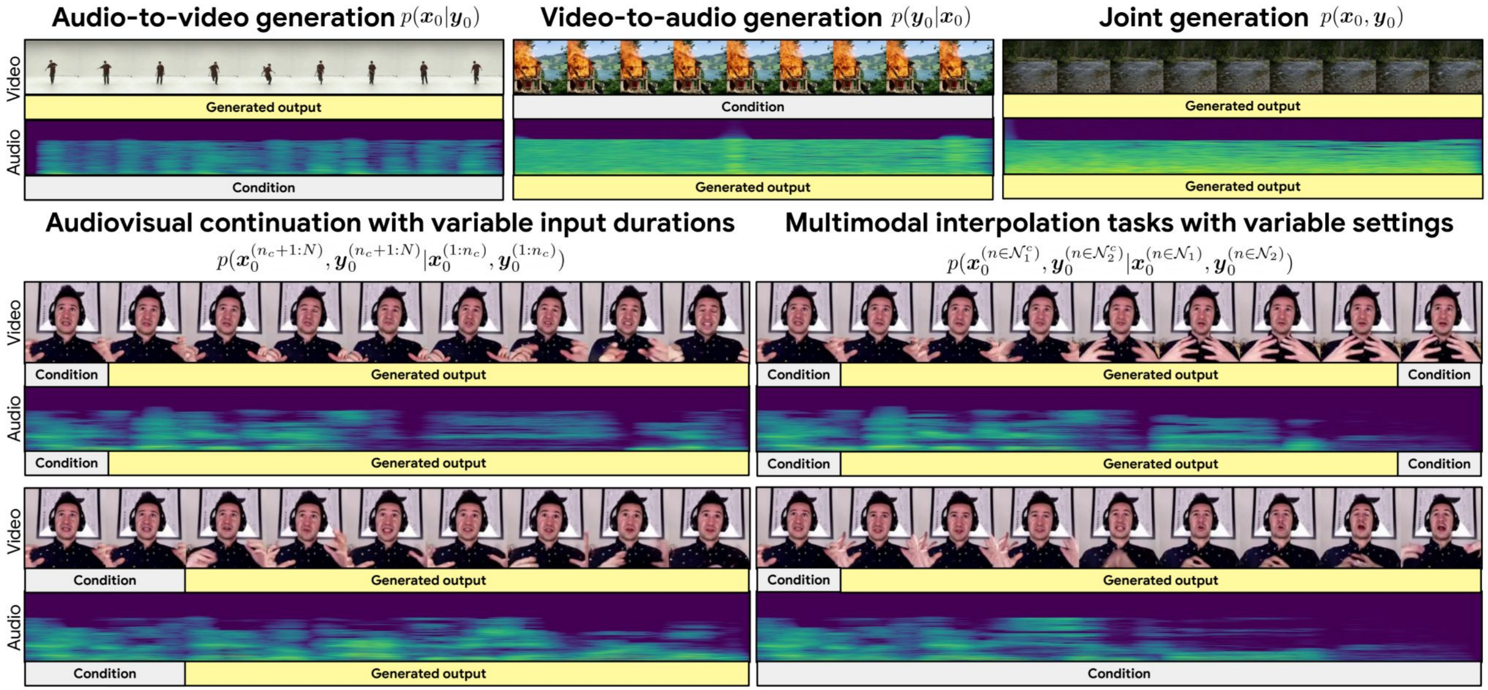
[1] Imagen 3 Team, Google. "Imagen 3", arXiv 2024, https://deepmind.google/technologies/imagen-3
[2] Google DeepMind. "Veo", https://deepmind.google/technologies/veo

Motivation

- Despite notable advancements, generating sequences across multiple modalities, like video and audio, remains challenging and is an open research area.
- Such capability would enable creating realistic, expressive, and controllable multimedia content, while also fostering cross-modal understanding of temporal signals.

Motivation

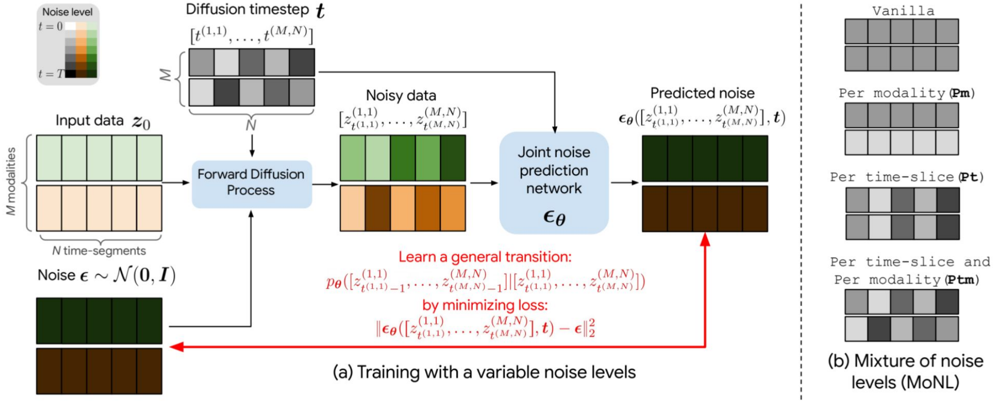
- Training diffusion models for audiovisual sequences allows for a range of generation tasks by learning conditional distributions of various input-output combinations of the two modalities.
- Training separate models for each variation is expensive and impractical.



MoNL & AVDiT

- In this work, we propose a novel Mixture of Noise Levels (MoNL) to effectively learn the arbitrary conditional distributions in the audiovisual space.
- We apply this approach for audiovisual generation by developing a latent-based audiovisual diffusion transformer (AVDiT).

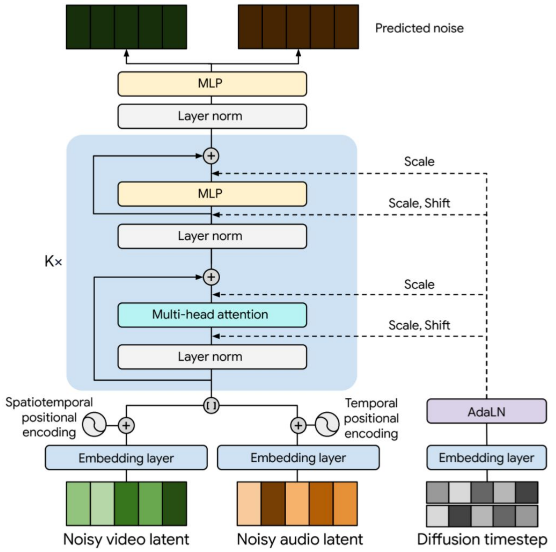
Training with MoNL



(a) Training with a variable noise levels

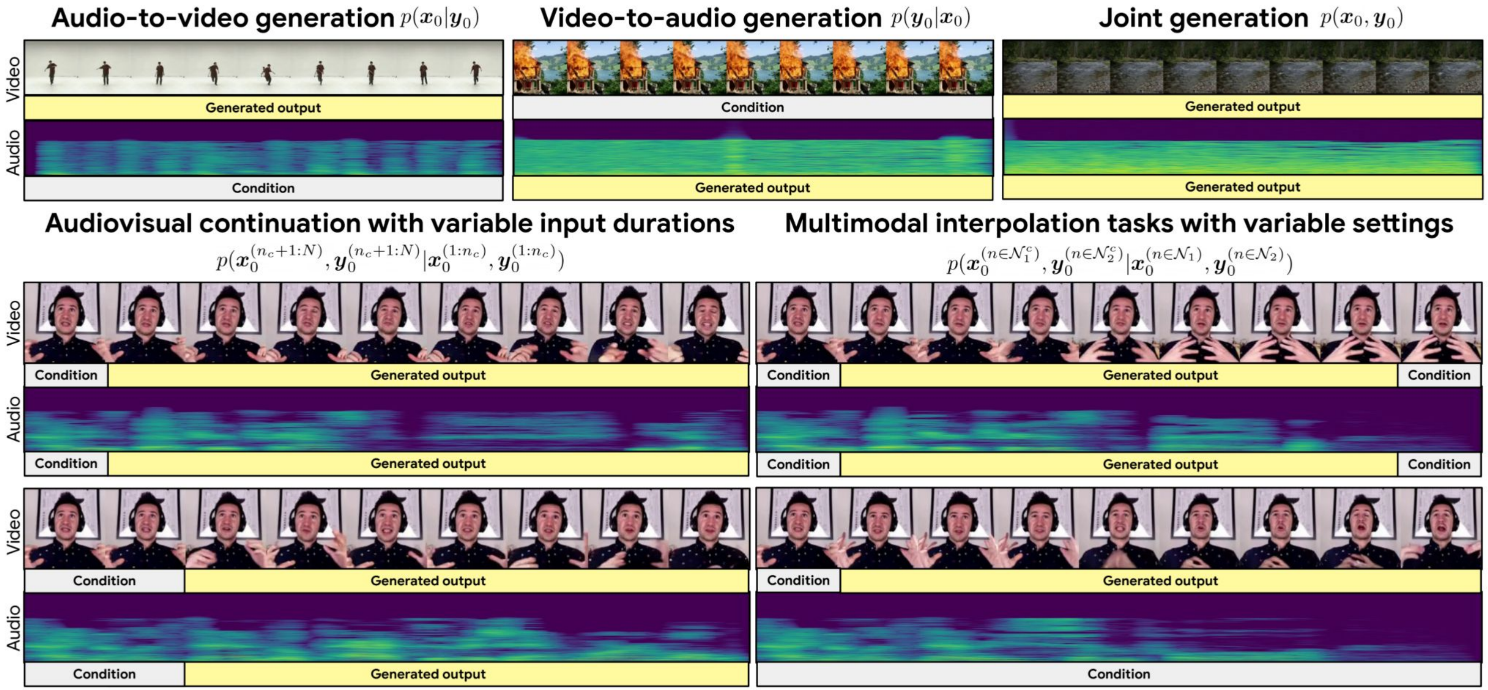
(b) Mixture of noise levels (MoNL)

Schematic of AVDiT



MoNL & AVDiT

- Our method lets a single model handle diverse audio-video generation tasks, creating temporally consistent audio-video sequences and saving training time and resource.

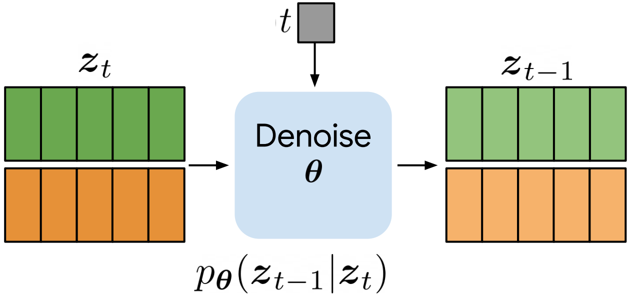


Mixture of Noise Levels (MoNL)

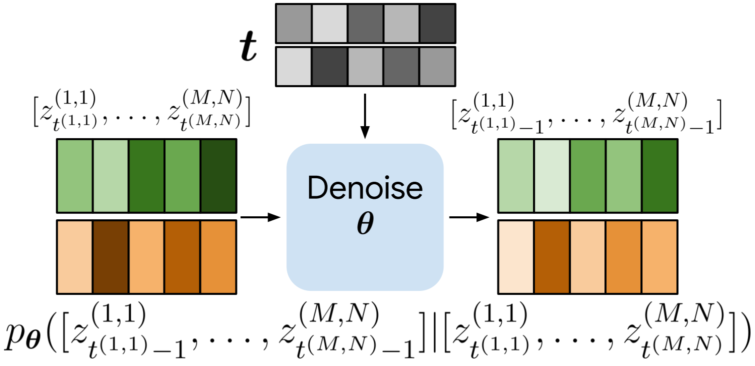
- Our goal is now to learn a general transition matrix between the various modalities and time-segments at each step:

$$p_{\boldsymbol{\theta}}([z^{(1,1)}_{t^{(1,1)}-1}, \ldots, z^{(M,N)}_{t^{(M,N)}-1}] \| [z^{(1,1)}_{t^{(1,1)}}, \ldots, z^{(M,N)}_{t^{(M,N)}}])$$

Vanilla Diffusion



MoNL(Ours)



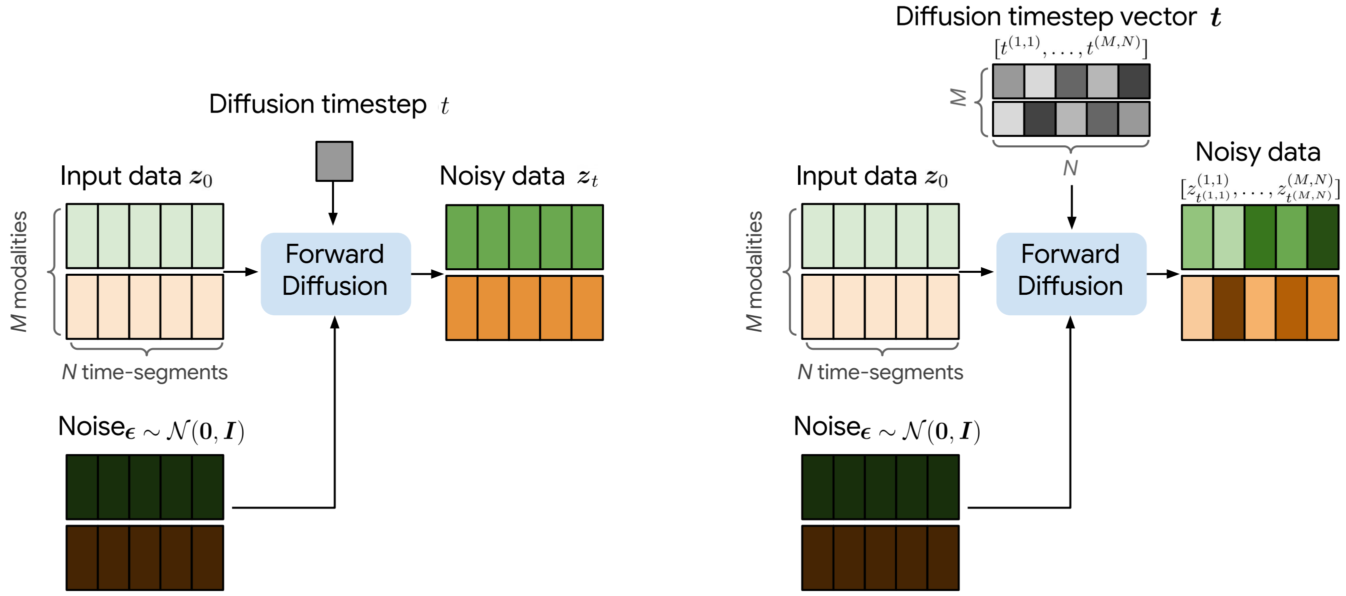
Mixture of Noise Levels (MoNL)

- During the forward diffusion, we can add variable level of noise to each element of input data as :

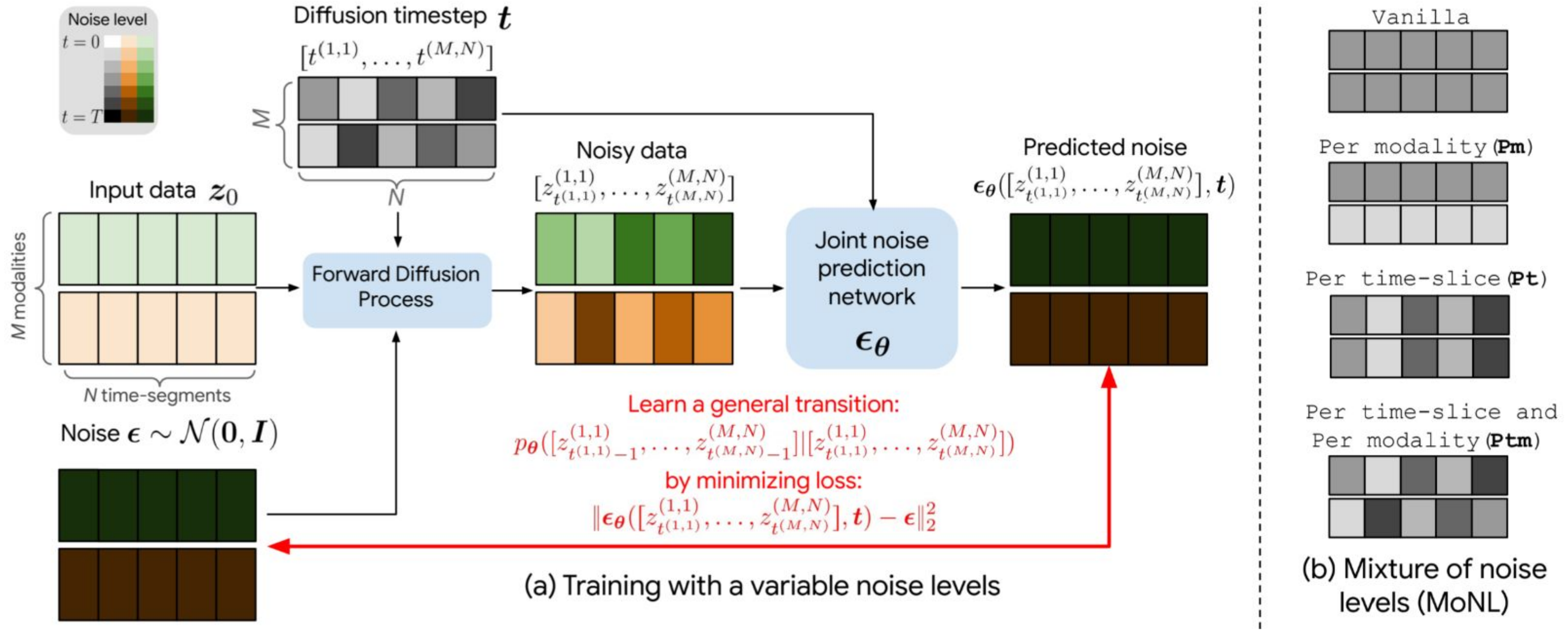
$$z_{t(m,n)}^{(m,n)} = \sqrt{\overline{\alpha}_{t(m,n)}}\, z_0^{(m,n)} + \sqrt{1-\overline{\alpha}_{t(m,n)}}\, \epsilon^{(m,n)}$$

Vanilla Diffusion

MoNL(Ours)



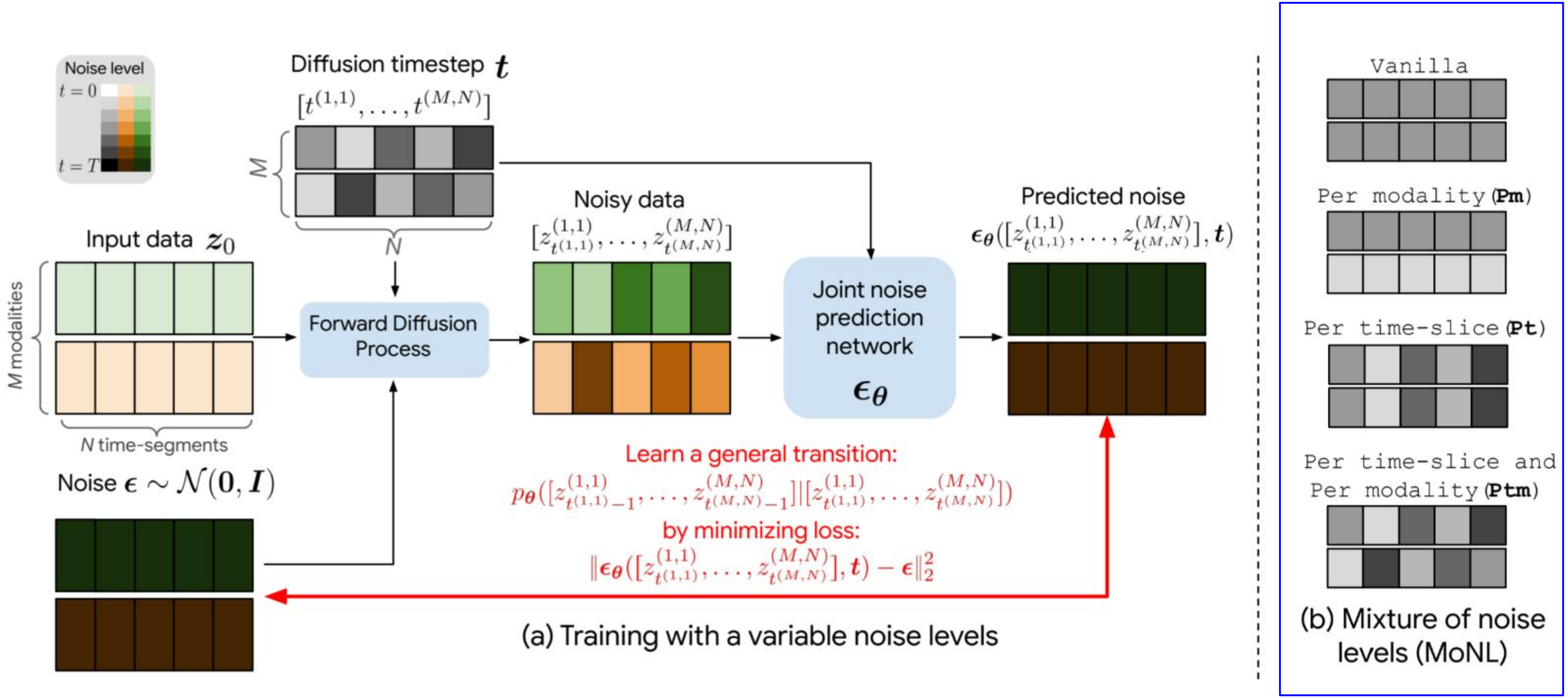
Mixture of Noise Levels (MoNL)



(a) Training with a variable noise levels

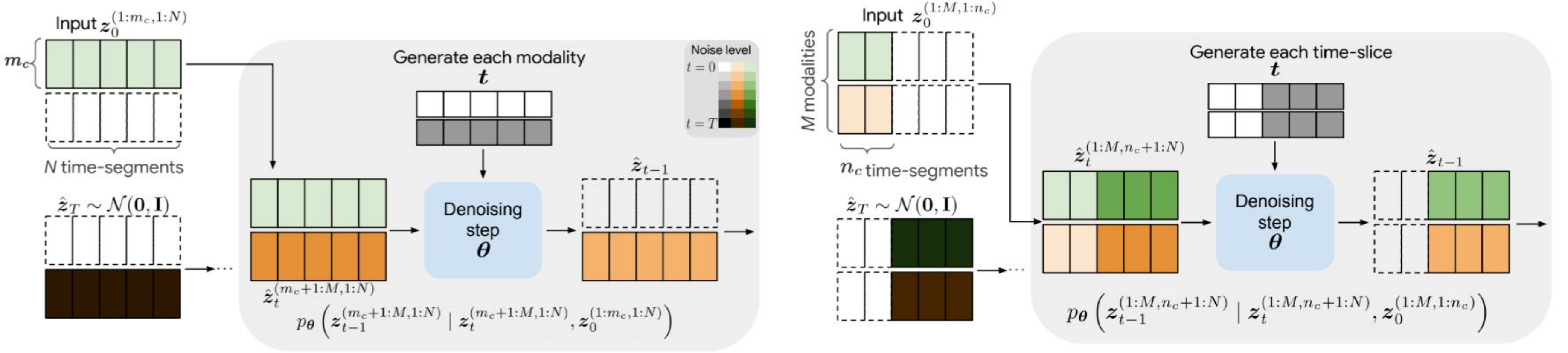
(b) Mixture of noise levels (MoNL)

Mixture of Noise Levels (MoNL)



Mixture of Noise Levels (MoNL) - Conditional Inference

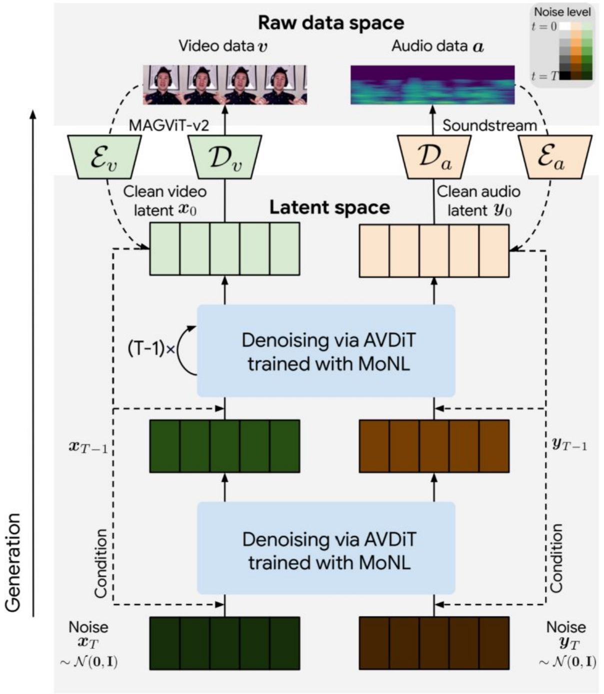
- We achieve arbitrary conditional distributions by selectively injecting inputs during inference based on the task specification
  - Clean inputs → conditional portions with $t^{(m,n)} = 0$
  - Noisy inputs → generating desired portions with the current diffusion step $t^{(m,n)} = t$



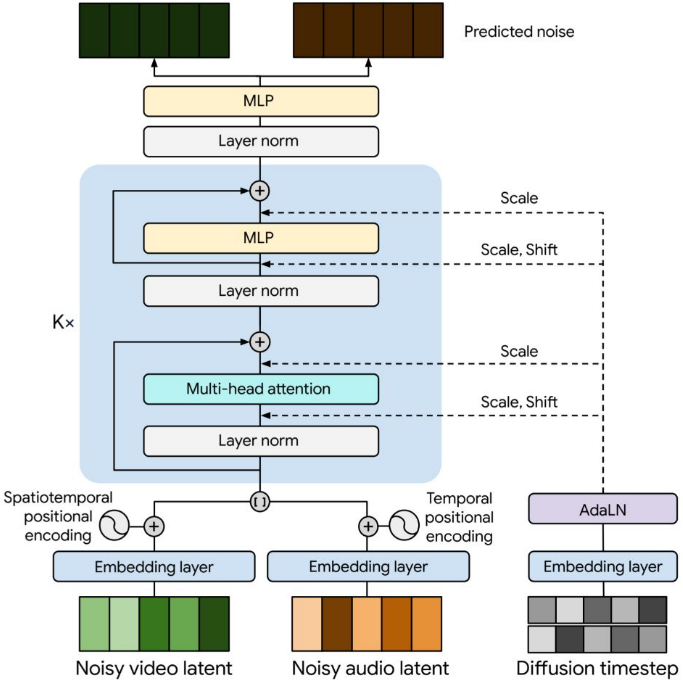
(a) Conditioning across modalities: cross-modal generation

(b) Conditioning across time-segments: multimodal interpolation

Audiovisual Latent Diffusion Transformer (AVDiT)

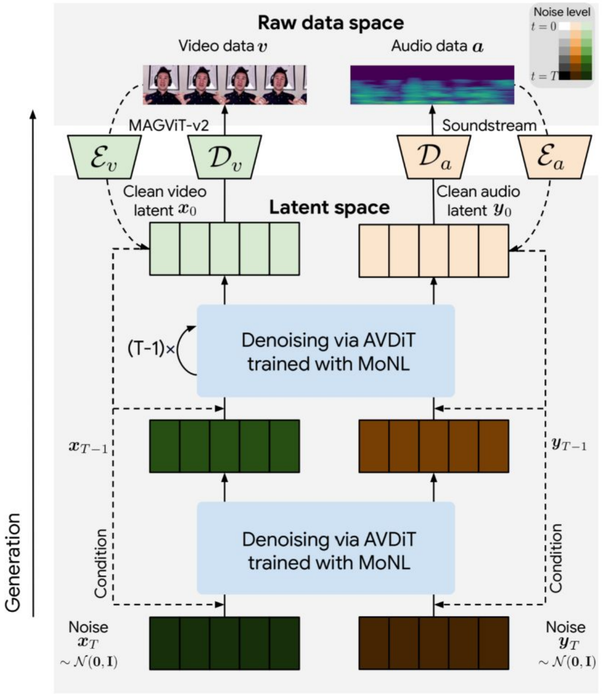


(a) Latent diffusion with mixture of noise levels (MoNL) and audiovisual diffusion transformer (AVDiT)



(b) Audio-video diffusion transformer (AVDiT)

Audiovisual Latent Diffusion Transformer (AVDiT)

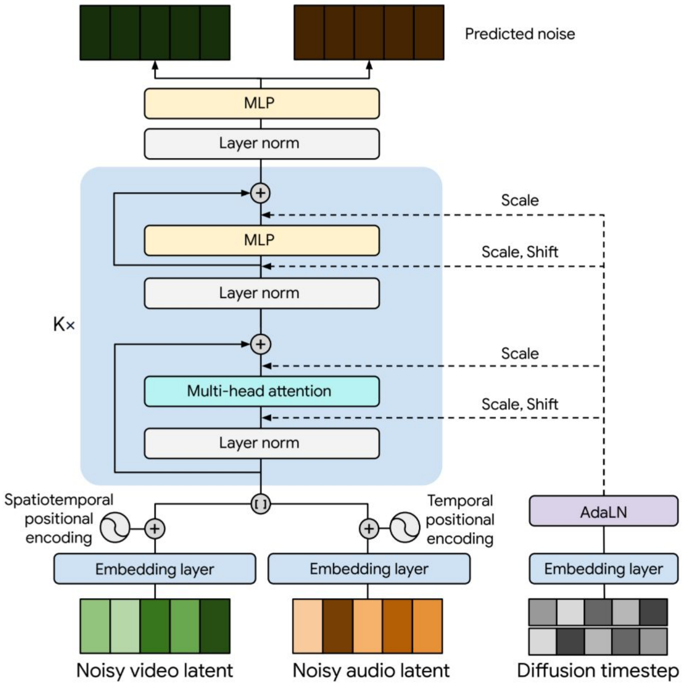


(a) Latent diffusion with mixture of noise levels (MoNL) and audiovisual diffusion transformer (AVDiT)

- We implement MoNL in the low-dimensional latent space learned by the MAGVIT-v2[1] for video and the SoundStream[2] for audio.
- Importantly, the temporal structure in these latent representations enables us to apply variable noise levels.

[1] Yu et al. "Language Model Beats Diffusion--Tokenizer is Key to Visual Generation." arXiv 2023
[2] Zeghidour et al. "Soundstream: An end-to-end neural audio codec." T-ASLP 2021

Audiovisual Latent Diffusion Transformer (AVDiT)



(b) Audio-video diffusion transformer (AVDiT)

- We also introduce a transformer-based network for joint noise prediction.
- Transformers are a natural fit for multimodal generation as they can
  - efficiently integrate multiple modalities[1] and their interactions,
  - capture intricate spatiotemporal dependencies[2],
  - have shown impressive video generation[3] capabilities.

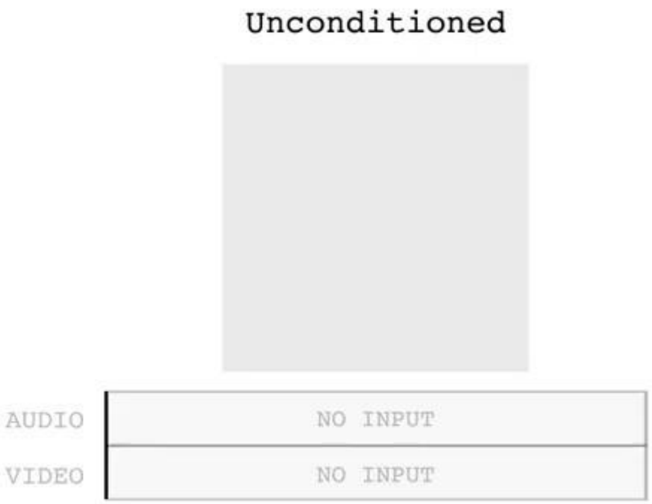
[1] Georgescu, Mariana-Iuliana, et al. "Audiovisual masked autoencoders." *CVPR 2023*
[2] Zhou, Luowei, et al. "Unified vision-language pre-training for image captioning and vqa." *AAAI 2020*
[3] Gupta, Agrim, et al. "Photorealistic video generation with diffusion models." *arXiv* 2023

Demo of our Model Trained on Monologue Dataset

Audiovisual demos

avdit2024.github.io

Demo of our Model Trained on Landscape



Audiovisual demos

avdit2024.github.io

Demo of our Model Trained on AIST++

Audiovisual demos

avdit2024.github.io

Effectiveness of MoNL - Quantitative Comparison

- On average across all tasks, AVDiT trained with MoNL outperforms all baselines, demonstrating its versatility to learn diverse conditional distributions.

Table 1: Comparison of AVDiT trained with mixture of noise levels (MoNL) on the Monologues dataset for unconditional joint generation (`Joint`), cross-modal (`A2V`, `V2A`) and multimodal interpolation (`AV-inpaint`, `AV-continue`) tasks. FAD $= 2.7$ and FVD $= 3.3$ for groundtruth autoencoder reconstructions of the inputs. Fréchet metrics estimated with N=25k.

Setting / Task	Joint		A2V	V2A	AV-inpaint		AV-continue		Average	
	FAD ↓	FVD ↓	FVD ↓	FAD ↓	FAD ↓	FVD ↓	FAD ↓	FVD ↓	FAD ↓	FVD ↓
Conditional (task-specific)	7.1	**63.6**	49.4	11.5	5.3	15.9	7.4	12.1	7.8	35.3
Per modality	7.0	84.4	**34.1**	**4.7**	6.2	213.6	4.5	92.1	5.6	106.1
Vanilla	7.1	**63.6**	53.3	8.1	8.1	226.8	6.1	140.8	7.4	121.1
MoNL (Ours)	**6.4**	77.6	40.2	5.3	**4.6**	**11.8**	**3.1**	**8.8**	**4.9**	**34.6**
Ablations										
Per time-segment	6.6	96.3	124.5	12.1	5.1	28.2	5.0	72.3	7.2	80.3
Per time-segment Per modality	7.0	84.5	52.5	5.9	5.4	22.9	4.8	61.2	5.7	55.3
`Pt/Pm/Ptm`	9.0	90.1	43.1	5.1	5.2	13.4	4.1	16.9	5.9	40.9

Effectiveness of MoNL - User Study

- Pairwise Mann-Whitney U tests were conducted with Bonferroni correction for multiple comparisons to assess statistical difference.
- Across all axes (AV-quality, AV-alignment, Person consistency), raters preferred samples generated from MoNL over that of Vanilla or Per-modality (Pm) approaches.

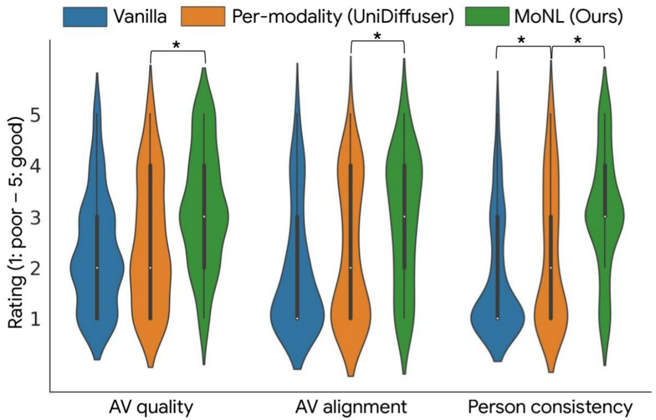


Figure 8: Comparative analysis across AVDiT models from the user study on AV quality, AV alignment and person consistency. The * indicates statistically significant pairwise difference at $p < 0.01$ after multiple correction.

Comparison with MM-Diffusion

- MM-Diffusion (MMD)[1]
  - The sole published work with a released model that tackles both audio and video generation within a single model.
  - While a direct comparison between U-Nets and our transformer architecture is inherently challenging due to their distinct design principles, we show that MoNL AVDiT surpasses this strong U-Net baseline.
- On Landscape dataset

Figure 2: Comparing conditional inference for AV-continuation for MM-Diffusion (left) and Ours (right) on Landscape dataset. Our approach excels at generating temporally consistent sequences.

- On AIST++ dataset

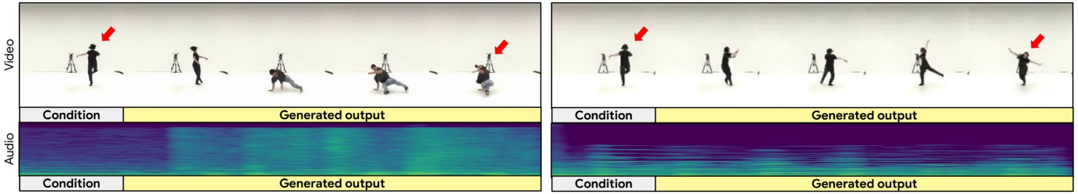


Figure 7: Unlike MM-Diffusion (left) where clothes and appearance is altered in the continuation (red arrow), our AVDiT with MoNL (right) maintains subject consistency in the AIST++ dataset.

[1] Ruan et al. "MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation." CVPR 2023.

Comparison with MM-Diffusion - Quantitative & User Study

- MoNL AVDiT outperformed MMD in terms of the FAD and FVD metrics across all tasks on the AIST++ and Landscape datasets.
- On the Landscape dataset, AV-align results demonstrate that our model achieves better alignment compared to MMD.
- Our MoNL AVDiT outperformed MMD in user studies overall.

Table 2: Quantitative comparison between our AVDiT with MoNL and MM-Diffusion (MMD).

		AIST++			Landscape			
Task	Method	FAD ↓	FVD ↓	KVD ↓	FAD ↓	FVD ↓	KVD ↓	AV align ↑
Reconstruction		0.90	11.72	0.96	0.76	16.41	-0.25	0.60
A2V	MMD	-	184.45	33.91	-	238.33	15.14	0.54
	Ours	-	**38.04**	**5.27**	-	**86.79**	**4.30**	**0.57**
V2A	MMD	13.30	-	-	13.60	-	-	0.50
	Ours	**1.11**	-	-	**0.78**	-	-	**0.51**

Table 3: User study of comparison between our model and MM-Diffusion (MMD) on the AIST++ dataset.

	Preference of ours over MMD		
	AV align	AV quality	Person consistency
AV-continue	0.69	0.71	0.93
A2V	0.77	0.61	0.75
V2A	0.61	0.49	0.60
Joint	0.74	0.72	0.81

Conclusion

- We propose a unified approach for multimodal diffusion using a mixture of noise levels (MoNL) for generating and manipulating sequences across modalities and time.
- This empowers a single model to handle diverse tasks like audio-video continuation, interpolation, and cross-modal generation.
- We show that an audiovisual latent diffusion transformer (AVDiT) trained with MoNL achieves state-of-the-art performance in audiovisual-sequence generation, providing new opportunities for expressive and controllable multimedia content creation.

Future works

- Our measures on subject consistency and synchrony of gestures and vocal tone were qualitative. Quantitative metrics to capture these joint spaces are part of our future work.
- We are also working on super-resolution systems to address visual quality and text conditioning to further optimize speech quality.

Thank you.

Poster: Session 5, Dec. 13(Fri) 11:00 AM
Audiovisual demos: avdit2024.github.io


Gwanghyun Kim*
Seoul National University
gwang.kim@snu.ac.kr


Alonso Martinez*


Yu-Chuan Su*


Brendan Jou
Google DeepMind


Jose Lezama
Google DeepMind


Agrim Gupta
Google DeepMind


Lijun Yu*


Lu Jiang*


Aren Jansen
Google DeepMind


Jacob Walker
Google DeepMind


Krishna Somandepalli§
Google DeepMind
ksoman@google.com

Equal contribution
*Work done while at Google.
§Corresponding author