Diffusion Autoencoders with Perceivers for Long, Irregular and Multimodal Astronomical Sequences

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Astronomical datasets continue to swell in volume and complexity, driven by large-scale surveys. Beyond the cataloging of galaxies and stars – photometrically by SDSS [York et al., 2000] and spectroscopically by DESI [Abareshi et al., 2022] – imaging surveys such as the Asteroid Terrestrial Last Alert System [ATLAS; Tonry et al., 2018], the Zwicky Transient Facility [ZTF; Bellm et al., 2018, Masci et al., 2018, Graham et al., 2019, Dekany et al., 2020], and the Young Supernovae Experiment [YSE; Jones et al., 2021, Aleo et al., 2023] now deliver nearly continuous coverage of the night sky in search of temporally evolving phenomena.

To extract meaningful structure from these massive datasets, self-supervised learning (SSL) has emerged as a powerful paradigm. SSL objectives are now widely used to pre-train large astrophysical networks [Zhang et al., 2024, Rizhko and Bloom, 2025], improving robustness to out-of-distribution data and enabling superior performance when fine-tuned on downstream tasks [Rizhko and Bloom, 2025].

A common SSL approach is reconstruction, in which the model learns to reproduce its input. Score-based generative models have achieved high-fidelity reconstructions in the image domain [Ho et al., 2020, Dhariwal and Nichol, 2021], but their latent spaces typically lack alignment with high-level semantic features. To address this limitation, Preechakul et al. [2022] introduced the diffusion autoencoder, which combines an autoencoder for dimensionality reduction with a diffusion model conditioned on the encoder latent space. This hybrid approach produces both semantically meaningful features and reconstructions of exceptional quality.

While diffusion autoencoders have shown promise for feature learning on image data, they are not directly applicable to astronomical time-series data. Time-domain astrophysics, in particular, involves long sequences of irregularly sampled, noisy photometric and spectroscopic data. These datasets demand custom architectures for scalable representation learning.

2 Background

Diffusion autoencoders. Originally proposed for images by Preechakul et al. [2022], diffusion autoencoders encode data with an encoder and reconstruct them using a conditional diffusion model. Since the encoding guides every denoising step, they can capture fine details more effectively than, for example, variational autoencoders [Kingma and Welling, 2013]. However, the original diffusion autoencoder relied on U-Net [Preechakul et al., 2022, Dhariwal and Nichol, 2021], which is better suited to regular modalities like images.

Perceiver. Perceiver and Perceiver-IO [Jaegle et al., 2021b,a] provide a general framework to (1) encode irregularly sampled sequences into a latent representation and (2) query outputs from this latent. This makes them a natural fit for integration with diffusion transformers [Dhariwal and Nichol, 2021], enabling scalable representation learning with diffusion autoencoders.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Related work. Autoencoding and dimensionality reduction have a long history in representation learning. Early models that remain widely used include variational autoencoders [VAEs, Kingma and Welling, 2013] and their variants, such as hierarchical models [Vahdat and Kautz, 2020] and discrete latent spaces [Van Den Oord et al., 2017, Razavi et al., 2019]. These models remain common in physics applications, though they often suffer from posterior collapse [Van Den Oord et al., 2017, Higgins et al., 2017] and generally show weaker generative ability than GANs [Goodfellow et al., 2020] or diffusion models [e.g., DDPM Ho et al., 2020].

Researchers have explored combining VAEs with diffusion models to improve generative quality, for example by learning a diffusion prior [Wehenkel and Louppe, 2021] or training diffusion models on VAE latent spaces [Kwon et al., 2022, Yan et al., 2021]. Beyond VAEs, masked autoencoders [MAEs, He et al., 2022] have recently gained attention as efficient learners for images and videos. MAEs reconstruct masked regions from the unmasked context, a strategy well-suited to modalities with strong local structure like images or audio, but less effective for data with long-range dependencies (researchers have also suggested that diffusion models can be interpreted as MAEs; Wei et al. 2023). Despite their impressive performance in image and audio domains, these methods often struggle to encode high-frequency structure in irregularly sampled long sequences, such as those obtained by astronomical spectrographs.

3 Diffusion autoencoder with perceiver

In this section, we introduce our **d**iffusion **a**uto**e**ncoder with **p**erceiver (daep[1]). A unimodal daep has three components: tokenizer, encoder, and diffusion decoder.

Tokenizers. We represent raw data as a sequence of tokens in the model dimension. We treat data as a collection of measurements at specific locations with accompanying metadata. Formally, we define $(\boldsymbol{v}, \boldsymbol{s}, \boldsymbol{m})$, where $\boldsymbol{v}$ are measurement values (e.g., flux of an astrophysical source), $\boldsymbol{s}$ is location information (e.g., wavelength, time, band), and $\boldsymbol{m}$ is observational metadata (e.g., instrument, observation time of spectra). We adapt the perceiver strategy [Jaegle et al., 2021b] by linearly projecting $\boldsymbol{v}$ using fixed sinusoidal embeddings for continuous parts of $\boldsymbol{s}$ (e.g., time) and categorical embeddings for discrete parts (e.g., color bands). We concatenate value and positional embeddings and project them to model dimension. We represent metadata as extra tokens appended to the sequence. For images, we add a small CNN at the beginning of the encoder to reduce the number of tokens. We show three example tokenizers in fig. 1.

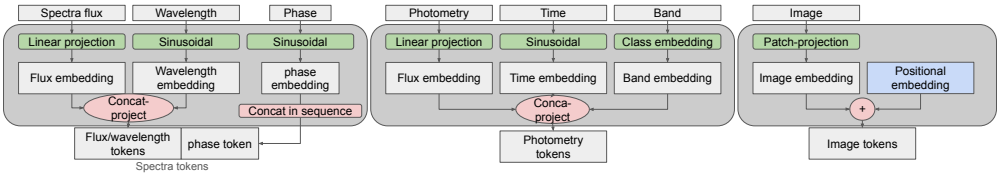

Figure 1: Tokenizers used in our empirical studies, from left to right: spectra (flux across wavelengths), light curves (brightness in different colors over time), and images.

Unimodal encoders. We use perceiver encoders [Jaegle et al., 2021b] to map token sequences into compact bottleneck representations. Tokens act as Keys and Values in cross attention, while bottleneck representations serve as Queries. Self-attention is applied only among bottleneck sequences. We repeat these perceiver blocks several times, optionally sharing weights. This design handles variable-length sequences with linear cost in sequence length, making it efficient for processing long and irregular data. Finally, we project bottleneck sequences from the model dimension to a fixed bottleneck dimension. We illustrate the encoder in fig. 2.

Perceiver IO–based diffusion decoder. Our decoder builds on diffusion transformers [Peebles and Xie, 2023], particularly cross-attention conditioning. We encode diffusion time with fixed sinusoidal embeddings passed through an MLP, as in Peebles and Xie [2023], and concatenate it with the

[1] Code available here.

conditioning representation for the score model. The score model, which predicts added noise, is a perceiver-IO: noisy data is tokenized, concatenated with conditioning tokens, and used as Keys and Values in cross attention. A latent sequence serves as Queries with self-attention, then acts as Keys and Values in a second cross-attention stage with positional information as Queries. We repeat these blocks, optionally sharing weights. The schematic is shown in fig. 2. While Jaegle et al. [2021a] recommend latent lengths of 128–512, this can be longer than data sequence for some tasks. In such cases, we use a single-stage perceiver decoder without a latent sequence, directly connecting noisy tokens to noise prediction through cross attention.

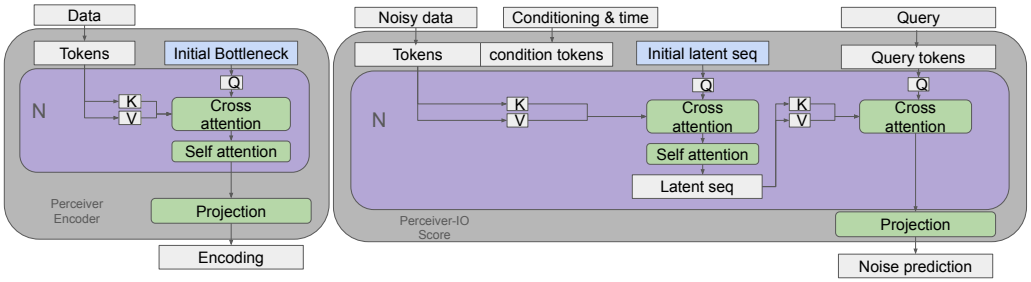

Figure 2: Schematic of the perceiver encoder and perceiver-IO score model used in daep.

Training and sampling. We train with the score-matching loss from DDPM [Ho et al., 2020] using 1,000 denoising steps. At inference, we adopt deterministic DDIM [Song et al., 2020] for faster sampling with 200 steps. Similar to Preechakul et al. [2022], our model is not a generative model since it requires the bottleneck representation. However, following Preechakul et al. [2022] and Wehenkel and Louppe [2021], we can train another DDIM to sample from the bottleneck distribution, enabling prior generation.

4 Unimodal experiments

High-resolution spectra of variable stars. We used data from v2.0 DR9 of the Large Sky Area Multi-Object Fiber Spectroscopic Telescope [LAMOST, Cui et al., 2012], specifically the dataset consolidated by Rizhko and Bloom [2025]. The dataset contains spectra of variable stars with on average $\sim$2,500 measurements and up to $\sim$4,000 per star. We trained our model and a benchmark β-VAE ($\beta = 0.1$) with the same perceiver encoder and decoder to encode spectra into a four-token sequence of dimension eight. Full architectural details are provided in appendix A.1. We show two enlarged test examples in fig. 3, with additional examples in fig. 6. Our model reconstructions captured finer spectral features compared to the VAE baseline.

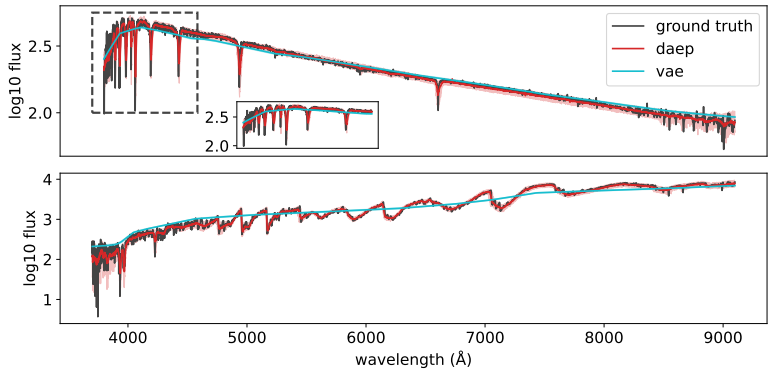

Figure 3: Two example reconstructions of long variable star spectra. Our daep model captured finer spectroscopic features, while the VAE mostly reproduced the overall continuum, likely due to posterior collapse.

Spectra and photometry of supernovae. We used data from the Zwicky Transient Facility Bright Transient Survey [ZTFBTS, Bellm et al., 2018]. In this survey, supernovae are measured in two bands—green (g) and red (r)—along with spectra, all sampled irregularly. We encoded the light curves into a two-token sequence of dimension two. Example reconstructions from our model are shown in fig. 4, with VAE results in fig. 7. Our method achieved more accurate reconstructions than the VAE baseline. Implementation details are provided in appendix A.2, with additional spectral results in appendix B.3.

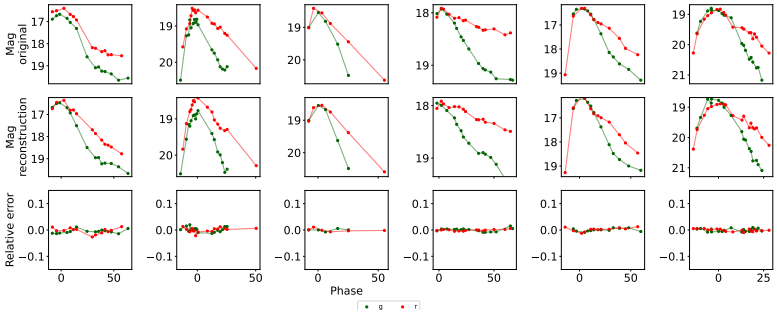

Figure 4: Light curve reconstruction from two latent tokens of dimension two using our daep model.

Images of galaxies. We used data from the Galaxy10 DECals dataset. Our method achieves moderately superior reconstructions compared to the VAE baseline, particularly for high-frequency features. Further details and results are provided in appendix A.3 and appendix B.4.

5 Towards multimodality

Modality mixing and training for multimodal data. To learn joint representations from multiple modalities, we used a late mixing strategy. We first encoded each modality with a perceiver encoder, added a learnable modality embedding, concatenated them along the sequence dimension, and then applied another perceiver encoder as a "mixer" to produce a single compact bottleneck sequence. Because the perceiver encoder does not require fixed-length input, we trained with modality dropping [Neverova et al., 2015, Liu et al., 2022] so the multimodal model can handle missing modalities. We always decoded all modalities using modality-specific diffusion decoders. A schematic of this architecture with two modalities is shown in fig. 5.

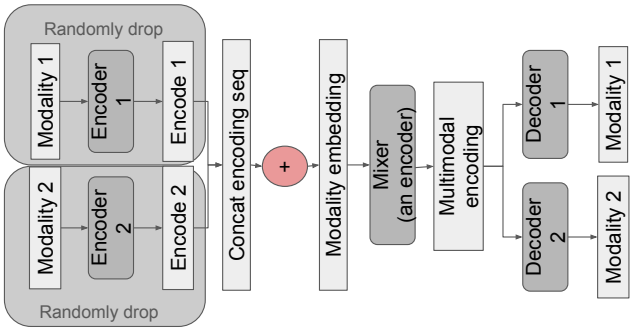

Figure 5: Late mixing and modality dropping for the multimodal daep model.

Simulated supernova spectra and photometry. We used data from simulated, idealized type Ia supernova light curves and spectra provided by Goldstein and Kasen [2018] and Shen and Gagliano [2025]. We trained a multimodal daep and evaluated it on cross-modality generation. Details are deferred to appendix A.4 and appendix B.5.

References

Behzad Abareshi, J Aguilar, S Ahlen, Shadab Alam, David M Alexander, R Alfarsy, L Allen, C Allende Prieto, O Alves, J Ameel, et al. Overview of the instrumentation for the dark energy spectroscopic instrument. *The Astronomical Journal*, 164(5):207, 2022.

P_D Aleo, K Malanchev, S Sharief, D_O Jones, G Narayan, R_J Foley, V_A Villar, C_R Angus, V_F Baldassare, M_J Bustamante-Rosell, et al. The young supernova experiment data release 1 (yse dr1): light curves and photometric classification of 1975 supernovae. *The Astrophysical Journal Supplement Series*, 266(1):9, 2023.

Eric C Bellm, Shrinivas R Kulkarni, Matthew J Graham, Richard Dekany, Roger M Smith, Reed Riddle, Frank J Masci, George Helou, Thomas A Prince, Scott M Adams, et al. The zwicky transient facility: system overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 131(995):018002, 2018.

Xiang-Qun Cui, Yong-Heng Zhao, Yao-Quan Chu, Guo-Ping Li, Qi Li, Li-Ping Zhang, Hong-Jun Su, Zheng-Qiu Yao, Ya-Nan Wang, Xiao-Zheng Xing, et al. The large sky area multi-object fiber spectroscopic telescope (lamost). *Research in Astronomy and Astrophysics*, 12(9):1197, 2012.

Richard Dekany, Roger M Smith, Reed Riddle, Michael Feeney, Michael Porter, David Hale, Jeffry Zolkower, Justin Belicki, Stephen Kaye, John Henning, et al. The zwicky transient facility: Observing system. *Publications of the Astronomical Society of the Pacific*, 132(1009):038001, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Daniel A Goldstein and Daniel Kasen. Evidence for sub-chandrasekhar mass type ia supernovae from an extensive survey of radiative transfer models. *The Astrophysical Journal Letters*, 852(2): L33, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Matthew J Graham, SR Kulkarni, Eric C Bellm, Scott M Adams, Cristina Barbarino, Nadejda Blagorodnova, Dennis Bodewits, Bryce Bolin, Patrick R Brady, S Bradley Cenko, et al. The zwicky transient facility: science objectives. *Publications of the Astronomical Society of the Pacific*, 131(1001):078001, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021a.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021b.

DO Jones, RJ Foley, G Narayan, Jens Hjorth, ME Huber, PD Aleo, KD Alexander, CR Angus, Katie Auchettl, VF Baldassare, et al. The young supernova experiment: survey goals, overview, and operations. *The Astrophysical Journal*, 908(2):143, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Han Liu, Yubo Fan, Hao Li, Jiacheng Wang, Dewei Hu, Can Cui, Ho Hin Lee, Huahong Zhang, and Ipek Oguz. Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 444–453. Springer, 2022.

Frank J Masci, Russ R Laher, Ben Rusholme, David L Shupe, Steven Groom, Jason Surace, Edward Jackson, Serge Monkewitz, Ron Beck, David Flynn, et al. The zwicky transient facility: Data processing, products, and archive. *Publications of the Astronomical Society of the Pacific*, 131 (995):018003, 2018.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2015.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Mariia Rizhko and Joshua S Bloom. Astrom3: A self-supervised multimodal model for astronomy. *The Astronomical Journal*, 170(1):28, 2025.

Yunyi Shen and Alexander T Gagliano. Variational diffusion transformers for conditional sampling of supernovae spectra. *arXiv preprint arXiv:2505.03063*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

J. L. Tonry, L. Denneau, A. N. Heinze, B. Stalder, K. W. Smith, S. J. Smartt, C. W. Stubbs, H. J. Weiland, and A. Rest. ATLAS: A High-cadence All-sky Survey System. *Publications of the Astronomical Society of the Pacific*, 130(988):064505, June 2018. doi: 10.1088/1538-3873/aabadf.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Antoine Wehenkel and Gilles Louppe. Diffusion priors in variational autoencoders. *arXiv preprint arXiv:2106.15671*, 2021.

Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16284–16294, 2023.

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

Donald G York, Jennifer Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

Gemma Zhang, Thomas Helfer, Alexander T Gagliano, Siddharth Mishra-Sharma, and V Ashley Villar. Maven: a multimodal foundation model for supernova science. *Machine Learning: Science and Technology*, 5(4):045069, 2024.

A Implementation details

A.1 LAMOST model details

Data preprocessing We enforced a 3-σ quality cut, i.e., only measurements exceeding 3 times the measurment error are kept for modeling. After all quality cuts, we have 17,063 training stars. We took arcsinh of flux before modeling and after generation we calculate the original flux. Flux and wavelength are standardized to a z-score using the mean and standard deviation of the whole training set after calculating arcsinh of flux.

Architectural details Training details We used same learning rate of $2.5e-4$ for both models and

model	bottleneck len	bottleneck dim	enc. layers	dec. layers	model dim	# heads	hidden seq len
daep	4	8	4	4	128	8	256
VAE	4	8	4	4	128	8	256

Table 1: Architectural choices used in LAMOST experiments.

trained for 2000 epochs and 200 epochs respectively for daep and VAE, both training loss converged. We set $\beta = 0.1$ for VAE.

A.2 ZTF model details

Light curve preprocessing We first enforced a 3-σ cut on measurements, then used a Gaussian process to find the peak time of red band as the 0 phase. We align time to be relative to the peak time. We only kept events that light curves cover the peak.

Spectra preprocessing We enforced a 3-σ quality cut for both spectra and lightcurve, i.e., only measurements exceeding 3 times the measurement error are kept for modeling. After all cuts we have 2,934 events left in training. We took the base-10 logarithm of the flux before modeling, and after generation we calculate the original flux. We also apply a median filter to filter out noise. Flux and wavelength are then standardized to a z-score using the mean and standard deviation of the whole training set.

Architectural details

model	bottleneck len	bottleneck dim	enc. layers	dec. layers	model dim	# heads
daep	4	4	4	4	128	4
VAE	4	4	4	4	128	4

Table 2: Architectural choices used in ZTF spectra experiments. We used a single stage decoder (skipping latent sequence) since the sequence is short.

model	bottleneck len	bottleneck dim	enc. layers	dec. layers	model dim	# heads
daep	2	2	4	4	128	4
VAE	2	2	4	4	128	4

Table 3: Architectural choices used in ZTF light curve experiments. We used a single stage decoder (skipping latent sequence) since the sequence is short.

Training details Different from LAMOST experiment, we augment our data by 5 folds and adding noise to flux measurement and randomly mask part of the measurements. We used same learning rate of $2.5e-4$ for both models and trained for 2000 epochs and 200 epochs respectively for daep and VAE, both training loss converged. We set $\beta = 0.1$ for VAE.

A.3 Galaxy10 model details

Preprocessing We normalize the pixel values assuming mean 0.5 and std 0.5 and reshape to size $64 \times 64 \times 3$.

Architectural details

model	bottleneck len	bottleneck dim	enc. layers	dec. layers	model dim	# heads	patch size
daep	8	8	4	4	128	4	4
VAE	8	8	4	4	128	4	4

Table 4: Architectural choices used in Galaxy10 experiments. We used a single stage decoder (skipping latent sequence) since the sequence is short.

Training details We augment our data by 3 folds with random flipping. We used same learning rate of $2.5e-4$ for both models and trained for 2000 epochs and 200 epochs respectively for daep and VAE, both training loss converged. We set $\beta = 0.5$ for VAE.

A.4 Multimodal spectra and photometry

Data preprocessing. We did not perform further processing beyond those used in Shen and Gagliano [2025].

Architectural details We have the first stage encoder for both light curve and spectra to have model dimension 256, 4 layers, 4 heads, and encode to 64 tokens. The mixer has 4 layers and 4 heads and model dimension 256, during encoder we allow the concatenated sequence to have self attention. We encode to a bottleneck sequence of 4 tokens of dimension 4 each.

Training details. In each batch we randomly dropped a modality with probability 0.2, but retain at least one modality. We trained with learning rate $2.5e-4$ and for 2000 epochs. The loss converged.

B Further experimental results

B.1 More high-resolution spectra

In fig. 6, we show additional spectra reconstructions using daep and VAE baselines. Our method consistently captures higher-frequency information details compared to the VAE baseline.

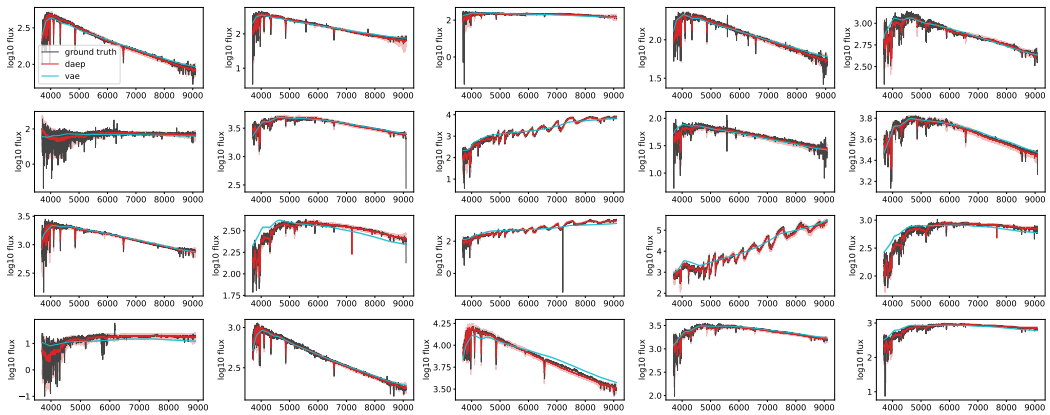

Figure 6: Reconstructions of additional LAMOST variable star spectra. Our method (red) captures more high-frequency absorption features than the VAE baseline with the same-sized bottleneck representation (blue).

B.2 ZTF light curves

In fig. 7, we show ZTF light curve reconstructions with the VAE baseline. Our method performs moderately better.

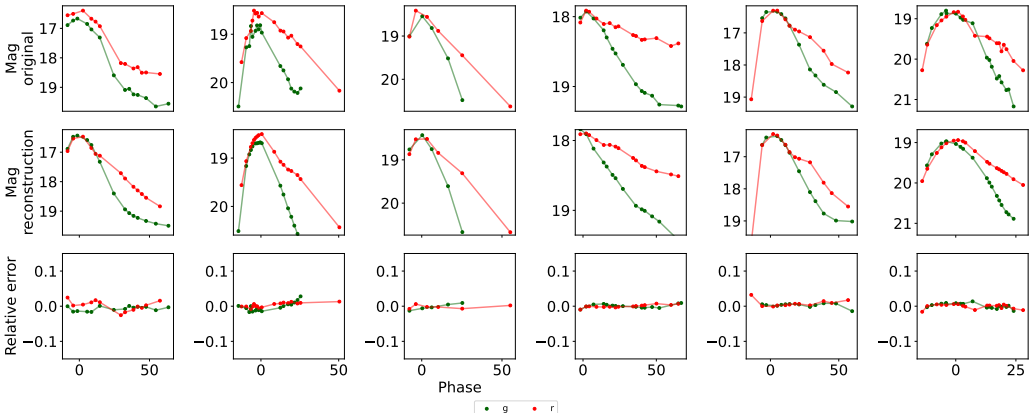

Figure 7: Reconstruction of ZTF light curves using a baseline VAE model (middle row) compared to the ground truth observations (top row). Residuals are given in the bottom row.

B.3 ZTF spectra

In figs. 8 and 9, we show spectra reconstructions on ZTF data. Our method captures finer details and produces better-covered posteriors than the VAE baseline.

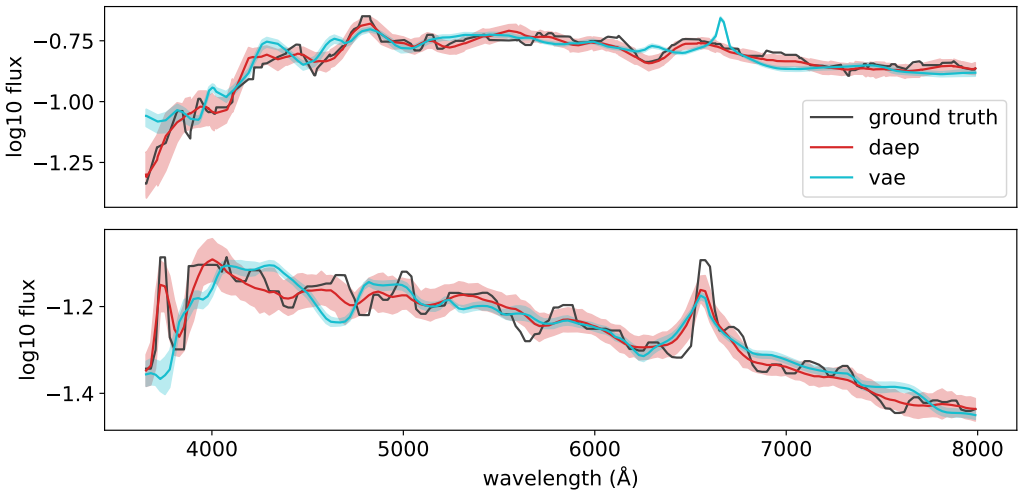

Figure 8: Zoomed-in ZTF spectra reconstructions with daep and VAE. Our method captures finer details and maintains better posterior coverage.

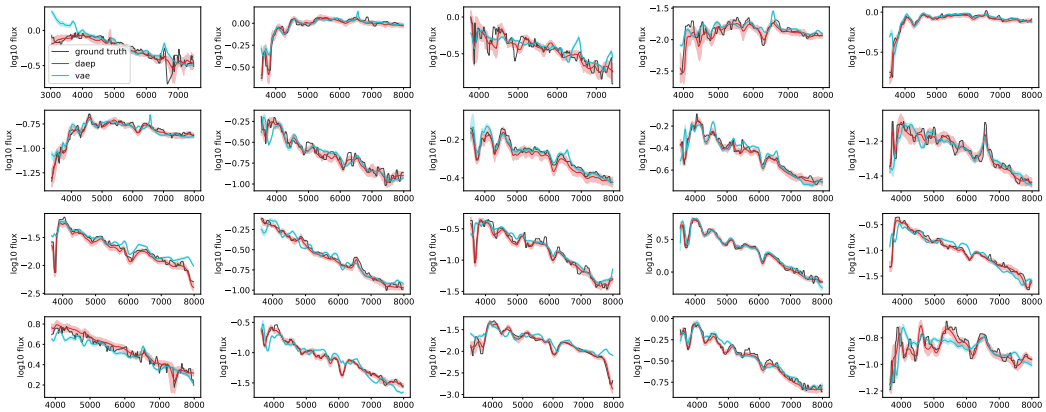

Figure 9: Additional ZTF spectra reconstructions with daep and VAE. Our method captures finer details and maintains better posterior coverage.

In fig. 10, we compare latent representations (after t-SNE) from daep and VAE, colored by event type. Interestingly, the daep latent space appears more continuous, while the VAE with $\beta = 0.1$ shows more holes.

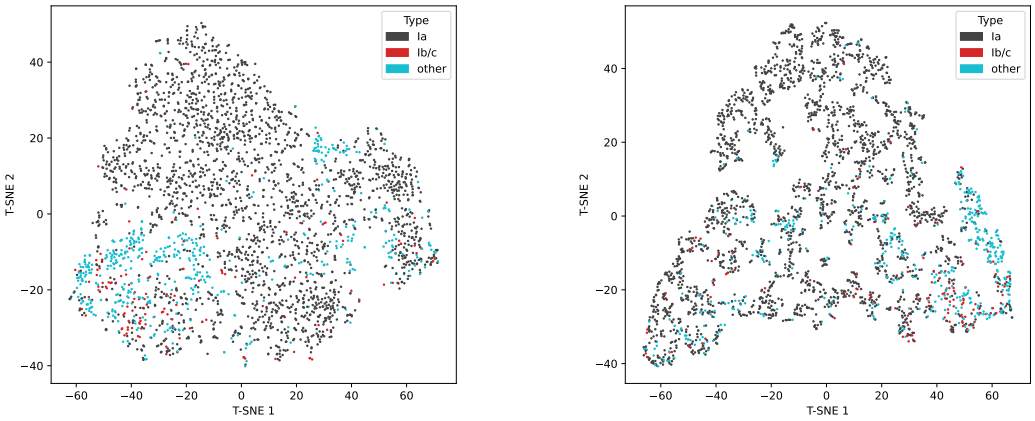

Figure 10: Latent representations of ZTF spectra from daep (left) and VAE (right). The latent space for daep appears more well-regularized compared to a VAE of comparable dimensionality.

B.4 Galaxy10

In figs. 11 and 12, we show reconstructions on the Galaxy10 dataset. Our method captures slightly finer-scale structures such as spiral arms compared to the VAE baseline.

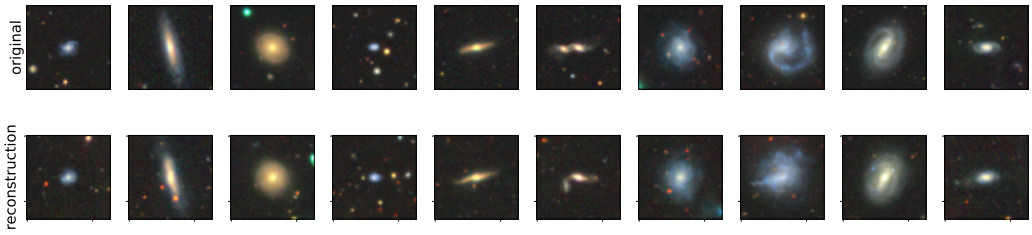

Figure 11: Reconstructions on Galaxy10 test data using daep.

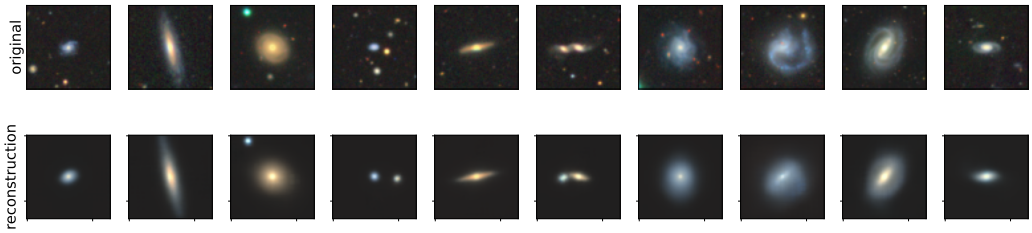

Figure 12: Reconstructions on Galaxy10 test data using a VAE. Reconstructions are heavily smoothed and retain fewer high-frequency structure than daep reconstructions.

B.5 Multimodal spectra and LSST photometry of supernovae

We demonstrate cross-modality generation using the learned encoder. In fig. 13, we show inference from light curves to spectra, while in fig. 14, we show inference from spectra to light curves. Our model performs cross-modality inference directly from the bottleneck representation.

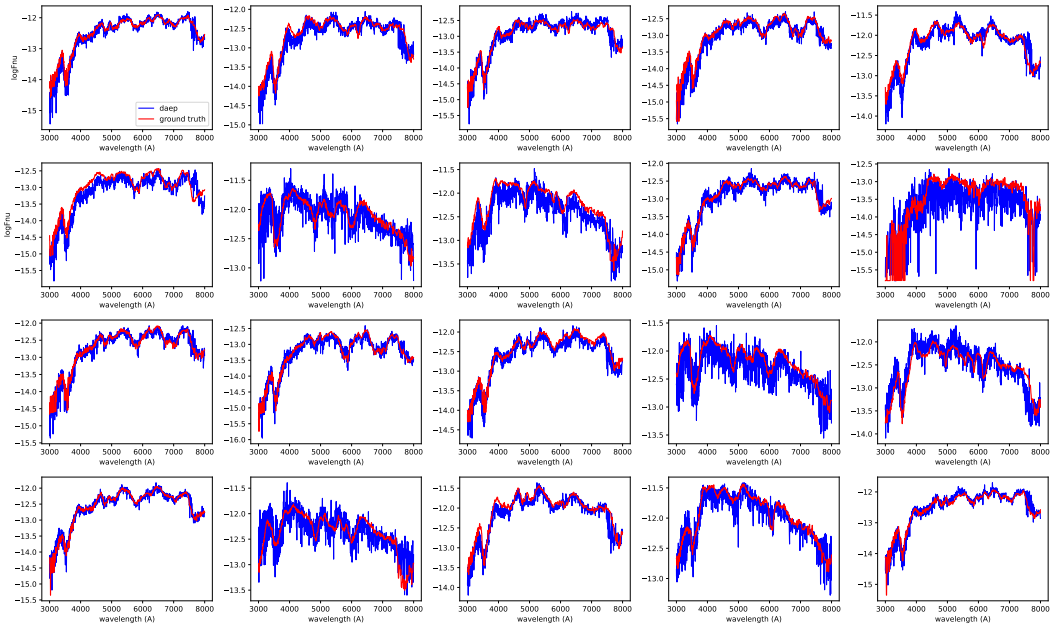

Figure 13: Cross-modality inference from light curves to spectra in the Goldstein simulation.

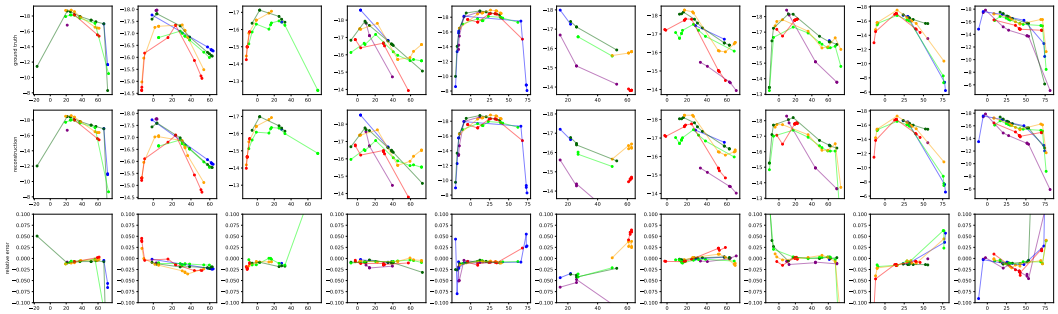

Figure 14: Cross-modality inference from spectra to light curves in the Goldstein simulation.

C Discussion

We have presented a perceiver-based diffusion autoencoder that scalably learns bottleneck representations from irregularly sampled sequences. Its latent spaces appear more regular than those of a β-VAE baseline (Figure 10), while achieving better reconstruction on multiple time-domain datasets. Further, the results for cross-modality generation show that the daep architecture can be valuable for the development of multi-modal models in astrophysics. In the future, we plan to further evaluate the learned representations on downstream tasks such as classification and clustering.