VARIATIONAL DIFFUSION TRANSFORMERS FOR CONDITIONAL SAMPLING OF SUPERNOVAE SPECTRA

Yunyi Shen
EECS, MIT
Cambridge, MA, USA
yshen99@mit.edu

Alexander T. Gagliano
IAIFI, MIT
CfA, Harvard University
Cambridge, MA, USA
gaglian2@mit.edu

ABSTRACT

Type Ia Supernovae (SNe Ia) have become the most precise distance indicators in astrophysics due to their incredible observational homogeneity. Increasing discovery rates, however, have revealed multiple sub-populations with spectroscopic properties that are both diverse and difficult to interpret using existing physical models. These peculiar events are hard to identify from sparsely sampled observations and can introduce systematics in cosmological analyses if not flagged early; they are also of broader importance for building a cohesive understanding of thermonuclear explosions. In this work, we introduce `DiTSNe-Ia`, a variational diffusion-based generative model conditioned on light curve observations and trained to reproduce the observed spectral diversity of SNe Ia. In experiments with realistic light curves and spectra from radiative transfer simulations, `DiTSNe-Ia` achieves significantly more accurate reconstructions than the widely used SALT3 templates across a broad range of observation phases (from 10 days before peak light to 30 days after it). `DiTSNe-Ia` yields a mean squared error of 0.108 across all phases—five times lower than SALT3's 0.508—and an after-peak error of just 0.0191, an order of magnitude smaller than SALT3's 0.305. Additionally, our model produces well-calibrated credible intervals with near-nominal coverage, particularly at post-peak phases. `DiTSNe-Ia` is a powerful tool for rapidly inferring the spectral properties of SNe Ia and other transient astrophysical phenomena for which a physical description does not yet exist.

1 INTRODUCTION

Spectral diversity of Type Ia Supernovae. Type Ia supernovae (SNe Ia) are the thermonuclear explosions of white dwarfs in binary systems as they approach the Chandrasekhar limit (Hoyle & Fowler, 1960; Liu et al., 2023). Since their use in confirming the accelerated expansion of the universe (Riess et al., 1998; Perlmutter et al., 1999), SNe Ia have been widely adopted as cosmological distance indicators. SNe Ia are *standardizable* candles: they exhibit correlations in their observational properties that can be corrected to infer their distance (such as the "Phillips relation" between an explosion's maximum luminosity and its dimming rate; Pskovskii, 1977; Phillips, 1993; Phillips et al., 1999).

Growing sample sizes of SNe Ia have complicated the story. Wide-field imaging surveys such as the Zwicky Transient Facility (Bellm et al., 2018) have discovered otherwise-typical SNe Ia with additional observational oddities (e.g., early and late-time re-brightening episodes), along with multiple sub-populations that defy the expected observational correlations. These include both red and underluminous SNe Ia (named "SNe Ia 1991bg-like" after their archetypal event; as well as SNe Iax, alternatively known as "SNe 2002cx-like"; Filippenko et al., 1992a); and overluminous SNe Ia (SNe Ia 1991T-like events; Filippenko et al., 1992b). Spectra of these SNe, which measure their wavelength-specific emission, have revealed important differences between the temperatures, velocities, and compositions of these more peculiar explosions. Nonetheless, our growing understanding of this observational diversity has only raised more questions about the physics of these explosions (Soker, 2019; Liu et al., 2023).

Legacy Survey of Space and Time. The upcoming Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST, Ivezić et al., 2019), which will image the full Southern Sky every 3-4 days for ten years beginning in 2025, is expected to discover $\sim$1 M SNe each year (Ivezić et al., 2019). Half of them will be SNe Ia. While discovery rates of these explosions will break exponential scaling as a result of the survey, our spectroscopic resources will only be able to observe the smallest subset (at most $\sim$0.1%). Rapid and unbiased predictions of the spectroscopic behavior of discovered SNe Ia from measurements of their time-evolving brightness (light curves) will allow astronomers to rapidly prioritize the most scientifically valuable events for spectroscopic observation. These data, in turn, will improve our ability both to precisely standardize the most homogeneous SNe Ia for cosmology and to launch exploratory studies into the explosion properties of the most diverse.

Goal: spectra inference and generation. Motivated by the need for low-latency physical inference in SN Ia science, we consider the task of inferring spectra from light curves, with both obtained at optical wavelengths. These measurements are correlated slices of an event's spectral energy distribution (SED): light curves measure the evolution of an explosion's flux at specific wavelengths over time, while spectra measure flux changes in wavelength at a single moment in time (Zhang et al., 2024). As a result, one can infer an explosion's spectral properties in time by sampling from a posterior of SN Ia spectra conditioned on its light curve and time. This casts the inference problem as a conditional generation task.

2 A MODEL FOR CONDITIONAL SPECTRA GENERATION

Variational diffusion transformer for spectra. We used a variational diffusion model for spectral generation (Kingma et al., 2021). We used a transformer (DiT) architecture for the score model (Peebles & Xie, 2023) because of the flexibility of modeling sequences that are not from a regular grid such as light curves and spectra. Each DiT block consists of one attention block and one cross-attention block separated by one layer normalization block. Each (cross)attention block also consists of multi-headed (cross)attention, layer normalization, and a Multi-Layer Perceptron (MLP), as well as skip connections (Figure 1).

An ongoing challenge in training neural networks from astronomical observations is that the sequence may not be regularly-spaced over time and wavelength. Our model addresses this challenge by embedding wavelength/time ("positional") information first with a set of fixed sinusoidal functions, where we employ exponentially-changing frequencies as in the original transformer model (Vaswani, 2017) taking input of (normalized) wavelengths/times.

We start with a sinusoidal embedding of dimension $2D_{\text{sin}}$; P denotes the sinusoidal encoding of positional data (wavelength and time), with $s \in \mathcal{S}$ for dimension $2i, 2i+1$ and $i = 0, \ldots, D_{\text{sin}}$. The final positional embedding $\boldsymbol{E}$ is then given by

$$\begin{aligned} &P_{2i}(s) = \sin\left(\frac{s}{10,000^{2i/D_{\text{sin}}}}\right) \quad P_{2i+1}(s) = \cos\left(\frac{s}{10,000^{2i/D_{\text{sin}}}}\right) \\ &\boldsymbol{E}(s) = MLP(\boldsymbol{P}(s)) \end{aligned} \tag{1}$$

We then pass the encoding of size $[L_{spec}, 2D_{\text{sin}}]$, with L_{spec} being the length of spectra, through an MLP to reshape it into the model dimension D. This is a similar treatment as diffusion time in Peebles & Xie (2023).

For our spectral observations, we embed wavelength data with the above treatment, add it to a linear embedding of flux, and project to the model dimension. This will yield an internal representation of size $[L_{spec}, D]$ with D being the dimension of the model and L_{spec} being the length of the spectrum. We then pass it through diffusion transformer blocks to predict the added noise. We discuss our treatment of light curve observations below.

Conditioning and photometric embedding. In our conditional generation framework, there are three datasets to condition on: the diffusion time of the model, the phase of the observation (the time relative to the supernova's maximum light, in days), and the light curve observations themselves. We adopt a cross-attention strategy in our model due to the correlated information between these data and the complexity of the unstructured light curve sequences. Diffusion time and phase were first passed through the fixed sinusoidal functions (Figure 1) and then projected to the model dimension

using an MLP with learnable weights. Each of them yield a representation of size $[1, D]$ for model dimension D.

Light curve data were embedded similar to spectra: we first embedded phase using fixed sinusoidal functions, passed the output to a learnable MLP, and then added these to a linear embedding of the brightness measurements in magnitudes. To the combined vector, we also add an embedding of the transmission filter of the observation. These transmission filters were treated as categorical variables and the embeddings are simple look-up tables. This filter information encodes the range of wavelengths at which the brightness of the supernova was measured. This will yield a sequence embedding for the light curve of size $[L_{LC}, D]$, with L_{LC} being the sequence length of the light curves and D being the common model dimension size.

Representations of light curve, phase and diffusion time are then concatenated in the sequence dimension to get a conditional representation of size $[L_{LC} + 2, D]$.

We show an overview of the `DiTSNe-Ia` architecture in Figure 1. The model has a dimension $D = 256$. The MLP used in each DiT block consists of a single hidden layer with dimension 512. In both multi-headed attention and cross-attention, we use 4 heads and 6 layers of DiT blocks.

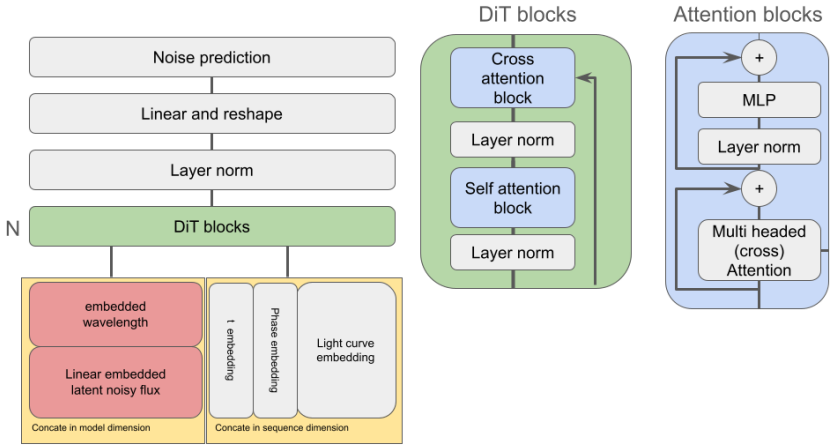

Figure 1: The architecture of our diffusion transformer for generation of SN Ia spectra conditioned on light curve data. Positional embeddings are used to preserve correlations from data irregularly sampled in both time and wavelength.

Loss and training. We train our model to minimize the variational lower bound (VLB) on the marginal likelihood (as in Equation 11 of Kingma et al., 2021).

3 Experiments

Data, preprocessing, and training scheme. We use the model grid of Goldstein & Kasen (2018) to train and validate `DiTSNe-Ia`. Goldstein & Kasen (2018) ran 4,500 simulations of SNe Ia with the radiation transport code `Sedona` (Kasen, 2006) to produce a full SED surface in temporal bins of 1 day and wavelength bins of 30 Å. The simulated SNe Ia span a broad range of ejecta masses ($M_{ej} \in [0.7\ M_\odot, 2.5\ M_\odot]$), velocities $\left(v_K \in [8000\ \mathrm{km\ s^{-1}}, 15000\ \mathrm{km\ s^{-1}}]\right)$, masses of the exploding star ($m \in [1\ M_\odot, 2.5\ M_\odot]$) (where the traditional Chandrasekhar limit corresponding to a thermonuclear explosion is $\approx 1.4\ M_\odot$); and fractions of the ejecta composed of both radioactive ^{56}Ni ($f_{\mathrm{Ni}} \in [0.1, 0.8]$) and C/O ($f_{\mathrm{CO}} \in [0.00, 0.07]$). We split the sample into 80% training, 20% testing. We then generated LSST-like light curves by integrating the SED surfaces through the published transmission curves of the LSST filters from the SVO Filter Profile Service[1]. We model the light curve cadence using the baseline v3.3 simulations of LSST's Wide-Fast-Deep survey and assume a random distribution of events in space across the survey footprint and in time since survey start. We have assumed that all SNe occur at a distance of 10 pc; in future work, we plan to adopt

[1] `http://svo2.cab.inta-csic.es/`

a SN Ia volumetric rate to evaluate the robustness of the model on light curves with realistic noise properties.

We impose quality cuts on our generated light curves, requiring at least 10 total measurements across all filters and with observations in any band before and after the explosion's peak brightness. For our simulated spectra, we calculate the flux directly from the SED surface in 10-day windows from 10 days before peak light to 30 days following peak light; we apply a median filter with bandwidth 3 to these data to mimic the finite resolution of a spectrograph. We then train the model using the log of the flux values in units of $\mathrm{erg\ s^{-1}\ cm^{-2}\ Hz^{-1}}$.

We standardized all flux, wavelength, phase, and magnitude values to z-scores for training, and convert them back to physical units at inference time. We also pad the light curves to 10 measurements in each band, resulting in uniform light curve arrays of length 60.

We train `DiTSNe-Ia` in `jax` (Bradbury et al., 2018) using the AdamW optimizer (Loshchilov et al., 2017). We used a warm-up cosine decay schedule as the learning rate with 500 warm-up steps and 3000 decay steps, a maximum learning rate of 3×10^{-3}, an initial learning rate of 10^{-5}, and a batch size of 64.

We tested our method and the baseline model on 300 SNe from the test set at five phases (-10 to 30 days relative to peak light in 10-day bins). For each light curve and phase combination, we took 100 samples of each conditional generation, used averages as point estimates, and calculate equal-tail credible intervals of each prediction (CI).

Test metric We used three metrics to test the proposed method, one for prediction accuracy and two for uncertainty quantification

- **Residual** of point estimate: the difference between point estimate of the model and ground truth spectra at different days after peak. We also quantify the accuracy using the mean squared error (MSE).
- **Coverage**: the empirical probability that the posterior CI contains the ground truth value as a function of wavelength.
- **CI width**: the difference between the upper quantile and lower quantile of the CI for a given level.

Baseline. We compare our prediction to SALT3 (Kenworthy et al., 2021; Guy et al., 2007), a popular template-matching technique for SN Ia spectra. The SALT3 model consists of linear combinations of SN Ia SED components, whose contributions are determined by parameters (x_0, x_1, and c) fit to light curves in multiple filters. We used the implementation of the SALT3 surfaces in the python package `sncosmo` (Barbary et al., 2016) and fit the surface to our re-sampled light curves with an MCMC to quantify fit uncertainties.

Results. We provide a summary of our results in Table 1, and in Figure 2. We find that `DiTSNe-Ia` achieves better predictive performance across the test set while having a lower average bias than the SALT3 method. `DiTSNe-Ia` yields an overall mean squared error of 0.108—five times lower than SALT3's 0.508—and an after-peak error of just 0.0191, an order of magnitude smaller than SALT3's 0.305. Uncertainty quantification achieves coverage near nominal values in predictions greater than 10 days after peak light with an overall coverage of 0.845 for 90% interval (see also last row in Figure 2); although the predictions at earlier phases (especially 10 days before peak) are under-covered. This is unsurprising given the limited photometric information available at early times. SALT3, in contrast, has small CIs but is severely under-covered.

Figure 3 shows the spectra posteriors for two randomly-selected SNe Ia in the test set as a function of phase. Qualitatively, we find that our method not only recovers the broad shape of each spectrum (which encodes information about the temperature and explosion energy of the SN); but also finer-scale features (which encode the explosion's composition and velocity structure) more precisely than the baseline model. We plan to quantify the constraints on these physical properties in a future study.

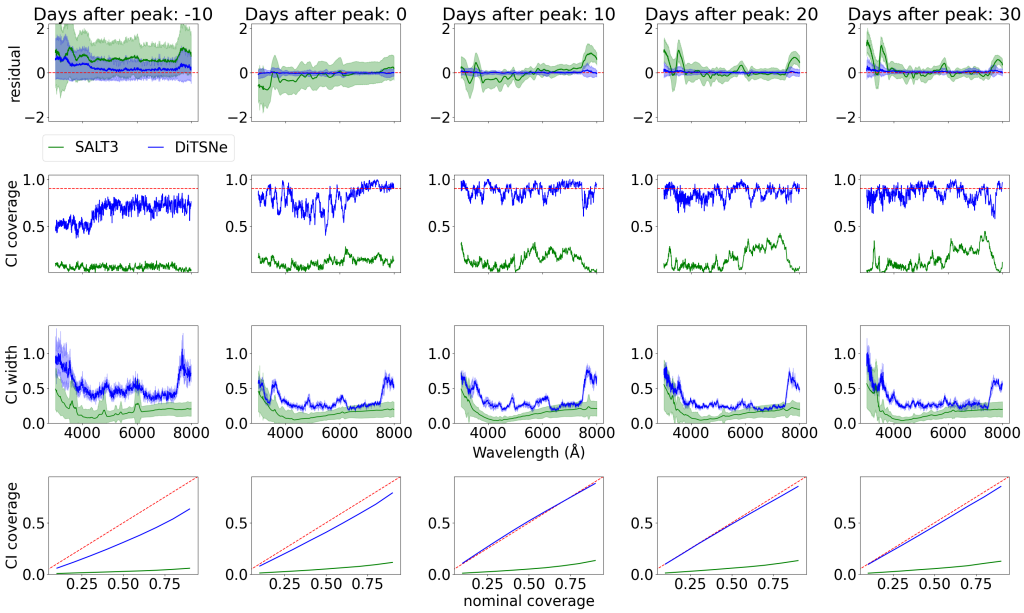

Figure 2: `DiTSNe` Prediction - Ground Truth and associated 1σ uncertainties (First Row) for spectra from the test set predicted over five different phases and conditioned on realistic LSST light curves. The prediction of the baseline SALT3 model is shown in green. Our method achieves a lower mean error rate than SALT3 at all phases. We also show coverage plots (Second Row) from the model, finding values near nominal starting +10 days after peak light; as well as the width of the width of the 90% CI from `DiTSNe` predictions (Third Row) and its realized coverage as a function of nominal coverage (Fourth Row).

Metric	Method	-10 days	0 days	10 days	20 days	30 days	Mean	After Peak
MSE	SALT3	1.49	0.467	0.232	0.247	0.279	0.508	0.305
	DiTSNe-Ia	**0.465**	**0.0183**	**0.015**	**0.0164**	**0.0267**	**0.108**	**0.0191**
CI Coverage	SALT3	0.0639	0.115	0.134	0.137	0.131	0.116	0.129
	DiTSNe-Ia	**0.634**	**0.79**	**0.883**	**0.854**	**0.853**	**0.803**	**0.845**
CI Width	SALT3	**0.187**	**0.154**	**0.157**	**0.165**	**0.174**	**0.165**	**0.161**
	DiTSNe-Ia	0.548	0.332	0.353	0.336	0.369	0.387	0.347

Table 1: Average MSE, CI coverage, and CI widths over all SNe in the test set as a function of phase (number of days relative to the peak brightness of each SN). Our method shows better performance than the SALT3 baseline but can be under-covered. Coverage and widths are shown for the 90% CI of predicted `DiTSNE-Ia` spectra. SALT3, in contrast, has shorter CIs but is severely under-covered.

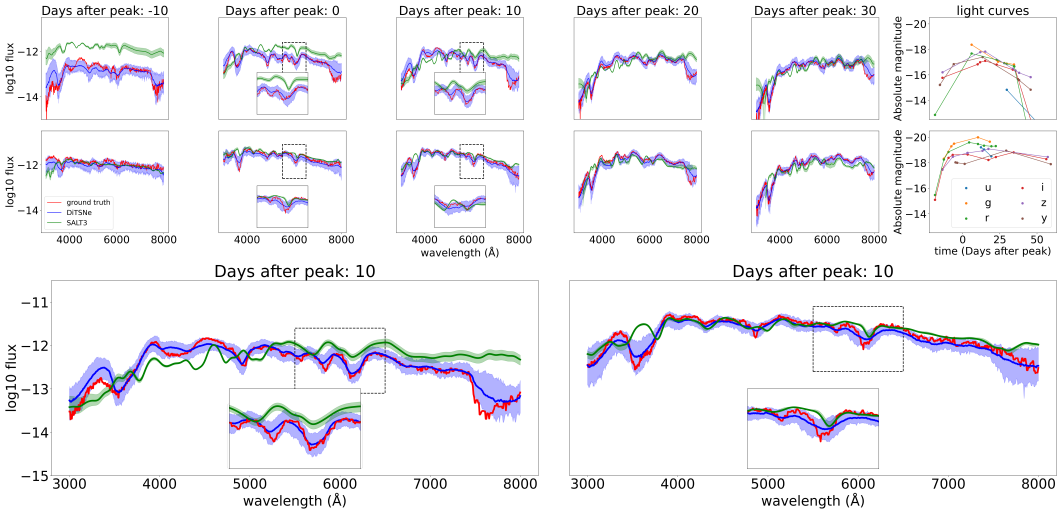

Figure 3: Upper: Spectra and light curves for two randomly chosen SNe Ia from the test set, with zoom-in of finer-scale features. The flux was calculated in units of erg $s^{-1}cm^{-2}$ Hz^{-1} and light curves are in absolute magnitude. Lower: Zoom-in view of each model's predictions 10 days after peak. Our method can recover finer-scale spectral features associated with the physical properties of the explosion that are not well-captured by SALT3.

4 Discussion

We have presented `DiTSNe-Ia`, a state-of-the-art generative model for inferring SN Ia spectra at multiple phases during an explosion using light curve information under idealized observing conditions. The method performs well when applied to simulated SNe Ia from Goldstein & Kasen (2018), demonstrating the potential of diffusion-based generative models in SN science.

Our follow-up work will focus on connecting the model's performance, measured by MSE and CI coverage, to physical parameters in the simulation to identify systematic patterns. Several challenges remain. For instance, although our approach is motivated by observational constraints, the current training data are still based on simulations, potentially producing a model that underestimates the true diversity observed in real data. Additionally, we assume that the training light curves and spectra are corrected for observational effects such as redshift. However, in practice, redshift is not always fully known and carries its own uncertainties.

Another area for improvement is the representation of light curves, particularly in incorporating translation invariance—a property useful when the exact time of SN peak is unknown. Relative positional encoding (Shaw et al., 2018) does not directly generalize to cross-attention mechanisms, as it requires both sequences to exist in the same space with well-defined positional distances, which is not the case for light curves and spectra. A promising direction may be to apply convolutional layers before the attention mechanism to extract translation-invariant features, a strategy commonly used in the computer vision community (Wu et al., 2021; Xiao et al., 2021). We will investigate these improvements to the model in future work.

References

Kyle Barbary, Tom Barclay, Rahul Biswas, Matt Craig, Ulrich Feindt, Brian Friesen, Danny Goldstein, Saurabh Jha, Steve Rodney, Caroline Sofiatti, et al. Sncosmo: Python library for supernova cosmology. *Astrophysics Source Code Library*, pp. ascl–1611, 2016.

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.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Alexei V Filippenko, Michael W Richmond, David Branch, Martin Gaskell, William Herbst, Charles H Ford, Richard R Treffers, Thomas Matheson, Luis C Ho, Arjun Dey, et al. The subluminous, spectroscopically peculiar type ia supernova 1991bg in the elliptical galaxy ngc 4374. *Astronomical Journal (ISSN 0004-6256), vol. 104, no. 4, p. 1543-1556, 1684.*, 104:1543–1556, 1992a.

Alexei V Filippenko, Michael W Richmond, Thomas Matheson, Joseph C Shields, E Margaret Burbidge, Ross D Cohen, Mark Dickinson, Matthew A Malkan, Brant Nelson, Jochen Pietz, et al. The peculiar type ia sn 1991t-detonation of a white dwarf? *Astrophysical Journal, Part 2-Letters (ISSN 0004-637X), vol. 384, Jan. 1, 1992, p. L15-L18.*, 384:L15–L18, 1992b.

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.

Julien Guy, P Astier, S Baumont, D Hardin, R Pain, N Regnault, S Basa, RG Carlberg, A Conley, S Fabbro, et al. SALT2: using distant supernovae to improve the use of type ia supernovae as distance indicators. *Astronomy & Astrophysics*, 466(1):11–21, 2007.

Fred Hoyle and William A Fowler. Nucleosynthesis in supernovae. *Astrophysical Journal, vol. 132, p. 565*, 132:565, 1960.

Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019.

Daniel Kasen. Secondary maximum in the near-infrared light curves of type ia supernovae. *The Astrophysical Journal*, 649(2):939, 2006.

WD Kenworthy, DO Jones, M Dai, R Kessler, D Scolnic, D Brout, MR Siebert, JDR Pierel, KG Dettman, G Dimitriadis, et al. SALT3: An improved type Ia supernova model for measuring cosmic distances. *The Astrophysical Journal*, 923(2):265, 2021.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Zheng-Wei Liu, Friedrich K Röpke, and Zhanwen Han. Type ia supernova explosions in binary systems: A review. *Research in Astronomy and Astrophysics*, 23(8):082001, 2023.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Saul Perlmutter, Goldhaber Aldering, Gerson Goldhaber, Richard A Knop, Peter Nugent, Patricia G Castro, Susana Deustua, Sebastien Fabbro, Ariel Goobar, Donald E Groom, et al. Measurements of ω and λ from 42 high-redshift supernovae. *The Astrophysical Journal*, 517(2):565, 1999.

Mark M Phillips. The absolute magnitudes of type ia supernovae. *Astrophysical Journal, Part 2-Letters (ISSN 0004-637X), vol. 413, no. 2, p. L105-L108.*, 413:L105–L108, 1993.

MM Phillips, Paulina Lira, Nicholas B Suntzeff, RA Schommer, Mario Hamuy, and José Maza. The reddening-free decline rate versus luminosity relationship for type ia supernovae. *The Astronomical Journal*, 118(4):1766, 1999.

Iu P Pskovskii. Light curves, color curves, and expansion velocity of type i supernovae as functions of the rate of brightness decline. *Soviet Astronomy, vol. 21, Nov.-Dec. 1977, p. 675-682. Translation. Astronomicheskii Zhurnal, vol. 54, Nov.-Dec. 1977, p. 1188-1201.*, 21:675–682, 1977.

Adam G Riess, Alexei V Filippenko, Peter Challis, Alejandro Clocchiatti, Alan Diercks, Peter M Garnavich, Ron L Gilliland, Craig J Hogan, Saurabh Jha, Robert P Kirshner, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The astronomical journal*, 116(3):1009, 1998.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Noam Soker. Supernovae ia in 2019 (review): A rising demand for spherical explosions. *New Astronomy Reviews*, 87:101535, 2019.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22–31, 2021.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021.

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.