DINOspec: Efficient Multimodal Alignment of Vision and Spectral Foundation Models for Astronomy

Erica Lastufka
University of Geneva
Geneva, Switzerland 1211
erica.lastufka@unige.ch

Mariia Drozdova
University of Geneva
Geneva, Switzerland 1211
mariia.drozdova@unige.ch

Daniel Schaerer
University of Geneva
Geneva, Switzerland 1211
daniel.schaerer@unige.ch

Svyatoslav Voloshynovskiy
University of Geneva
Geneva, Switzerland 1211
Svyatoslav.Voloshynovskyy@unige.ch

Abstract

Astronomical observations provide multimodal views of physical systems, with images and spectra capturing complementary properties of celestial objects. Scientific foundation models can learn powerful representations from these observations, but representations learned by separate models remain difficult to combine. We investigate whether laten space representations learned by separate vision and spectral models can be aligned without retraining their encoders. We introduce DINOspec, a multimodal framework that aligns a frozen DINOv3 image encoder with a pre-trained AION-1 spectral encoder using lightweight adapters and contrastive learning on 20,472 paired images and spectra of astronomical objects. DINOspec improves galaxy morphology classification (F1: 0.72→0.78) and spectral type classification (F1: 0.70→0.74) while training at few as 21M parameters. Redshift prediction shows no consistent improvement over the AION-1 baseline of $R^2 \approx 0.9$, demonstrating that potential benefits of alignment depends on the configuration and downstream task. These results provide an empirical study of composing astronomical representations from independently pretrained foundation models through lightweight alignment modules.

1 Introduction

Observational astrophysics provides a natural setting for studying multimodal representation learning, as different instruments measure complementary aspects of the same physical systems. Multi-band imaging, spectroscopy, photometry, and source catalogs each contain different information about astronomical objects, but combining these observations requires representations compatible across modalities. Recent astronomical foundation models have begun addressing this challenge. AstroCLIP [Parker et al., 2024] aligns optical images and spectra, AstroM3 [Rizhko and Bloom, 2025] extends contrastive learning to photometry, spectra, and metadata, and AION-1 [Parker et al., 2025] integrates five heterogeneous datasets using modality-specific tokenizers and cross-modal transformer layers.

Preprint.

A central challenge for scientific foundation models is that representations learned from different instruments or datasets are difficult to reuse and combine. Astronomical observations differ substantially in wavelength coverage, spatial resolution, and noise properties, producing domain shifts between instruments or even surveys. Existing multimodal architectures typically rely on jointly trained modality-specific components, particularly tokenizers optimized for the statistics of the training data. Even unimodal benchmarks show that visual encoders of various architectures perform differently across data regimes [Lastufka et al., 2025]. Consequently, incorporating new data sources into existing multimodal models would require retraining substantial portions of it rather than reusing existing physical representations.

In this work, we investigate whether independently learned scientific representations can be aligned through lightweight trainable modules. This separates representation learning from cross-modal alignment, allowing use of existing foundation models without modifying their learned representations. We build on the Locked-image Tuning (LiT) strategy introduced by DINO.txt [Jose et al., 2024], which demonstrated that frozen visual representations can be adapted to new semantic spaces using alignment layers. Our method, DINOspec, combines a frozen DINOv3 [Siméoni et al., 2025] visual encoder with a spectral encoder operating on tokens from the pre-trained AION-1 tokenizer. DINOspec does not introduce a new contrastive alignment objective; instead, it investigates the empirical behavior of lightweight alignment between independently pretrained astronomical image and spectral representations. We evaluate how adapter initialization, capacity, and training strategy affect downstream morphology classification, spectral classification, and redshift regression, providing evidence of when alignment improves downstream utility and when unimodal information already dominates the downstream task.

2 Data

Dataset name	Origin	Train	Test	Labels	Task
hsc	HSC *grz* images	86614			vision pre-training
sdss_desi	SDSS and DESI spectra	107223	102407		spectra pre-training
desi_xmatchGRZ	cross-matched HSC and DESI	16378	4094		vision and joint pre-training
GZ10_DECals	LegacySurvey *grz* images	14189	3546	10 classes	classification
GZ10_xmatch	HSC *grz* images	553	139	10 classes	classification
SDSS16_reg	SDSS spectra	14000	3000	redshift	regression
SDSS16_class	SDSS spectra	182000	39000	spectral class	classification

Table 1: The datasets used in this study.

For cross-modal alignment, we use paired imaging and spectroscopic observations of the same astrophysical objects. The pre-trained image and spectral encoders are learned from larger unimodal datasets (LVD-1689M for DINOv3 and SDSS/DESI data for AION-1), while we additionally investigate whether unimodal pre-training the alignment blocks improves joint representations. The datasets used for unimodal pre-training (when applicable), joint multi-modal alignment, and downstream task evaluation (described in Section 4) are summarized in Table 1. Further details of dataset construction are in Appendix A.2.

Image data originate from the Hyper Suprime-Cam's (HSC) deep and ultra-deep surveys [Aihara et al., 2018], which provides calibrated optical imaging in five bands (g, r, i, z, and y). We use the g, r, and z bands which cover a broad redshift range to construct RGB images using the Lupton et al. [2004] *asinh* intensity mapping. Image cutouts are extracted at a pixel scale of 0.162 arcsec pixel^{-1}, centered on cataloged sources, with a final size of 160×160 pixels. A total of 86,614 such cutouts were obtained.

Optical spectra were from the Sloan Digital Sky Survey (SDSS; Ahumada et al. [2020]) and the Dark Energy Spectroscopic Instrument (DESI; Collaboration et al. [2016]). SDSS provides medium-resolution optical spectra covering approximately 3650–10400 Å, while DESI provides spectra spanning approximately 3600–9800 Å with wavelength-dependent resolution. The combined SDSS and DESI dataset contained 209630 object spectra.

Finally, we constructed cross-matched HSC and DESI objects within a one arcsecond radius to obtain a dataset containing 20,472 cutouts. These paired observations are used for CLIP-style contrastive training, to align image and spectrum data in a shared representation space. Objects from this cross-

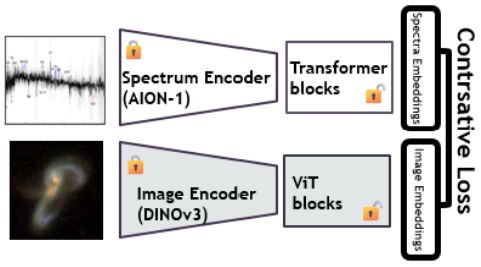

Figure 1: The DINOspec architecture.

Model Component	Parameter Count
DINOv3 Large 🔒	300M
AION-1 Base 🔒	300M
ViT block	7M
2x ViT block	14M
Embedding layer	1.18M
2 block Transformer	6.9M
8 block Transformer	27.8M
16 block Transformer	55.6M
2-layer MLP	0.33M

Table 2: Model parameter counts.

Configuration #	Assumption	Vision blocks	Transformer blocks	Training strategy
0	Strong spectra	Pre-trained	Embedding layer only	Joint training
1	Strong spectra	Random	Pre-trained	SM frozen for first 50 epochs
2	Strong vision	Pre-trained	Pre-trained	VM frozen for first 50 epochs
3	Strong vision	Pre-trained	Random	VM frozen for first 50 epochs
4		Pre-trained	Random	Joint training
5		Pre-trained	Pre-trained	Joint training
6		Random	Pre-trained	Joint training

Table 3: Experiment configurations. The vision backbone and spectral tokenizer are always frozen. The vision backbone plus any additional blocks compose the entire vision model (VM); likewise for the spectral model (SM). Initialization of additional blocks is either "Pre-trained" or "Random".

matched dataset were excluded from any downstream task evaluation datasets in order to prevent leakage.

3 Models

DINOspec adopts the structure introduced in DINO.txt [Jose et al., 2024], consisting of a frozen DINOv3-Large vision backbone and a spectral branch operating on tokens from the frozen AION-1 tokenizer. The vision branch appends transformer blocks to the DINOv3 output, while the spectral branch uses either a linear embedding layer or additional transformer blocks. The two branches are connected through trainable adapters, which align visual and spectral representations using the symmetric cross-entropy CLIP loss [Radford et al., 2021]. Table 2 shows different components of the DINOspec model which we investigated in this work, and the number of trainable parameters in each component. For comparison, the parameter counts of the DINOv3 and AION-1 models used are also included, although these models are used exclusively as frozen encoders.

We designed seven experimental configurations (Table 3) to investigate multimodal alignment under different assumptions about the relative strength of the vision and spectral encoders. Here, pre-training refers to unimodal training of randomly initialized transformer blocks using either the BERT reconstruction objective or the DINOv3 loss; details are provided in the Appendix.

Configuration 0 uses pre-trained vision blocks with a lightweight spectral embedding layer, while Configuration 1 introduces a pre-trained spectral model and evaluates alignment when the spectral branch guides the visual representation. Configurations 2 and 3 reverse this assumption, using pre-trained visual representations while varying the initialization of the spectral adapter. Finally, Configurations 4–6 make no specific assumption about the strength of either encoder, and compare different combinations of pre-trained and randomly initialized components when all trainable modules are optimized jointly. For each configuration, joint training was performed using the cross-matched dataset for 150 epochs and a learning rate of $5e-6$.

4 Evaluation

We evaluate the utility of aligned representations on both vision and spectral downstream tasks rather than direct measures of alignment such as retrieval or cross-modal matching. Images and spectra provide complementary, but not identical, information about astronomical objects; retrieval performance alone therefore does not establish whether an alignment preserves the modality-specific information required for scientific inference. Our evaluation measures downstream utility, rather than simply alignment quality.

Galaxy morphology classification used the 10-class Galaxy10 DECaLS dataset [Leung, 2026]. Since AION-1 requires a specific image preprocessing pipeline, we cross-matched Galaxy10 objects with available HSC imaging, resulting in 692 usable images (GZ10_xmatch in Table 1). For consistency, AION-1 was evaluated using only HSC images and no additional data modalities. Unless otherwise specified, image morphology classification used only the vision encoder and a 2-layer MLP classifier.

As a complementary assessment on a larger dataset and imaging source than the HSC subset used for the primary morphology evaluation, we evaluate morphology classification on the full Galaxy10 DECaLS dataset using Legacy Survey images. LegacySurvey images are have a larger pixel scale than HSC images, resulting in lower resolution, and the survey prioritized coverage over depth. Together with different response profiles in the same imaging bands, these galactic images have slightly different properties than the joint training distribution DINOspec models were exposed to. This experiment should be treated as a qualitative check of whether the results observed in the primary evaluation persist on a larger dataset, rather than as a controlled comparison of performance across imaging domains. Full results of this evaluation are in Appendix A.5.

Spectral tasks used the SDSS16 dataset [Zhong et al., 2024], which provides labels for redshift and spectral class. We evaluated redshift regression for galaxies and classification of stars, galaxies, and quasars (SDSS16_reg and SDSS16_class in Table 1). As for vision tasks, evaluation used only the corresponding spectral encoder with a 2-layer MLP classifier, modified for classification or regression.

Learning rates were selected through a small hyperparameter sweep for each task, with early stopping applied. Each evaluation was repeated seven times with different random seeds. We report F1 scores for classification and the coefficient of determination (R^2) for redshift prediction.

#	Transformer Blocks	Pre-trained ViT Blocks?	Trainable Parameters	Image F1	Spectra F1	Spectra R^2
1	2*	N + LoRA	21.4M	$\mathbf{0.796 \pm 5.8e-3}$	$0.708 \pm 3.4e-3$	$0.879 \pm 1.7e-2$
1	2*	N	21M	$0.779 \pm 6.1e-3$	$0.704 \pm 5.7e-3$	$0.882 \pm 1.4e-2$
1	16*	N	70M	$0.773 \pm 5.5e-3$	$0.651 \pm 1.4e-3$	$0.893 \pm 1.0e-3$
6	16	N	70M	$0.768 \pm 5.5e-3$	$0.659 \pm 3.5e-3$	$0.891 \pm 5.5e-4$
6	8	N	42M	$0.764 \pm 5.3e-3$	$0.663 \pm 4.7e-3$	$0.895 \pm 1.0e-3$
1	8*	N	42M	$0.758 \pm 5.3e-3$	$0.667 \pm 5.5e-3$	$0.895 \pm 6.2e-4$
6	2	N	21M	$0.747 \pm 4.6e-3$	$0.688 \pm 3.9e-3$	$0.884 \pm 1.6e-2$
2	2	Y*	21M	$0.732 \pm 4.2e-3$	$\mathbf{0.736 \pm 2.8e-3}$	$0.879 \pm 3.9e-3$
4	16ˆ	Y	70M	$0.731 \pm 6.1e-3$	$0.732 \pm 1.4e-3$	$0.860 \pm 4.4e-3$
0	-	Y	15M	$0.731 \pm 6.0e-3$	$0.697 \pm 1.0e-3$	$0.910 \pm 4.8e-3$
5	8	Y	42M	$0.728 \pm 8.6e-3$	$0.711 \pm 4.5e-3$	$0.898 \pm 1.1e-3$
4	8ˆ	Y	42M	$0.728 \pm 6.2e-3$	$0.717 \pm 2.2e-3$	$0.891 \pm 5.9e-3$
5	2	Y	21M	$0.724 \pm 5.7e-3$	$0.706 \pm 2.1e-3$	$0.890 \pm 3.0e-3$
3	2ˆ	Y*	21M	$0.724 \pm 5.6e-3$	$0.714 \pm 5.1e-3$	$0.873 \pm 2.3e-3$
	DINOv3 ViT-L	+ LoRA + MLP	15.33M	$0.757 \pm 1.9e-2$	-	-
	DINOv3 ViT-L	+ MLP	0.33M	$0.722 \pm 5.2e-3$	-	-
	AION-1 B	+ MLP	0.33M	$0.702 \pm 1.2e-2$	$0.697 \pm 1.0e-3$	$\mathbf{0.910 \pm 4.8e-3}$

Table 4: Evaluation results from joint pre-training (* frozen model first 50 epochs, ˆrandomly initialized). Baseline comparisons to vision (DINOv3) and multimodal (AION-1) models included.

5 Results

Table 4 summarizes performance across image and spectral tasks. DINOspec improves morphology and spectral classification over the DINOv3-L and AION-1 baselines while maintaining similar redshift prediction performance. When the spectral encoder is used to guide alignment with the image branch (Configurations 1 and 6), galaxy morphology classification sees the strongest improvements,

from the baseline F1=0.72 to F1=0.78, which out-performs DINOv3 even after LoRA fine-tuning. In contrast, spectral classification improves the most from F1=0.70 to F1=0.74 when alignment is performed after pre-training of the extra vision and spectral blocks, keeping the entire image branch frozen for the first 50 epochs of joint training and allowing the visual features to guide alignment. Notably, the same pre-trained components when trained jointly without restriction (Configuration 5) do not result in the same magnitude of improvement in either morphology or spectral classification.

Visual galaxy morphology classification is the task that benefits most from multimodal joint training. As there exists some domain shift given that the DINOv3 backbone was likely not exposed to significant samples of different galaxy types during pre-training, this is perhaps to be expected. The top result, that Configuration 1 with the pre-trained spectral model performs best on morphology classification, is confirmed by our additional evaluation of the full Galaxy10 DECaLS dataset using Legacy Survey images, as is the general finding that multimodal joint training benefits downstream morphology classification over baselines. LoRA adaptation of the vision backbone for that best-performing configuration achieved F1=0.796 on HSC images. Since this requires updating an additional $\approx 0.6M$ parameters, it is not directly comparable to the other configurations but demonstrates an efficient alternative for improving visual performance. Increasing the number of spectral adapter parameters did not consistently improve performance. This suggests that the AION-1 spectral encoder already contains sufficient information for alignment, a hypothesis that is supported by the fact that using only an embedding layer instead of multiple transformer blocks still results in an improvement in visual morphology classification.

Unlike morphology and spectral classification, redshift prediction shows no consistent improvement over the AION-1 baseline ($R^2 = 0.910$). Some configurations approach this performance, while others show a reduction in R^2. A possible explanation is that the spectral encoder already captures the information required for redshift estimation, while the image modality provides little additional signal. The use of Lupton-scaled grz images, which removes the exact flux ratios needed for photometric redshift estimation, may also limit the usefulness of images for redshift estimation.

Together, these results show that lightweight alignment can improve downstream performance without jointly training large multimodal encoders, provided sufficiently strong unimodal representations exist. The strong performance of DINOv3 across the HSC and Legacy Survey morphology evaluations also suggests that general-purpose vision backbones pretrained on natural images can be useful for astronomical applications, potentially offering greater flexibility across imaging surveys than more specialized models like AION-1. This highlights the potential of alignment modules to adapt broadly pretrained visual representations to astronomical data without requiring a dedicated vision encoder for each instrument or survey.

6 Conclusion

The increasing diversity of astrophysical surveys and instruments makes it impractical to rely exclusively on jointly trained multimodal architectures. Our results demonstrate that representation learning and cross-modal alignment can be separated, allowing independently developed foundation models to be reused. Adapter-based alignment using as few as 21M parameters successfully combines independently pre-trained vision and spectral models while preserving their individual capabilities. Improvements in morphology and spectral classification shows alignment is most effective when modalities contain complementary information about the physical object. Conversely, unchanged redshift prediction demonstrates that alignment does not provide additional information when one modality already contains the dominant signal for a task. The benefit of multimodal representation learning therefore depends on the relationship between modalities and the objectives.

DINOspec extends representation alignment methods to astronomical observations, where different instruments capture distinct physical processes across the electromagnetic spectrum. Rather than requiring a new jointly trained architecture for each combination of modalities, it provides a framework for connecting independently developed scientific foundation models through alignment modules each containing only $\mathcal{O}(10^6)$ parameters.

References

Romina Ahumada, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, Scott F. Anderson, Brett H. Andrews, Borja Anguiano, Riccardo Arcodia, Eric Armengaud, Marie Aubert, Santiago Avila, Vladimir Avila-Reese, Carles Badenes, Christophe Balland, Kat Barger, Jorge K. Barrera-Ballesteros, Sarbani Basu, Julian Bautista, Rachael L. Beaton, Timothy C. Beers, B. Izamar T. Benavides, Chad F. Bender, Mariangela Bernardi, Matthew Bershady, Florian Beutler, Christian Moni Bidin, Jonathan Bird, Dmitry Bizyaev, Guillermo A. Blanc, Michael R. Blanton, Médéric Boquien, Jura Borissova, Jo Bovy, W. N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Kevin Bundy, Martin Bureau, Adam Burgasser, Etienne Burtin, Mariana Cano-Díaz, Raffaella Capasso, Michele Cappellari, Ricardo Carrera, Solène Chabanier, William Chaplin, Michael Chapman, Brian Cherinka, Cristina Chiappini, Peter Doohyun Choi, S. Drew Chojnowski, Haeun Chung, Nicolas Clerc, Damien Coffey, Julia M. Comerford, Johan Comparat, Luiz da Costa, Marie-Claude Cousinou, Kevin Covey, Jeffrey D. Crane, Katia Cunha, Gabriele da Silva Ilha, Yu Sophia Dai, Sanna B. Damsted, Jeremy Darling, James W. Davidson, Roger Davies, Kyle Dawson, Nikhil De, Axel de la Macorra, Nathan De Lee, Anna Bárbara de Andrade Queiroz, Alice Deconto Machado, Sylvain de la Torre, Flavia Dell'Agli, Hélion du Mas des Bourboux, Aleksandar M. Diamond-Stanic, Sean Dillon, John Donor, Niv Drory, Chris Duckworth, Tom Dwelly, Garrett Ebelke, Sarah Eftekharzadeh, Arthur Davis Eigenbrot, Yvonne P. Elsworth, Mike Eracleous, Ghazaleh Erfanianfar, Stephanie Escoffier, Xiaohui Fan, Emily Farr, José G. Fernández-Trincado, Diane Feuillet, Alexis Finoguenov, Patricia Fofie, Amelia Fraser-McKelvie, Peter M. Frinchaboy, Sebastien Fromenteau, Hai Fu, Lluís Galbany, Rafael A. Garcia, D. A. García-Hernández, Luis Alberto Garma Oehmichen, Junqiang Ge, Marcio Antonio Geimba Maia, Doug Geisler, Joseph Gelfand, Julian Goddy, Violeta Gonzalez-Perez, Kathleen Grabowski, Paul Green, Catherine J. Grier, Hong Guo, Julien Guy, Paul Harding, Sten Hasselquist, Adam James Hawken, Christian R. Hayes, Fred Hearty, S. Hekker, David W. Hogg, Jon A. Holtzman, Danny Horta, Jiamin Hou, Bau-Ching Hsieh, Daniel Huber, Jason A. S. Hunt, J. Ider Chitham, Julie Imig, Mariana Jaber, Camilo Eduardo Jimenez Angel, Jennifer A. Johnson, Amy M. Jones, Henrik Jönsson, Eric Jullo, Yerim Kim, Karen Kinemuchi, Charles C. Kirkpatrick IV, George W. Kite, Mark Klaene, Jean-Paul Kneib, Juna A. Kollmeier, Hui Kong, Marina Kounkel, Dhanesh Krishnarao, Ivan Lacerna, Ting-Wen Lan, Richard R. Lane, David R. Law, Jean-Marc Le Goff, Henry W. Leung, Hannah Lewis, Cheng Li, Jianhui Lian, Lihwai Lin, Dan Long, Penélope Longa-Peña, Britt Lundgren, Brad W. Lyke, J. Ted Mackereth, Chelsea L. MacLeod, Steven R. Majewski, Arturo Manchado, Claudia Maraston, Paul Martini, Thomas Masseron, Karen L. Masters, Savita Mathur, Richard M. McDermid, Andrea Merloni, Michael Merrifield, Szabolcs Mészáros, Andrea Miglio, Dante Minniti, Rebecca Minsley, Takamitsu Miyaji, Faizan Gohar Mohammad, Benoit Mosser, Eva-Maria Mueller, Demitri Muna, Andrea Muñoz-Gutiérrez, Adam D. Myers, Seshadri Nadathur, Preethi Nair, Kirpal Nandra, Janaina Correa do Nascimento, Rebecca Jean Nevin, Jeffrey A. Newman, David L. Nidever, Christian Nitschelm, Pasquier Noterdaeme, Julia E. O'Connell, Matthew D. Olmstead, Daniel Oravetz, Audrey Oravetz, Yeisson Osorio, Zachary J. Pace, Nelson Padilla, Nathalie Palanque-Delabrouille, Pedro A. Palicio, Hsi-An Pan, Kaike Pan, James Parker, Romain Paviot, Sebastien Peirani, Karla Peña Ramírez, Samantha Penny, Will J. Percival, Ismael Perez-Fournon, Ignasi Pérez-Ràfols, Patrick Petitjean, Matthew M. Pieri, Marc Pinsonneault, Vijith Jacob Poovelil, Joshua Tyler Povick, Abhishek Prakash, Adrian M. Price-Whelan, M. Jordan Raddick, Anand Raichoor, Amy Ray, Sandro Barboza Rembold, Mehdi Rezaie, Rogemar A. Riffel, Rogério Riffel, Hans-Walter Rix, Annie C. Robin, A. Roman-Lopes, Carlos Román-Zúñiga, Benjamin Rose, Ashley J. Ross, Graziano Rossi, Kate Rowlands, Kate H. R. Rubin, Mara Salvato, Ariel G. Sánchez, Laura Sánchez-Menguiano, José R. Sánchez-Gallego, Conor Sayres, Adam Schaefer, Ricardo P. Schiavon, Jaderson S. Schimoia, Edward Schlafly, David Schlegel, Donald P. Schneider, Mathias Schultheis, Axel Schwope, Hee-Jong Seo, Aldo Serenelli, Arman Shafieloo, Shoaib Jamal Shamsi, Zhengyi Shao, Shiyin Shen, Matthew Shetrone, Raphael Shirley, Víctor Silva Aguirre, Joshua D. Simon, M. F. Skrutskie, Anže Slosar, Rebecca Smethurst, Jennifer Sobeck, Bernardo Cervantes Sodi, Diogo Souto, David V. Stark, Keivan G. Stassun, Matthias Steinmetz, Dennis Stello, Julianna Stermer, Thaisa Storchi-Bergmann, Alina Streblyanska, Guy S. Stringfellow, Amelia Stutz, Genaro Suárez, Jing Sun, Manuchehr Taghizadeh-Popp, Michael S. Talbot, Jamie Tayar, Aniruddha R. Thakar, Riley Theriault, Daniel Thomas, Zak C. Thomas, Jeremy Tinker, Rita Tojeiro, Hector Hernandez Toledo, Christy A. Tremonti, Nicholas W. Troup, Sarah Tuttle, Eduardo Unda-Sanzana, Marica Valentini, Jaime Vargas-González, Mariana Vargas-Magaña, Jose Antonio Vázquez-Mata, M. Vivek, David Wake, Yuting Wang, Benjamin Alan Weaver, Anne-Marie Weij-

mans, Vivienne Wild, John C. Wilson, Robert F. Wilson, Nathan Wolthuis, W. M. Wood-Vasey, Renbin Yan, Meng Yang, Christophe Yèche, Olga Zamora, Pauline Zarrouk, Gail Zasowski, Kai Zhang, Cheng Zhao, Gongbo Zhao, Zheng Zheng, Zheng Zheng, Guangtun Zhu, and Hu Zou. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *ApJS*, 249(1):3, June 2020. ISSN 0067-0049. doi: 10.3847/1538-4365/ab929e. URL `https://doi.org/10.3847/1538-4365/ab929e`.

Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, Stéphane Arnouts, Neta A Bahcall, Steven Bickerton, James Bosch, Kevin Bundy, Peter L Capak, James H H Chan, Masashi Chiba, Jean Coupon, Eiichi Egami, Motohiro Enoki, Francois Finet, Hiroki Fujimori, Seiji Fujimoto, Hisanori Furusawa, Junko Furusawa, Tomotsugu Goto, Andy Goulding, Johnny P Greco, Jenny E Greene, James E Gunn, Takashi Hamana, Yuichi Harikane, Yasuhiro Hashimoto, Takashi Hattori, Masao Hayashi, Yusuke Hayashi, Krzysztof G Hełminiak, Ryo Higuchi, Chiaki Hikage, Paul T P Ho, Bau-Ching Hsieh, Kuiyun Huang, Song Huang, Hiroyuki Ikeda, Masatoshi Imanishi, Akio K Inoue, Kazushi Iwasawa, Ikuru Iwata, Anton T Jaelani, Hung-Yu Jian, Yukiko Kamata, Hiroshi Karoji, Nobunari Kashikawa, Nobuhiko Katayama, Satoshi Kawanomoto, Issha Kayo, Jin Koda, Michitaro Koike, Takashi Kojima, Yutaka Komiyama, Akira Konno, Shintaro Koshida, Yusei Koyama, Haruka Kusakabe, Alexie Leauthaud, Chien-Hsiu Lee, Lihwai Lin, Yen-Ting Lin, Robert H Lupton, Rachel Mandelbaum, Yoshiki Matsuoka, Elinor Medezinski, Sogo Mineo, Shoken Miyama, Hironao Miyatake, Satoshi Miyazaki, Rieko Momose, Anupreeta More, Surhud More, Yuki Moritani, Takashi J Moriya, Tomoki Morokuma, Shiro Mukae, Ryoma Murata, Hitoshi Murayama, Tohru Nagao, Fumiaki Nakata, Mana Niida, Hiroko Niikura, Atsushi J Nishizawa, Yoshiyuki Obuchi, Masamune Oguri, Yukie Oishi, Nobuhiro Okabe, Sakurako Okamoto, Yuki Okura, Yoshiaki Ono, Masato Onodera, Masafusa Onoue, Ken Osato, Masami Ouchi, Paul A Price, Tae-Soo Pyo, Masao Sako, Marcin Sawicki, Takatoshi Shibuya, Kazuhiro Shimasaku, Atsushi Shimono, Masato Shirasaki, John D Silverman, Melanie Simet, Joshua Speagle, David N Spergel, Michael A Strauss, Yuma Sugahara, Naoshi Sugiyama, Yasushi Suto, Sherry H Suyu, Nao Suzuki, Philip J Tait, Masahiro Takada, Tadafumi Takata, Naoyuki Tamura, Manobu M Tanaka, Masaomi Tanaka, Masayuki Tanaka, Yoko Tanaka, Tsuyoshi Terai, Yuichi Terashima, Yoshiki Toba, Nozomu Tominaga, Jun Toshikawa, Edwin L Turner, Tomohisa Uchida, Hisakazu Uchiyama, Keiichi Umetsu, Fumihiro Uraguchi, Yuji Urata, Tomonori Usuda, Yousuke Utsumi, Shiang-Yu Wang, Wei-Hao Wang, Kenneth C Wong, Kiyoto Yabe, Yoshihiko Yamada, Hitomi Yamanoi, Naoki Yasuda, Sherry Yeh, Atsunori Yonehara, and Suraphong Yuma. The Hyper Suprime-Cam SSP Survey: Overview and survey design. *Publ Astron Soc Jpn Nihon Tenmon Gakkai*, 70(SP1):S4, January 2018. ISSN 0004-6264. doi: 10.1093/pasj/psx066. URL `https://doi.org/10.1093/pasj/psx066`.

DESI Collaboration, Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E. Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balland, Otger Ballester, Charles Baltay, Lucas Beaufore, Chris Bebek, Timothy C. Beers, Eric F. Bell, José Luis Bernal, Robert Besuner, Florian Beutler, Chris Blake, Hannes Bleuler, Michael Blomqvist, Robert Blum, Adam S. Bolton, Cesar Briceno, David Brooks, Joel R. Brownstein, Elizabeth Buckley-Geer, Angela Burden, Etienne Burtin, Nicolas G. Busca, Robert N. Cahn, Yan-Chuan Cai, Laia Cardiel-Sas, Raymond G. Carlberg, Pierre-Henri Carton, Ricard Casas, Francisco J. Castander, Jorge L. Cervantes-Cota, Todd M. Claybaugh, Madeline Close, Carl T. Coker, Shaun Cole, Johan Comparat, Andrew P. Cooper, M.-C. Cousinou, Martin Crocce, Jean-Gabriel Cuby, Daniel P. Cunningham, Tamara M. Davis, Kyle S. Dawson, Axel de la Macorra, Juan De Vicente, Timothée Delubac, Mark Derwent, Arjun Dey, Govinda Dhungana, Zhejie Ding, Peter Doel, Yutong T. Duan, Anne Ealet, Jerry Edelstein, Sarah Eftekharzadeh, Daniel J. Eisenstein, Ann Elliott, Stéphanie Escoffier, Matthew Evatt, Parker Fagrelius, Xiaohui Fan, Kevin Fanning, Arya Farahi, Jay Farihi, Ginevra Favole, Yu Feng, Enrique Fernandez, Joseph R. Findlay, Douglas P. Finkbeiner, Michael J. Fitzpatrick, Brenna Flaugher, Samuel Flender, Andreu Font-Ribera, Jaime E. Forero-Romero, Pablo Fosalba, Carlos S. Frenk, Michele Fumagalli, Boris T. Gaensicke, Giuseppe Gallo, Juan Garcia-Bellido, Enrique Gaztanaga, Nicola Pietro Gentile Fusillo, Terry Gerard, Irena Gershkovich, Tommaso Giannantonio, Denis Gillet, Guillermo Gonzalez-de Rivera, Violeta Gonzalez-Perez, Shelby Gott, Or Graur, Gaston Gutierrez, Julien Guy, Salman Habib, Henry Heetderks, Ian Heetderks, Katrin Heitmann, Wojciech A. Hellwing, David A. Herrera, Shirley Ho, Stephen Holland, Klaus Honscheid, Eric Huff, Timothy A. Hutchinson, Dragan Huterer, Ho Seong Hwang, Joseph Maria Illa Laguna, Yuzo Ishikawa, Dianna Jacobs, Niall Jeffrey, Patrick Jelinsky, Elise Jennings, Linhua Jiang, Jorge Jimenez, Jennifer Johnson, Richard Joyce, Eric Jullo, Stéphanie Juneau, Sami Kama,

Armin Karcher, Sonia Karkar, Robert Kehoe, Noble Kennamer, Stephen Kent, Martin Kilbinger, Alex G. Kim, David Kirkby, Theodore Kisner, Ellie Kitanidis, Jean-Paul Kneib, Sergey Koposov, Eve Kovacs, Kazuya Koyama, Anthony Kremin, Richard Kron, Luzius Kronig, Andrea Kueter-Young, Cedric G. Lacey, Robin Lafever, Ofer Lahav, Andrew Lambert, Michael Lampton, Martin Landriau, Dustin Lang, Tod R. Lauer, Jean-Marc Le Goff, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Su-Jeong Lee, Daniela Leitner, Michael Lesser, Michael E. Levi, Benjamin L'Huillier, Baojiu Li, Ming Liang, Huan Lin, Eric Linder, Sarah R. Loebman, Zarija Lukić, Jun Ma, Niall MacCrann, Christophe Magneville, Laleh Makarem, Marc Manera, Christopher J. Manser, Robert Marshall, Paul Martini, Richard Massey, Thomas Matheson, Jeremy McCauley, Patrick McDonald, Ian D. McGreer, Aaron Meisner, Nigel Metcalfe, Timothy N. Miller, Ramon Miquel, John Moustakas, Adam Myers, Milind Naik, Jeffrey A. Newman, Robert C. Nichol, Andrina Nicola, Luiz Nicolati da Costa, Jundan Nie, Gustavo Niz, Peder Norberg, Brian Nord, Dara Norman, Peter Nugent, Thomas O'Brien, Minji Oh, Knut A. G. Olsen, Cristobal Padilla, Hamsa Padmanabhan, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Antonella Palmese, Daniel Pappalardo, Isabelle Pâris, Changbom Park, Anna Patej, John A. Peacock, Hiranya V. Peiris, Xiyan Peng, Will J. Percival, Sandrine Perruchot, Matthew M. Pieri, Richard Pogge, Jennifer E. Pollack, Claire Poppett, Francisco Prada, Abhishek Prakash, Ronald G. Probst, David Rabinowitz, Anand Raichoor, Chang Hee Ree, Alexandre Refregier, Xavier Regal, Beth Reid, Kevin Reil, Mehdi Rezaie, Constance M. Rockosi, Natalie Roe, Samuel Ronayette, Aaron Roodman, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, Eduardo Rozo, Vanina Ruhlmann-Kleider, Eli S. Rykoff, Cristiano Sabiu, Lado Samushia, Eusebio Sanchez, Javier Sanchez, David J. Schlegel, Michael Schneider, Michael Schubnell, Aurélia Secroun, Uros Seljak, Hee-Jong Seo, Santiago Serrano, Arman Shafieloo, Huanyuan Shan, Ray Sharples, Michael J. Sholl, William V. Shourt, Joseph H. Silber, David R. Silva, Martin M. Sirk, Anze Slosar, Alex Smith, George F. Smoot, Debopam Som, Yong-Seon Song, David Sprayberry, Ryan Staten, Andy Stefanik, Gregory Tarle, Suk Sien Tie, Jeremy L. Tinker, Rita Tojeiro, Francisco Valdes, Octavio Valenzuela, Monica Valluri, Mariana Vargas-Magana, Licia Verde, Alistair R. Walker, Jiali Wang, Yuting Wang, Benjamin A. Weaver, Curtis Weaverdyck, Risa H. Wechsler, David H. Weinberg, Martin White, Qian Yang, Christophe Yeche, Tianmeng Zhang, Gong-Bo Zhao, Yi Zheng, Xu Zhou, Zhimin Zhou, Yaling Zhu, Hu Zou, and Ying Zu. The DESI Experiment Part I: Science,Targeting, and Survey Design, December 2016. URL http://arxiv.org/abs/1611.00036. arXiv:1611.00036 [astro-ph.IM].

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. DINOv2 Meets Text: A Unified Framework for Image- and Pixel-Level Vision-Language Alignment, December 2024. URL http://arxiv.org/abs/2412.16334. arXiv:2412.16334 [cs].

E. Lastufka, O. Bait, M. Drozdova, V. Kinakh, D. Piras, M. Audard, M. Dessauges-Zavadsky, T. Holotyak, D. Schaerer, and S. Voloshynovskiy. Examining vision foundation models for classification and detection in optical and radio astronomy. *A&A*, 703:A217, November 2025. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202553691. URL https://www.aanda.org/articles/aa/abs/2025/11/aa53691-25/aa53691-25.html.

Henry Leung. henrysky/Galaxy10, July 2026. URL https://github.com/henrysky/Galaxy10. original-date: 2021-03-16T20:23:19Z.

Robert Lupton, Michael R. Blanton, George Fekete, David W. Hogg, Wil O'Mullane, Alex Szalay, and Nicholas Wherry. Preparing Red-Green-Blue Images from CCD Data. *Publications of the Astronomical Society of the Pacific*, 116:133–137, February 2004. ISSN 0004-6280. doi: 10.1086/382245. URL https://ui.adsabs.harvard.edu/abs/2004PASP..116..133L. ADS Bibcode: 2004PASP..116..133L.

Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Ruben Ohana, Mariel Pettee, Bruno Regaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. AstroCLIP: A Cross-Modal Foundation Model for Galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, June 2024. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stae1450. URL http://arxiv.org/abs/2310.03024. arXiv:2310.03024 [astro-ph.IM].

Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Nathan Casserau, Pierre Cornette, Keiya Hirashima, Geraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay, Mariel Pettee, Bruno Regaldo-Saint Blancard, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. AION-1: Omnimodal Foundation Model for Astronomical Sciences, October 2025. URL http://arxiv.org/abs/2510.17960. arXiv:2510.17960 [astro-ph.IM].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

M. Rizhko and J. S. Bloom. AstroM3: A Self-supervised Multimodal Model for Astronomy. *AJ*, 170(1):28, June 2025. ISSN 1538-3881. doi: 10.3847/1538-3881/adcbad. URL https://dx.doi.org/10.3847/1538-3881/adcbad.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. URL http://arxiv.org/abs/2508.10104. arXiv:2508.10104 [cs.CV].

Fucheng Zhong, Nicola R Napolitano, Caroline Heneka, Rui Li, Franz Erik Bauer, Nicolas Bouche, Johan Comparat, Young-Lo Kim, Jens-Kristian Krogager, Marcella Longhetti, Jonathan Loveday, Boudewijn F Roukema, Benedict L Rouse, Mara Salvato, Crescenzo Tortora, Roberto J Assef, Letizia P Cassarà, Luca Costantin, Scott M Croom, Luke J M Davies, Alexander Fritz, Guillaume Guiglion, Andrew Humphrey, Emanuela Pompei, Claudio Ricci, Cristóbal Sifón, Elmo Tempel, and Tayyaba Zafar. Galaxy Spectra neural Network (GaSNet). II. Using deep learning for spectral classification and redshift predictions. *Mon Not R Astron Soc*, 532(1):643–665, July 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae1461. URL https://doi.org/10.1093/mnras/stae1461.

A Appendix

A.1 Code

Code for unimodal and joint pre-training, as well as evaluation, is available at `https://github.com/elastufka/dinospec_repo`.

A.2 Dataset construction

HSC, SDSS, and DESI data were sourced from the Multimodal Universe [] version 1, available at `https://users.flatironinstitute.org/~polymathic/data/MultimodalUniverse/v1/`

Because the images in the GZ10 evaluation dataset are grz LegacySurvey images and the AION-1 tokenizer requires $griz$ LegacySurvey images or $grizy$ HSC images as input, we chose to cross-match the GZ10 sources to available HSC data since we already had the required five-channel HSC images on hand. This resulted in a much smaller evaluation dataset than the original GZ10, with only 8 samples of the rarest class. We performed a class-balanced train-test split to ensure that all classes were represented in both the train and test sets. The use of this smaller dataset for evaluation is why the reported F1-score for the AION-1 baseline on the galaxy morphology classification task is lower than what is found in their paper.

Data and labels for the spectral classification and redshift prediction tasks are from `https://huggingface.co/datasets/Fucheng/GaSNet-II-SDSS-dataset`

A.3 Adapter pre-training

The additional transformer blocks used for multimodal alignment were optionally pre-trained before joint vision–spectra training. The underlying DINOv3 vision backbone and AION-1 spectral tokenizer were kept frozen during this stage. The two vision adapter blocks were initialized with random weights and then pre-trained using the DINOv3 self-distillation objective on the *hsc* dataset. Given student and teacher network outputs p_s and p_t, the objective minimizes the cross-entropy between their softened output distributions,

$$\mathcal{L}_{\mathrm{DINO}} = -\sum_{k} p_t(k) \log p_s(k), \tag{1}$$

where the teacher network is updated using an exponential moving average of the student parameters. After 50 pre-training epochs, the vision model (DINOv3-Large frozen backbone plus two ViT blocks) achieved a F1 score of $0.717 \pm 7.1e-03$ on optical galaxy classification.

The spectral adapter blocks were pre-trained using a masked-token prediction objective analogous to the BERT objective. Spectral tokens produced by the pre-trained AION-1 tokenizer were randomly masked, and the transformer blocks were trained to reconstruct the missing tokens from the surrounding context. The objective is given by

$$\mathcal{L}_{\mathrm{BERT}} = -\sum_{i \in M} \log P(x_i | \mathbf{x}_{\setminus M}), \tag{2}$$

where M denotes the set of masked token positions. Spectral adapter pre-training was performed on the *sdss_desi* training dataset. A validation dataset was used to prevent overfitting. The 16-block transformer was pre-trained for fewer epochs than the others due to memory constraints. Spectral adapters were pre-trained for longer than the vision adapter because without the the inference cost ($\sim$ 17 ms/image) of the DINOv3 backbone training proceeded much faster. Table 5 shows the performance of the pre-trained spectral adapters on the redshift prediction task.

The frozen DINOv3-L vision encoder provides a strong baseline for morphology classification, achieving an F1 score of $0.722 \pm 5.2 \times 10^{-3}$. Adding pre-trained vision adapter blocks without multimodal alignment does not improve performance, indicating that additional visual capacity alone is insufficient.

Model	Pre-training Epochs	z R^2 score
2 block Transformer	500	$0.912 \pm 8.1e-03$
8 block Transformer	500	$0.897 \pm 9.3e-03$
16 block Transformer	100	$0.906 \pm 3.4e-03$

Table 5: Spectral adapter pre-training

A.4 Additional joint training and evaluation details

Joint training was performed on paired image–spectrum observations. For each batch of N pairs, the vision and spectral encoders produce L2-normalized representations $\mathbf{v}i$ and $\mathbf{s}i$. Pairwise similarity logits are computed as $Sij = \exp(\gamma), \mathbf{v}i^\top \mathbf{s}j$, where γ is a learned logit-scale parameter. We optimize the symmetric cross-entropy CLIP objective,

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N} \sum i = 1^N \left[\log \frac{e^{Sii}}{\sum j = 1^N e^{S_{ij}}} + \log \frac{e^{S_{ii}}}{\sum_{j=1}^N e^{S_{ji}}} \right] \tag{3}$$

The first term performs image-to-spectrum classification and the second spectrum-to-image classification. Thus, each paired observation forms a positive pair, while the remaining observations in the batch serve as in-batch negatives.

Training was performed for 150 epochs using a learning rate of 5×10^{-6}, cosine learning-rate decay, a 10% warmup, weight decay of 0.01, and gradient clipping with a maximum norm of 1. Training used two distributed processes with 16 paired examples per process and four gradient-accumulation steps, corresponding to an effective batch size of 128 paired examples. Depending on the experimental configuration, either the vision or spectral branch was frozen for an initial training period before being jointly optimized; the vision backbone and AION-1 spectral tokenizer remained frozen throughout.

For downstream evaluation, each trained branch was evaluated independently using a 2-layer MLP prediction head. Classification and regression heads were trained separately for each task, with learning rates selected through a small hyperparameter sweep between $5e-6$ and $1e-5$ and early stopping. Table 6 shows the learning rates selected for each model and task. Each evaluation was repeated with seven random seeds.

A.5 Complete Results

In the main text, we only reported results that exceeded the morphology classification baseline of DINOv3-Large in the paper body. Complete results are shown below, including the results using the full Galaxy10 DECaLS dataset using LegacySurvey images. No results on spectral classification or redshift prediction are available for the DINOv3 baselines, because the DINOv3 model applicable to images only.

#	Transformer Blocks	Pre-trained ViT Blocks?	Image lr	Spectra lr	Redshift lr
1	2*	N + LoRA	$5e-5$	$4e-6$	$1e-5$
1	2*	N	$2.5e-5$	$5e-6$	$1e-5$
1	16*	N	$4e-5$	$2e-6$	$5e-6$
6	16	N	$2e-5$	$5e-6$	$5e-6$
6	8	N	$2e-5$	$5e-6$	$1e-5$
1	8*	N	$5e-5$	$3e-6$	$7e-6$
6	2	N	$4e-5$	$2e-6$	$9e-6$
2	2	Y*	$3e-5$	$5e-6$	$6e-6$
4	16^	Y	$4.5e-5$	$2e-6$	$9e-6$
0	-	Y	$3.5e-5$	$2e-6$	$9e-6$
5	8	Y	$2e-5$	$5e-6$	$8e-6$
4	8^	Y	$2e-5$	$5e-6$	$9e-6$
5	2	Y	$4e-5$	$2e-6$	$8e-6$
3	2^	Y*	$4e-5$	$3e-6$	$1e-5$
2	8	Y*	$3e-5$	$3e-6$	$9e-6$
4	2^	Y	$3.5e-5$	$4e-6$	$9e-6$
5	16	Y	$3.5e-5$	$2e-6$	$5e-6$
3	8^	Y*	$4e-5$	$3e-6$	$9e-6$

Table 6: Learning rates selected for each model and task (galaxy morphology classification, spectral type classification, and redshift prediction) after a small sweep (* frozen model first 50 epochs, ^randomly initialized).

#	Transformer Blocks	Pre-trained ViT Blocks?	HSC Image F1	LS Image F1	Spectra F1	Spectra R^2
1	2*	N + LoRA	$\mathbf{0.796 \pm 5.8e-3}$		$0.708 \pm 3.4e-3$	$0.879 \pm 1.7e-2$
1	2*	N	$0.779 \pm 6.1e-3$	$\mathbf{0.788 \pm 2.7e-3}$	$0.704 \pm 5.7e-3$	$0.882 \pm 1.4e-2$
1	16*	N	$0.773 \pm 5.5e-3$	$0.775 \pm 1.9e-3$	$0.651 \pm 1.4e-3$	$0.893 \pm 1.0e-3$
6	16	N	$0.768 \pm 5.5e-3$	$0.773 \pm 1.7e-4$	$0.659 \pm 3.5e-3$	$0.891 \pm 5.5e-4$
6	8	N	$0.764 \pm 5.3e-3$	$0.771 \pm 4.4e-4$	$0.663 \pm 4.7e-3$	$0.895 \pm 1.0e-3$
1	8*	N	$0.758 \pm 5.3e-3$	$0.776 \pm 1.3e-3$	$0.667 \pm 5.5e-3$	$0.895 \pm 6.2e-4$
6	2	N	$0.747 \pm 4.6e-3$	$0.769 \pm 3.0e-3$	$0.688 \pm 3.9e-3$	$0.884 \pm 1.6e-2$
2	2	Y*	$0.732 \pm 4.2e-3$	$0.772 \pm 1.0e-3$	$\mathbf{0.736 \pm 2.8e-3}$	$0.879 \pm 3.9e-3$
4	16^	Y	$0.731 \pm 6.1e-3$	$0.773 \pm 1.3e-3$	$0.732 \pm 1.4e-3$	$0.860 \pm 4.4e-3$
0	-	Y	$0.731 \pm 6.0e-3$	$0.773 \pm 7.0e-4$	$0.697 \pm 1.0e-3$	$0.910 \pm 4.8e-3$
5	8	Y	$0.728 \pm 8.6e-3$	$0.771 \pm 4.4e-4$	$0.711 \pm 4.5e-3$	$0.898 \pm 1.1e-3$
4	8^	Y	$0.728 \pm 6.2e-3$	$0.771 \pm 2.0e-4$	$0.717 \pm 2.2e-3$	$0.891 \pm 5.9e-3$
5	2	Y	$0.724 \pm 5.7e-3$	$0.772 \pm 1.5e-3$	$0.706 \pm 2.1e-3$	$0.890 \pm 3.0e-3$
3	2^	Y*	$0.724 \pm 5.6e-3$	$0.768 \pm 2.7e-3$	$0.714 \pm 5.1e-3$	$0.873 \pm 2.3e-3$
2	8	Y*	$0.722 \pm 6.0e-3$	$0.769 \pm 1.5e-3$	$0.691 \pm 2.1e-2$	$0.899 \pm 1.2e-3$
4	2^	Y	$0.721 \pm 4.5e-3$	$0.772 \pm 1.1e-3$	$0.709 \pm 4.1e-3$	$0.888 \pm 2.0e-3$
5	16	Y	$0.720 \pm 5.9e-3$	$0.772 \pm 7.8e-4$	$0.653 \pm 3.7e-3$	$0.884 \pm 1.1e-3$
3	8^	Y*	$0.717 \pm 8.5e-3$	$0.764 \pm 5.9e-3$	$0.731 \pm 2.8e-3$	$0.874 \pm 6.1e-3$
	DINOv3 ViT-L	+ LoRA + MLP	$0.757 \pm 1.9e-2$	-	-	-
	DINOv3 ViT-L	+ MLP	$0.722 \pm 5.2e-3$	$0.763 \pm 4.0e-3$	-	-
	AION-1 ViT-B	+ MLP	$0.702 \pm 1.2e-2$	$0.697 \pm 1.0e-3$	$0.651 \pm 1.3e-3$	$\mathbf{0.910 \pm 4.8e-3}$

Table 7: Evaluation results from joint pre-training (* frozen model first 50 epochs, ^randomly initialized). Baseline comparisons to vision (DINOv3) and multimodal (AION-1) models included.