Beyond Raw Signals: Undecoded Generative Latents as Privileged Synthetic Data

Cristian Sbrolli
Politecnico di Milano
cristian.sbrolli@polimi.it

Nicolas Michel
The University of Tokyo
michel@cvm.t.u-tokyo.ac.jp

Matteo Matteucci
Politecnico di Milano
matteo.matteucci@polimi.it

Toshihiko Yamasaki
The University of Tokyo
yamasaki@cvm.t.u-tokyo.ac.jp

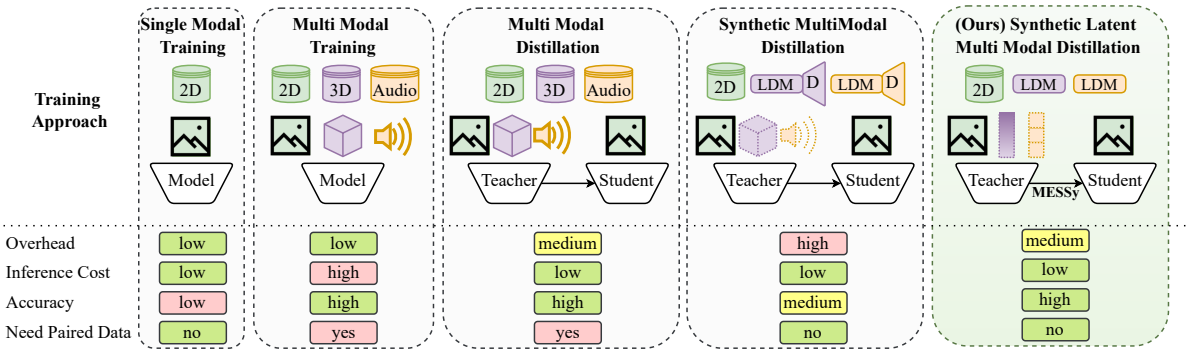

Figure 1. **Multimodal learning paradigms.** Traditional methods require scarce paired datasets or incur high inference costs. Standard synthetic distillation relies on the expensive generative decode-classifier encode loop. Our framework (right) bypasses this entirely, distilling semantically pure undecoded latents directly into a purely visual student by using Multilayer Explicit Simulated Sinesthesia (MESSy).

Abstract

While multimodal integration significantly improves computer vision models, deploying them incurs prohibitive inference costs and requires scarce, perfectly paired datasets. Recent methods address this data bottleneck by synthesizing missing modalities via generative AI, yet they introduce a severe inefficiency: the Decode-Encode Loop. Specifically, information-rich generative latents are decoded into noisy raw signals, forcing the downstream classifier to waste capacity re-encoding them. To bypass this bottleneck, we propose Direct Latent Augmentation (DLA), utilizing undecoded generative latents directly as privileged information. Furthermore, to transfer this dense knowledge to a purely visual student, we introduce Multilayer Explicit Simulated Synesthesia (MESSy). Instead of enforcing rigid representation matching, which forces the student to distort its native visual features to accommodate complex multimodal topologies, MESSy uses a predictive objective to safely internalize these physical priors. Empirical results demonstrate that our framework significantly outperforms raw data augmentation and traditional distillation. Ultimately, our approach yields highly accurate unimodal students with "synesthetic" latent structures that are inherently aligned with physical properties they have never directly observed.

1. Introduction

Multimodal architectures, fusing 2D images with additional modalities, consistently outperform unimodal models across multiple domains. However, processing multiple high-dimensional streams incurs substantial inference costs, and auxiliary modalities are rarely available in the wild. While cross-modal knowledge distillation solves the inference bottleneck by transferring a heavy teacher's knowledge to an image-only student, a paradigm known as Learning Using Privileged Information (LUPI), it fundamentally relies on scarce, perfectly paired multimodal datasets.

Generative AI can overcome this data scarcity by synthesizing missing modalities, an approach recently formalized as Learning Using Generated Privileged Information (LUGPI). Yet, current pipelines overlook a key structural feature: they force semantically dense generative latents through an expensive decoding step to produce "raw" data.

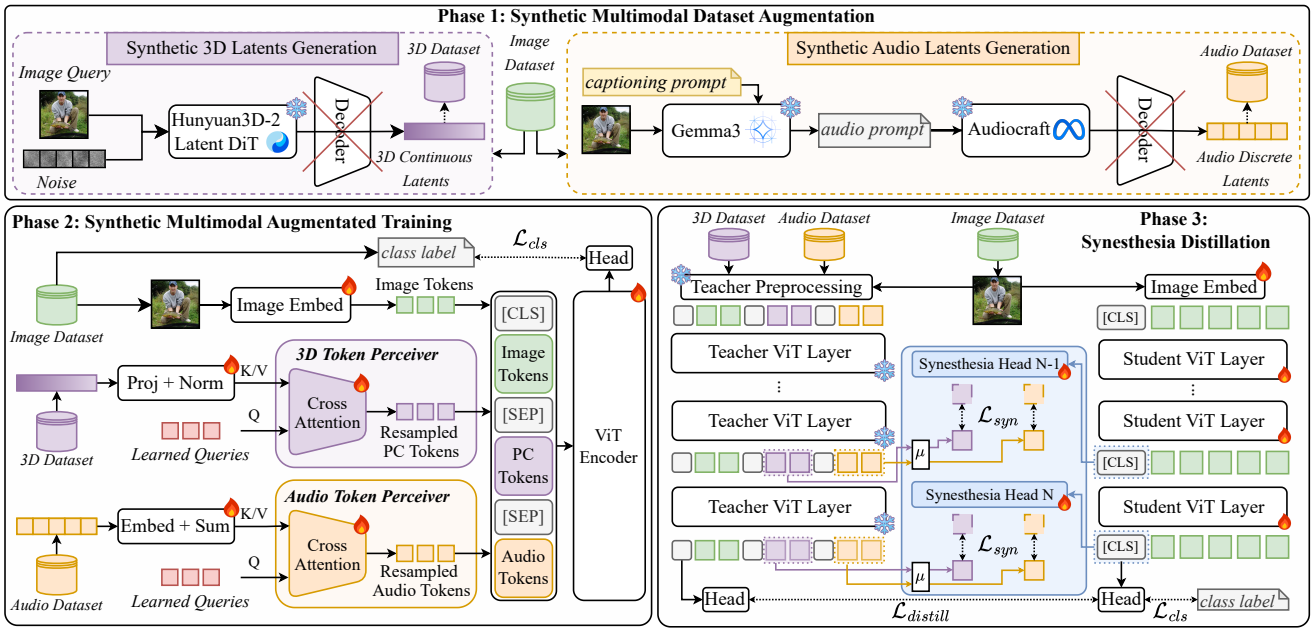

Figure 2. **Phase 1 (Synthetic Latent Augmentation):** We extract continuous 3D latents and discrete audio codes directly from generative models, bypassing the Decode-Encode Loop. **Phase 2 (Multimodal Teacher Training):** A Teacher classifies these augmented triplets, fusing standard image tokens with latents via modality-specific Perceiver blocks. **Phase 3 (MESSy Distillation):** We transfer this knowledge to a purely visual student. Using our MESSy objective, auxiliary heads predict the teacher's multimodal features across multiple layers, preventing capacity interference.

This creates a highly inefficient *Decode-Encode Loop*, injecting generative artifacts and perceptual noise that the downstream classifier must waste capacity re-encoding.

We propose Direct Latent Augmentation (DLA) to bypass this inefficiency, directly utilizing undecoded generative latents as pristine privileged information. However, transferring this dense knowledge to a unimodal student presents a challenge. Alongside standard logit-level knowledge distillation (KD), existing cross-modal methods typically supplement training by minimizing the direct Euclidean (L_2) distance between embedding spaces. This imposes a rigid geometric isomorphism causing *capacity interference*, where forcing a visual student to mimic high-dimensional audio and 3D spaces degrades its native discriminative ability. To resolve this, we introduce Multilayer Explicit Simulated Synesthesia (MESSy). Rather than identically matching spaces, MESSy employs lightweight predictive heads, allowing the student to predict the teacher's auxiliary features from its own visual tokens. This ensures the student internalizes physical priors without collapsing its native visual manifold.

2. Prior Work.

Traditional Learning Using Privileged Information (LUPI) [17] bridges multimodal training and unimodal inference via cross-modal distillation [6, 7], but it fundamentally relies on perfectly aligned real-world datasets that are expensive to scale. Recent generative augmentation pipelines, such as Learning Using Generated Privileged Information (LUGPI) [9], circumvent this by synthesizing auxiliary data. However, they operate exclusively in the raw data domain (text and image). This not only ignores critical physical modalities like audio and 3D, but the decoding step also injects perceptual noise that wastes downstream parameter capacity. Finally, to stabilize optimization, feature-level distillation often aligns multiple hidden layers [12, 13, 16]. In cross-modal LUPI, this is typically enforced by rigidly minimizing the L_2 distance between embeddings. We demonstrate that in asymmetric architectures, this strict geometric isomorphism induces *capacity interference*, limiting the visual student from fully exploiting the cross-modal distillation and motivating our predictive MESSy objective.

3. Method

3.1. Multimodal Latent Augmentation

Standard synthetic augmentation creates a deterministic Markov chain from the generated latent Z to the final classification features F: $Z \xrightarrow{\mathcal{D}} \hat{X} \xrightarrow{\mathcal{E}} F$. By the Data Processing Inequality, mutual information I with the label Y cannot increase: $I(Y;Z) \geq I(Y;\hat{X}) \geq I(Y;F)$. Decoding $\mathcal{D}$ injects perceptual noise that degrades the semantic signal. To

avoid this and fully leverage synthetic signals, we intercept the generative pipeline to extract latents directly.

Given an RGB image query X_v, we extract 3D and audio latents without synthesizing raw signals (Phase 1, Fig. 2). For 3D, we employ Hunyuan3D-2 Latent DiT [15] to obtain continuous latents $Z_{3D} \in \mathbb{R}^{N_{3D} \times D_{3D}}$. For audio, we use Gemma3 [14] to create acoustic captions, strictly filtering out class names, and pass this to Audiocraft [1], extracting undecoded discrete codes $Z_{aud} \in \mathbb{Z}^{N_{aud}}$. We report full generation details in the supplementary material.

To process these heterogeneous latents (Phase 2, Fig. 2), we design a Multimodal Teacher Network $\mathcal{T}$. Standard image patches E_v are fused with Z_{3D} and Z_{aud} using modality-specific Perceiver cross-attention blocks, which compress the variable-length latents into a fixed budget of resampled tokens (E_{3D}, E_{aud}). The final joint sequence $E_{input} = [[\texttt{CLS}], E_v, [\texttt{SEP}_1], E_{3D}, [\texttt{SEP}_2], E_{aud}]$ is processed by a ViT Encoder and optimized via cross-entropy.

MESSy Distillation In the final phase, our objective is to transfer the multimodal representations of the frozen Teacher $\mathcal{T}$ into a lightweight, purely visual Student $\mathcal{S}$. As discussed, standard cross-modal distillation enforces a strict geometric isomorphism (minimizing the L_2 distance between embeddings). This forces the lower-capacity student to distort its optimal visual decision boundaries to geometrically accommodate the missing modalities, inducing destructive capacity interference [18].

To resolve this bottleneck, we propose Multilayer Explicit Simulated Synesthesia (MESSy) (Phase 3 of Fig. 2). Rather than enforcing rigid L_2 spatial equality, MESSy requires the student to encode only the sufficient statistics necessary to reconstruct the auxiliary modalities. Let s^l denote the student's visual classification token at layer l, and let t_m^l denote the teacher's average-pooled token for modality $m \in \mathcal{M}$, where $\mathcal{M} = \{aud, 3D\}$ is the set of auxiliary modalities. We attach modality-specific lightweight projection heads H_m to the student. The MESSy objective minimizes the MSE over the deepest K layers:

$$\mathcal{L}_{MESSy} = \sum_{l=L-K}^{L} \sum_{m \in \mathcal{M}} ||H_m(s^l) - t_m^l||_2^2 \qquad (1)$$

During inference, the auxiliary heads H_m are discarded, leaving an enriched visual latent space with zero additional inference cost. We ablate K in the supplementary materials.

4. Experiments

Datasets. We evaluate on Imagenette2-320 [8], Caltech101 [5], and ImageNet-100 [3, 11], constructing our synthetic datasets by passing the original visual splits through our generative pipeline. Focusing on these benchmarks allows us to rigorously ablate modality combinations and perform expensive raw-vs-latent comparisons. . We execute 3 independent runs per experiment.

Table 1. Teacher representation efficiency. Directly utilizing generative latents outperforms fully decoded raw synthetic data.

Dataset	Image Only	Raw	Latents (Ours)
Imagenette2	79.63 ± 0.32	88.28 ± 0.65	$\mathbf{93.10 \pm 0.19}$
Caltech101	55.10 ± 0.27	59.08 ± 1.96	$\mathbf{66.47 \pm 1.42}$
ImageNet-100	61.57 ± 0.53	65.40 ± 0.90	$\mathbf{71.59 \pm 0.45}$

Table 2. Modality Contribution Analysis. Top-1 Accuracy (%) of the Teacher network under different latent combinations.

Modalities	Imagenette2	Caltech101	ImageNet-100
Image Only	79.63 ± 0.32	55.10 ± 0.27	61.57 ± 0.53
Audio Only	54.61 ± 0.85	23.61 ± 1.10	13.07 ± 0.54
3D Only	68.53 ± 0.43	37.23 ± 2.84	39.20 ± 0.73
Image + Audio	87.37 ± 0.24	59.58 ± 0.83	65.76 ± 0.04
Image + 3D	87.80 ± 0.15	65.15 ± 0.24	70.03 ± 0.53
Full (Image+3D+Audio)	$\mathbf{93.10 \pm 0.19}$	$\mathbf{66.47 \pm 1.42}$	$\mathbf{71.59 \pm 0.45}$

Architecture & Training. Teacher ($\mathcal{T}$) and Student ($\mathcal{S}$) utilize the ViT-B/16 architecture. The Teacher employs Perceiver resamplers (100 tokens per modality) and an aggressive modality dropout rate of $p = 0.8$. The MESSy distillation objective extracts features from the last $K = 3$ layers. For the Raw synthetic data baseline, the Teacher encodes 3D point clouds using a PointNet++ style tokenizer, and audio Mel-spectrograms via a standard 2D convolutional patch embedding.

Teacher Performance: The Latent Advantage. Before evaluating downstream accuracy, we quantify the computational savings of bypassing the decode-encode loop during offline dataset augmentation. Bypassing the mesh decoder for 3D generation reduces inference time by 87.7% and saves over 600 MB of peak VRAM. Similarly, extracting discrete audio tokens prior to waveform decoding yields a 20.6% reduction in generation time and cuts peak VRAM usage by 23.8% (full details in supplementary materials).

Crucially, these efficiency gains translate to improved downstream classification accuracy (Table 1). Training the Teacher directly on generative latents significantly outperforms fully decoded raw synthetic data. On Caltech101, the Latent Teacher achieves an absolute improvement of $+7.39\%$ over the Raw Teacher and $+11.37\%$ over the Image-Only baseline. These consistent gains validate that generative decoding injects perceptual noise that degrades the semantic signal, while undecoded latents provide highly discriminative privileged information.

Orthogonal Modality Contributions. To validate that the extracted generative latents provide genuine, complementary physical priors rather than acting as a generative shortcut, we ablate the contribution of individual modalities (Table 2). 3D geometry and audio provide orthogonal su-

Method	Imagenette2	Caltech101	ImageNet-100
Image Baseline	79.63 ± 0.32	55.10 ± 0.27	61.57 ± 0.53
MM Teacher (UB)	93.10 ± 0.19	66.47 ± 1.42	71.59 ± 0.45
KD	80.66 ± 0.20	56.49 ± 0.42	63.97 ± 0.59
LUGPI	81.85 ± 0.51	60.12 ± 0.76	65.70 ± 0.68
MESSy (Ours)	$\mathbf{84.23 \pm 0.40}$	$\mathbf{63.52 \pm 0.91}$	$\mathbf{68.54 \pm 0.83}$

Table 3. Top-1 Accuracy (%) comparison of unimodal distillation paradigms. Students evaluate purely on visual inputs at inference.

pervision. On ImageNet-100, integrating either 3D latents (Image+3D) or audio latents (Image+Audio) independently improves upon the Image-Only baseline. Fusing all modalities yields the highest performance across all benchmarks.

Crucially, we also evaluate networks trained solely on the isolated synthetic modalities to ensure the model is not simply exploiting a generative shortcut. We find that the synthetic modalities perform poorly on their own, falling far below the real Image-Only baseline. This confirms that the generative latents do not trivially leak the target label; instead, they act as true complementary physical priors that help the visual encoder resolve ambiguities.

4.1. MESSy Distillation Efficacy

We evaluate transferring our robust multimodal knowledge to a unimodal Student ($\mathcal{S}$) in Table 3. Our Latent-Distilled Student (MESSy) achieves the best unimodal performance, outperforming the image-only baseline by $+8.42\%$ on Caltech101 and $+6.97\%$ on ImageNet-100.

We compare MESSy against two baselines: standard Knowledge Distillation (KD) and LUGPI. KD yields marginal gains, not capturing complex multimodal semantics. LUGPI enforces strict L_2 representation matching, which improves accuracy but bottlenecks the student. By relaxing this rigid spatial constraint using predictive multi-layer heads, MESSy outperforms LUGPI by a further $+3.40\%$ on Caltech101 and $+2.84\%$ on ImageNet-100. This confirms our hypothesis (Sec. 3.1): strict geometric isomorphism causes capacity interference, whereas MESSy allows the student to internalize multimodal statistics without compromising its optimal visual manifold.

4.2. Emergent Synesthesia

To confirm the emergence of artificial synesthesia, the student's visual latent space must inherently reflect the topological structure of the unobserved physical modalities. We verify this by measuring the Normalized Mutual Information (NMI) between the student's visual clusters and ground-truth multimodal clusters. To capture both foundational physics and high-level semantics, we construct these clusters using two feature types: *Hand-Crafted (HC)* (e.g., acoustic MFCCs [2], 3D covariance metrics) and *Deep* embeddings (via CLAP [4] and PointNet [10]). To account for

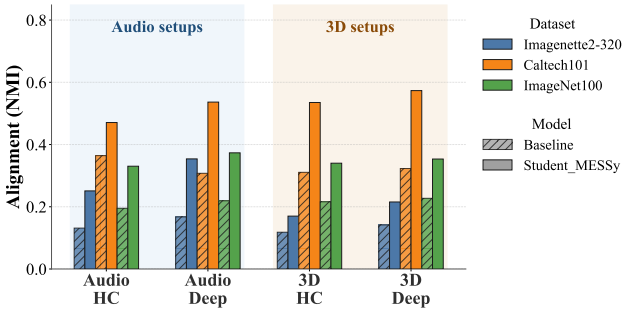

Figure 3. Emergent Alignment (NMI) between the models' visual latent spaces and the hidden modality topologies.

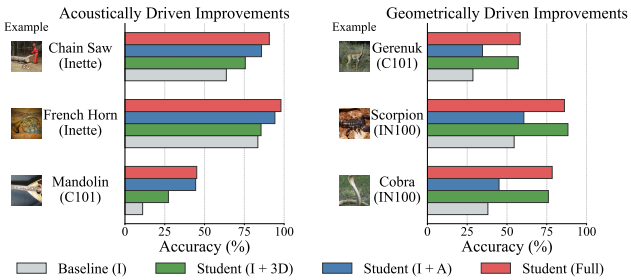

Figure 4. Qualitative analysis on class average improvement.

varying semantic diversity, we scale the number of clusters relative to the dataset's class vocabulary, using 50 clusters for Imagenette2 and 150 for Caltech101 and ImageNet-100, ensuring low overlap with the original class labels (we report full overlap analisys in supplementary materials).

As shown in Fig. 3, the baseline visual model exhibits poor structural alignment with acoustic and geometric spaces. In stark contrast, our MESSy student exhibits massive NMI alignment across both Deep and HC features.

4.3. Qualitative Analysis: Disentangling Priors

We performed a qualitative class-level analysis comparing students distilled from bimodal and Full teachers (Fig.4). Looking at the top improving classes (avg accuracy), we find that our model selectively leverages the simulated priors: acoustically driven improvements dominate classes with distinct sound profiles (e.g., *Chain Saw*, *French Horn*), while geometrically driven improvements dominate species with complex 3D morphological traits (e.g., the *Gerenuk*).

5. Conclusion

We introduced Direct Latent Augmentation (DLA) to bypass the inefficient decode-encode loop of generative multimodal training, utilizing pristine latents directly as privileged information. To seamlessly transfer this dense knowledge, our MESSy distillation objective allows a purely visual student to internalize acoustic and geometric priors without suffering from capacity interference.

References

[1] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[2] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980. 4

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4

[5] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 3

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[7] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for rgb-d detection. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 5032–5039. IEEE, 2016. 2

[8] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet. https://github.com/fastai/imagenette, 2019. 3

[9] Rafael-Edy Menadil, Mariana-Iuliana Georgescu, and Radu Tudor Ionescu. Learning using generated privileged information by text-to-image diffusion models. In *International Conference on Pattern Recognition*, pages 423–438. Springer, 2024. 2

[10] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[11] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[12] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. 2

[13] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4323–4332, 2019. 2

[14] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek

Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. 3

[15] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 3

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[17] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 2

[18] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. 3