Exploring zero-shot translation in multi-modal vision models

Johann Wenckstern (389139)
CS-503 Final Project Report

***Abstract*—Recent advancements in multi-modal learning have produced powerful foundation models capable of handling diverse data types, but these models depend heavily on large aligned multi-modal datasets, which are scarce in fields like biomedicine. In this work, we study the question whether zero-shot modality translation, i.e. synthesizing one modality from another in the absence of aligned training pairs, is possible. We retrain 4M, a recent multi-modal generative model, on a dataset spanning seven modalities while withholding aligned samples of different modality pairs. Our experiments show that these models can perform zero-shot translation between visual and textual modalities, provided that shared binding modalities are available during training. However, performance drops as binding modalities are removed, and translation from images to captions remains notably limited.**

I. Introduction

Recent efforts in multi-modal learning have led to the development of powerful multi-modal foundation models, which are capable of processing and generating a broad spectrum of data modalities—including RGB images, geometric and semantic representations [1], textual inputs [2], [3], and intermediate neural network feature maps [4]. These models demonstrate impressive cross-modal capabilities, such as inferring missing modalities from different combinations of input modalities.

However, training such models relies on the availability of large-scale aligned multi-modal datasets. While uni-modal data, is abundant and easy to obtain, the availability of aligned samples across two or more modalities is limited. This data scarcity is especially noticeable in specialized fields such as biomedicine, where acquiring, labeling, and aligning data from multiple imaging technologies (e.g., histology, spatial proteomics, spatial transcriptomics) is expensive or even technically impossible. Past work has sought to address this challenge in two primary ways: by training on fragmented datasets where only some modality pairs are available [3], or by employing pseudo-labeling with expert models to generate aligned synthetic data samples [4], [5].

For computer vision, this pseudo-labeling has proven to be an effective strategy, due to the rich information content of RGB images, which enables almost deterministic translation to other modalities. Consequently, the different synthesized modalities from a single image are still aligned. In domains like biomedicine, there may be no single, "binding" base modality that adequately captures the necessary information for all others. Instead, different tests capture fundamentally distinct biological signals. For example, histological imaging reveals tissue structure, while spatial proteomics and transcriptomics capture molecular and cellular information. These modalities are not trivially aligned, and information in one often cannot be deterministically inferred from another. This motivates the central research question of this work:

Given n modalities, $M_1, M_2, ..., M_n$, for which aligned data is available for all pairs of modalities except (M_a, M_b) for some $i \neq j$ can we train a generative model capable of synthesizing M_b directly from M_a and vice-versa?

Beyond the practical motivation, there is a deeper conceptual reason for studying this type of zero-shot modality translation: success in this setting would indicate that a model has learned to internally construct joint unified representations across all modalities — an open and ambitious goal in multi-modal learning. In this sense, zero-shot translation can serve as both a benchmark and a tool for probing whether such unified latent spaces have been learned.

To address the posed question, we retrain 4M, a recent multi-modal generative model—on a dataset covering seven visual and textual modalities, while systematically excluding data samples containing specific modality pairs (M_a, M_b). Our results show that, for visual modality pairs, the model can synthesize M_b directly from M_a with only minor performance degradation compared to the non-zero-shot setting. Through ablation studies, we further demonstrate that this zero-shot capability relies on the presence of binding modalities with aligned samples; as the number of binding modalities decreases, so does zero-shot performance. Finally, we evaluate zero-shot translation between visual and textual modalities and find that zero-shot capabilities emerge also between such.

II. Related Work

Viewing different languages as modalities, the ability for zero-shot translation has been observed in NLP as an emergent capability of language models [6]. However, this phenomenon has been little studied in the context of vision and multi-modal research. Recent multi-modal generative models such as 4M [4] and Unified-IO [3] do not benchmark this capability. Unified-IO trains on fragmented datasets, whereas 4M circumvents this by creating pseudo-labels using expert models. Unified-IO does, however, investigate

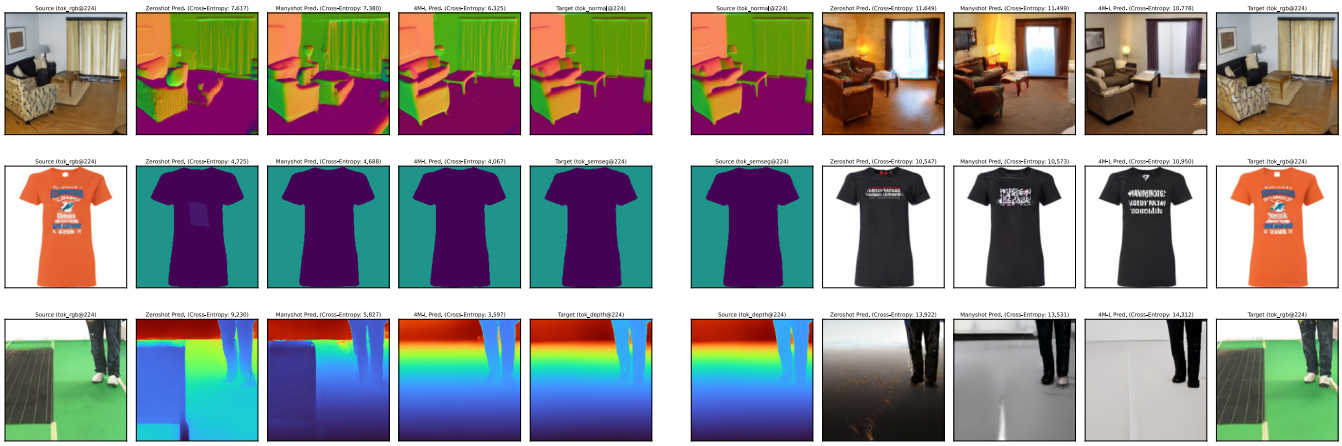

Figure 1. **Visual Modalities.** Comparison of zero-shot translation examples with corresponding many-shot and 4M-L predictions.

how omitting certain modality pairings during training affects performance on other modality translation tasks at test time. To our knowledge, the only work directly addressing the same general zero-shot translation question as ours is LoReTTa [7]. Despite initially posing the same question, LoReTTa takes a different approach. Their proposed method, termed transitive pretraining, involves a two-step process: first training only on aligned modality pairs, then using the pre-trained model to pseudo-label missing pairings during a second phase of training. In the context of representation learning, ImageBind explores how aligning various modalities with only RGB images can be used to learn shared representations across modalities. Our work is distinct from this, as it focuses on a generative task rather than on global representation learning. Finally, learning a unified joint representation of multi-modal data, also known as a platonic representation [8], is an active area of research. Our work establishes a novel criterion for evaluating such representations and demonstrates that those learned by 4M exhibit emergent capabilities that satisfy it.

III. Method

We investigate the zero-shot cross-modal translation capabilities of 4M [4], a multi-modal generative vision model. We selected the 4M framework as it supports a wide range of image- and sequence-based modalities, and permits to train and infer on flexible combinations of input and target modalities.

The original 4M model was trained on CC12M [9], pseudo-labeled using expert models to include for every sample seven modalities, namely RGB, Depth, Normal Surfaces, Semantic Segmentation, Bounding Boxes, Captions, CLIP features [10]. We refer to this set of modalities in the following as $\mathcal{M}$. Further, for each training sample, the number of input and target tokens per modality were sampled from a Dirichlet distribution with uniform weights, resulting in a mix of all modalities.

To evaluate zero-shot translation between a modality pair (M_a, M_b), we retrain 4M using a modified token-sampling strategy on the multi-modal CC12M dataset. For each training sample, we draw the input and target token counts from a Dirichlet distribution with weights: either $\alpha \in \mathbb{R}^7$ where $\alpha_i = 0.5\delta_{ia}$ or $\beta \in \mathbb{R}^7$ where $\beta_i = 0.5\delta_{ib}$. Crucially, we fix the sampled distribution for each sample across training epochs, as suggested by feedback, to prevent information leakage. This effectively partitions the pseudo-labeled CC12M dataset into two disjoint subsets: one excluding M_a and the other excluding M_b.

By default, we adopt the 4M-B architecture and train the model for 100 billion tokens to reduce computational costs. In contrast to the original 4M paper, we exclude non-quantized RGB tokens from training. Including both quantized and non-quantized RGB representations would otherwise unnecessarily complicate zero-shot evaluations involving RGB. Apart from these modification, our training setup follows the original 4M procedure [4].

IV. Experiments

A. Zero-shot translation of visual modalities

In our first experiment, we evaluate the zero-shot translation capabilities of 4M across pairs of visual modalities:

- RGB $\leftrightarrow$ Normal
- RGB $\leftrightarrow$ Depth
- RGB $\leftrightarrow$ Semantic Segmentation

Since RGB is the most information-rich modality, we assume that translating from RGB to another modality is relatively easy, whereas translating to RGB is more challenging. By including RGB in each modality pair, we are able to assess the performance on both relatively easy and difficult translation tasks. For each of these pairs, we train a 4M-B instance following the procedure outlined in Section III.

We evaluate translation performance on a held-out test set both qualitatively and quantitatively. Qualitatively, we ana-

lyze representative generation examples; quantitatively, we compute the average cross-entropy between model-predicted token probabilities and the pseudo-labeled ground truth over 2500 samples. The 4M framework supports a wide range of generation configurations. The specific generation schedules used in our experiments are detailed in Appendix Table III.

As a baseline we consider a 4M-B instance trained with an equal token budget of 100 billion on all seven modalities with constant Dirichlet weights $\gamma \in \mathbb{R}^7$ where $\gamma_i = 0.5$. We will refer to this baseline in the following as the *many-shot* setting. Further, to estimate potential performance improvements through scaling, we compare also to 4M-L with weights pretrained on 500 billion tokens provided by the original 4M study [4].

Our findings indicate that 4M is generally capable of zero-shot predictions across all tested tasks. However, the quality of these predictions varies considerably between individual samples. Selected successful examples are shown in Fig. 1 A more extensive and diverse set of examples can be found in the Appendix in Figures 4-5. Qualitatively, we observe that across all tasks the zero-shot performance is slightly worse than in the many-shot setting. The average cross-entropy scores, indicated in Table I, support this observation. Further, we remark that this visual performance gap appears comparable to the difference observed between the many-shot 4M-B model and the larger 4M-L model. This suggests that scaling the model and training could further improve zero-shot performance.

Task	**0-Shot**	**Many-Shot**	**4M-L**
RGB → Normal	8.211	7.026	6.637
RGB → Depth	8.284	7.021	6.574
RGB → Sem. Seg.	6.807	6.347	5.997
Normal → RGB	10.327	10.400	10.333
Depth → RGB	12.276	10.666	10.575
Sem. Seg. → RGB	11.917	11.785	11.446

Table I
CROSS-ENTROPY LOSSES FOR ZERO-SHOT TRANSLATION OF VISUAL MODALITIES.

B. Zero-shot translation under systematic reduction of modalities

The results of previous experiments suggest that the multi-modal masked training objective encourages the model to learn a regularized latent representation of the observed modalities $\mathcal{M}$, such that even the two modalities M_a and M_b without aligned training samples yield structurally compatible representations. We hypothesize that this is possible because the training dataset contains further modalities $\mathcal{M} \setminus \{M_a, M_b\}$, for which aligned samples with both M_a and M_b are available. These modalities might be implicitly utilized to align the latent representations of M_a and M_b, comparable, to the explicit binding training objective used in ImageBind [11]. In the previous experiment, we had five such binding modalities. Next, we study the robustness of the zero-shot translations, if less binding modalities are available during training.

To this aim, we choose to focus on the task of zero-shot translating from $M_a = \text{RGB}$ to $M_b = \text{Depth}$ and retrain 4M-B on the following sets of binding modalities:

- No Binding Mod:
  $\mathcal{M} \setminus \{M_a, M_b\} = \emptyset$
- One Binding Mod.:
  $\mathcal{M} \setminus \{M_a, M_b\} = \{\text{Normal}\}$
- Three Binding Modalities:
  $\mathcal{M} \setminus \{M_a, M_b\} = \{\text{Normal, Sem. Seg., CLIP}\}$
- Five Binding Modalities:
  $\mathcal{M} \setminus \{M_a, M_b\} = \begin{Bmatrix}\text{Normal, Sem. Seg.,} \\ \text{CLIP, Capt., BBox}\end{Bmatrix}$,

where M_a is RGB and M_b is Depth.

Figures 2 and 10 present a comparison of zero-shot predictions from these runs, while Table II reports the corresponding average cross-entropy scores on the test set.

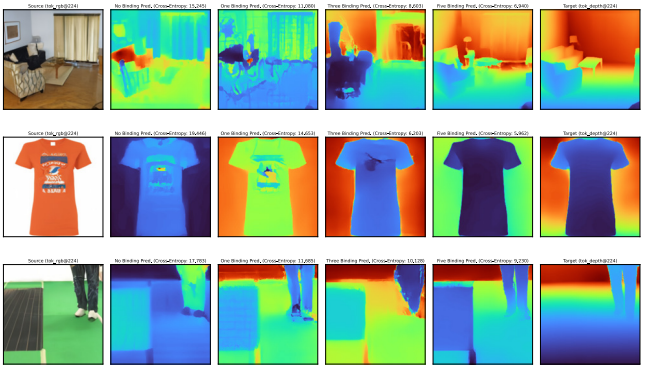

Figure 2. **RGB → Depth** zero-shot predictions for different numbers of binding modalities seen during training.

#Binding Modalities	**0**	**1**	**3**	**5**
Average Cross-Entropy	14.128	11.810	9.209	8.265

Table II
AVERAGE CROSS-ENTROPY FOR ZERO-SHOT RGB TO DEPTH TRANSLATION GIVEN DIFFERENT NUMBERS OF BINDING MODALITIES

The illustrated examples, along with their individual and average test cross-entropy scores, show a clear and consistent trend: incorporating more binding modalities during training leads to improved zero-shot translation performance. This is particularly noteworthy given that, with an increasing number of binding modalities, the target modality is observed less frequently during training due to the fixed total token budget of 100 billion tokens across all runs.

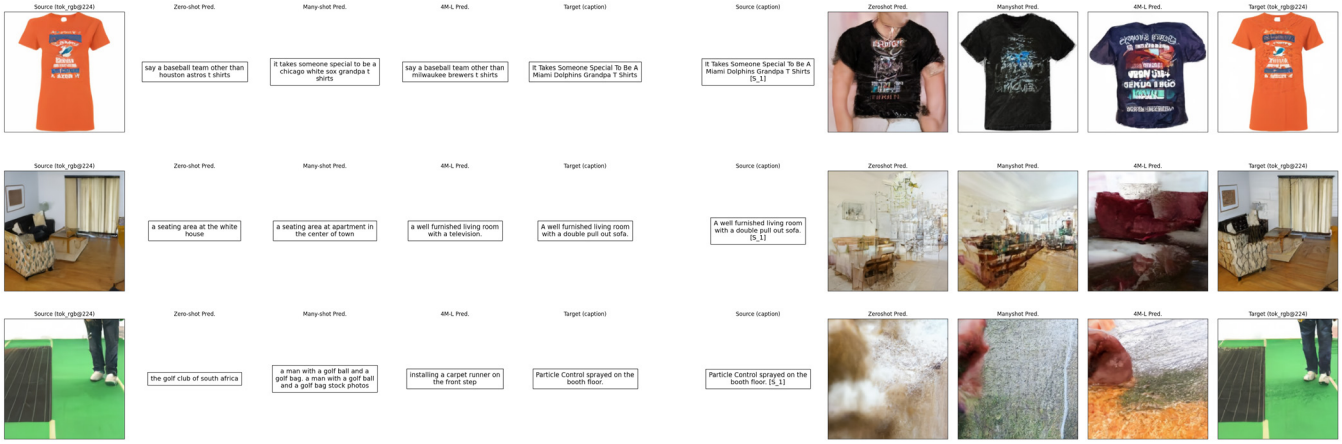

Figure 3. **RGB ↔ Caption**. Comparison of zero-shot translation examples with corresponding many-shot and 4M-L predictions.

The results of this experiment further suggest that zero-shot performance may be improved by incorporating additional modalities.

C. Zero-shot translation between RGB and Caption

Thus far, our experiments have focused on the zero-shot translation capabilities between visual modalities. These modalities share inherent structural and semantic similarities, for example, they are all spatially resolved and capture object silhouettes. We now turn our attention to evaluating zero-shot translation between visual and textual modalities, where the modality gap is presumed to be more substantial. To this end, we retrain 4M while withholding all samples that contain aligned RGB images and caption data, following the same procedure as in our previous experiments.

Figures 3, 11 and 12 show zero-shot translation results for RGB→Caption and Caption→RGB, again compared to the many-shot setting and 4M-L. The employed generation schedules are indicated in the Appendix.

For Caption→RGB, we observe that the zero-shot predictions depict identifiable scenes that are semantically aligned with the input captions, although the visual quality of these images is quite poor. Image quality improves marginally in the many-shot setting and notably in the predictions from 4M-L. For RGB→Caption, both zero-shot and many-shot predictions yield diverse results. Some generated captions are meaningful and well-aligned with the input image, while others are nonsensical. Only the captions generated by 4M-L consistently demonstrate decent quality, suggesting that this task benefits significantly from increased model scale.

V. Conclusion and Limitations

In this project, we demonstrate that zero-shot translation capabilities between visual and textual modality pairs without aligned samples emerge when training a multi-modal generative model on a diverse set of modalities. Through an ablation study, we show that this capability critically depends on the presence and quantity of intermediary binding modalities for which paired data is available. We also observe that zero-shot predictions are slightly lower in quality compared to many-shot settings, but find indications that performance can be improved through further scaling of the model and training regime, as well as by incorporating additional modalities. Moreover, our work introduces a new criterion for assessing whether a multi-modal model has learned a unified joint representation.

A key limitation, and an opportunity for future work, is that our analysis has been conducted solely on the pseudo-labeled CC12M dataset. Consequently, the practicality of our findings must still be demonstrated on naturally fragmented, real-world multi-modal datasets without any pseudo-labeling, which is the ultimate goal we aim to address.

VI. Contributions and Acknowledgment

This project was conducted by a single author. The author thanks Zhitong Gao and Roman Bachmann for helpful discussions and feedback, and Mingqiao Ye for proving the utilized dataset.

References

[1] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *European Conference on Computer Vision*. Springer, 2022, pp. 348–367.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[3] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," *arXiv preprint arXiv:2206.08916*, 2022.

[4] D. Mizrahi, R. Bachmann, O. Kar, T. Yeo, M. Gao, A. Dehghan, and A. Zamir, "4m: Massively multimodal masked modeling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 363–58 408, 2023.

[5] R. Bachmann, O. F. Kar, D. Mizrahi, A. Garjani, M. Gao, D. Griffiths, J. Hu, A. Dehghan, and A. Zamir, "4m-21: An any-to-any vision model for tens of tasks and modalities," *Advances in Neural Information Processing Systems*, vol. 37, pp. 61 872–61 911, 2024.

[6] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[7] M. Tran, Y. Dicente Cid, A. Lahiani, F. Theis, T. Peng, and E. Klaiman, "Training transitive and commutative multimodal transformers with loretta," *Advances in Neural Information Processing Systems*, vol. 36, pp. 42 918–42 931, 2023.

[8] M. Huh, B. Cheung, T. Wang, and P. Isola, "The platonic representation hypothesis," *arXiv preprint arXiv:2405.07987*, 2024.

[9] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[11] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 180–15 190.

VII. APPENDIX

Task	Sampling Scheme	Sampling Steps	Temp.	CFG Scale	CFG Schedule	Top P	Top K
RGB → Normal	ROAR	1	0.01	2.0	Constant	0.8	0
RGB → Depth	ROAR	1	0.01	2.0	Constant	0.8	0
RGB → Sem. Seg.	ROAR	1	0.01	2.0	Constant	0.8	0
Normal → RGB	ROAR	8	0.01	2.0	Constant	0.8	0
Depth → RGB	ROAR	8	0.01	2.0	Constant	0.8	0
Sem. Seg. → RGB	ROAR	8	0.01	2.0	Constant	0.8	0
RGB → Caption	AR	-	0.3	1.0	Constant	0.8	0
Caption → RGB	ROAR	25	3.0	2.0	Constant	0.8	0

Table III
GENERATION SCHEDULES

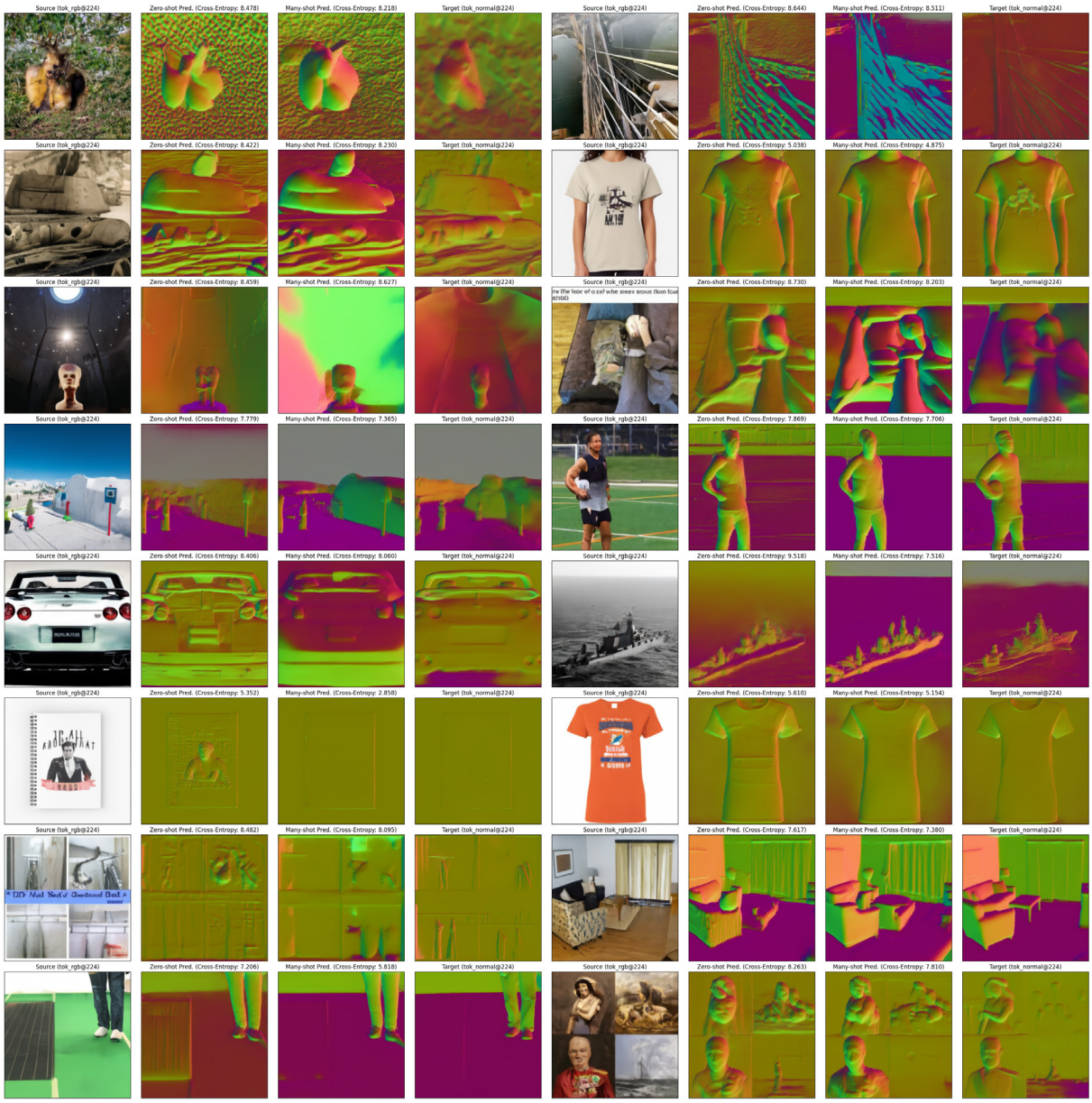

Figure 4. **RGB → Normal**. For each sample, we show from left to right: decoded tokenized RGB input, zero-shot normal prediction, many-shot normal prediction, and the decoded ground truth.

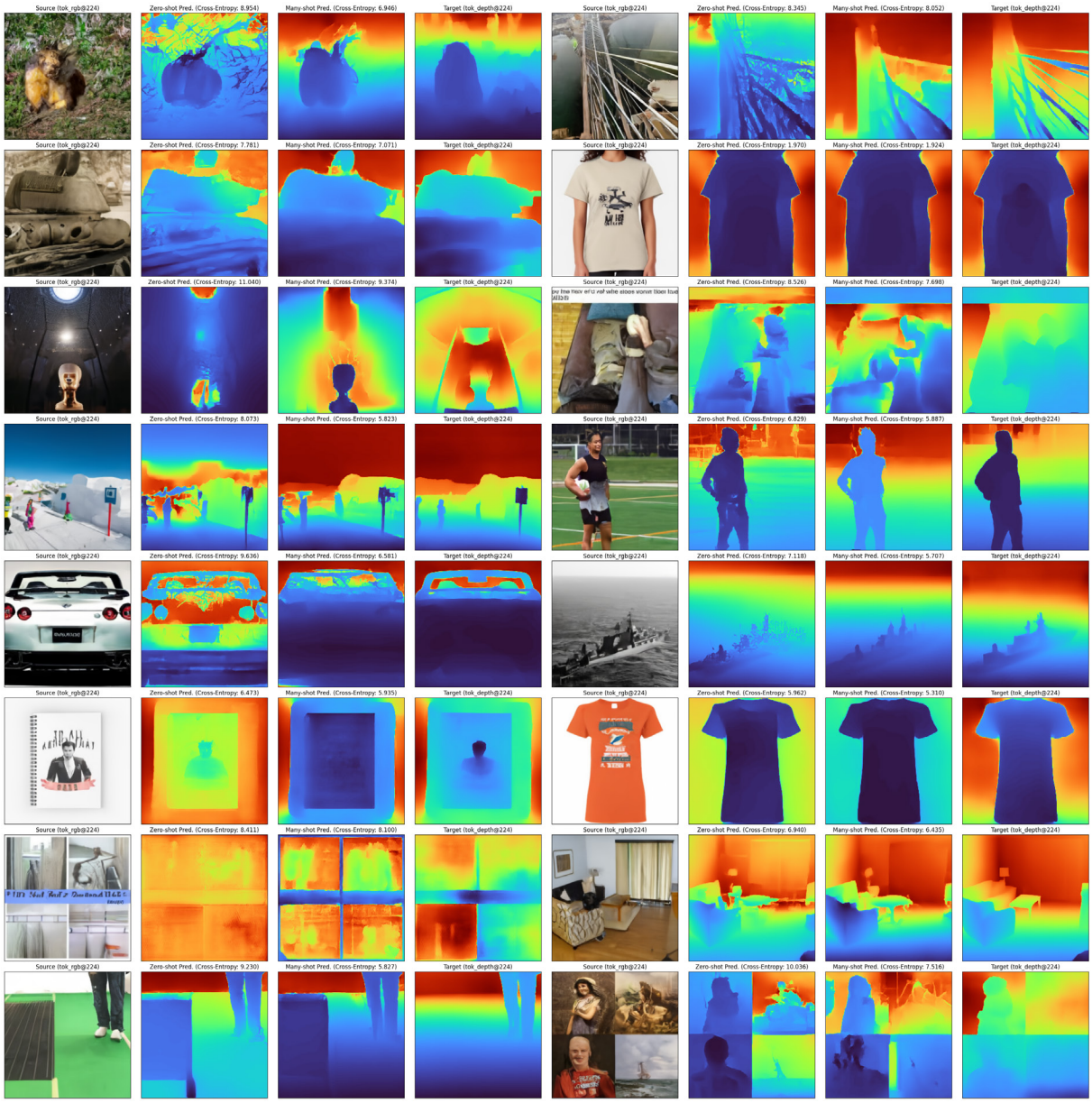

Figure 5. **RGB → Depth**. For each sample, we show from left to right: decoded tokenized RGB input, zero-shot depth prediction, many-shot depth prediction, and the decoded ground truth.

Figure 6. **RGB → Sem. Seg.**. For each sample, we show from left to right: decoded tokenized RGB input, zero-shot semantic segmentation, many-shot semantic segmentation, and the decoded ground truth.

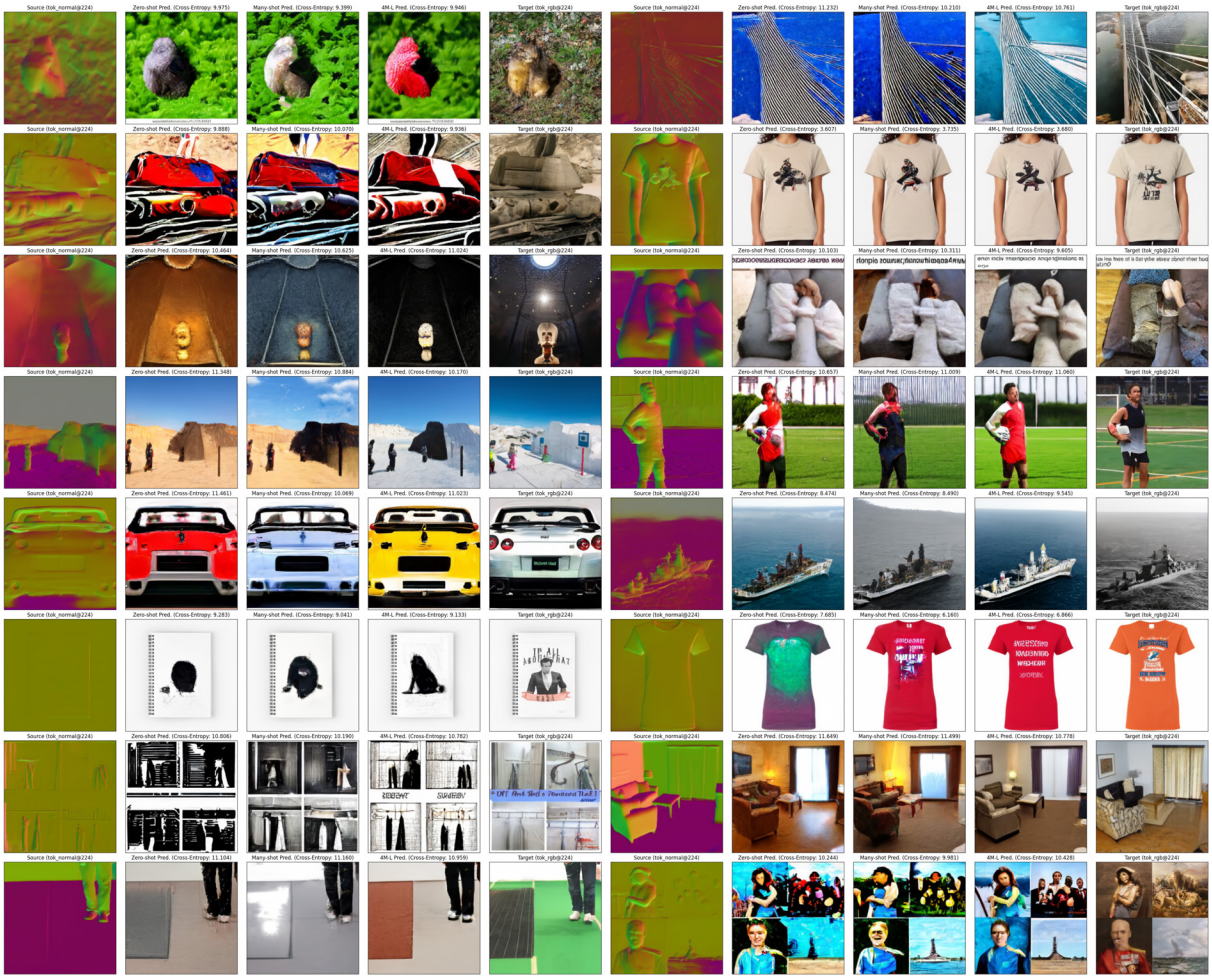

Figure 7. **Normal → RGB**. For each sample, we show from left to right: decoded tokenized Normal input, zero-shot RGB prediction, many-shot RGB prediction, and the decoded ground truth.

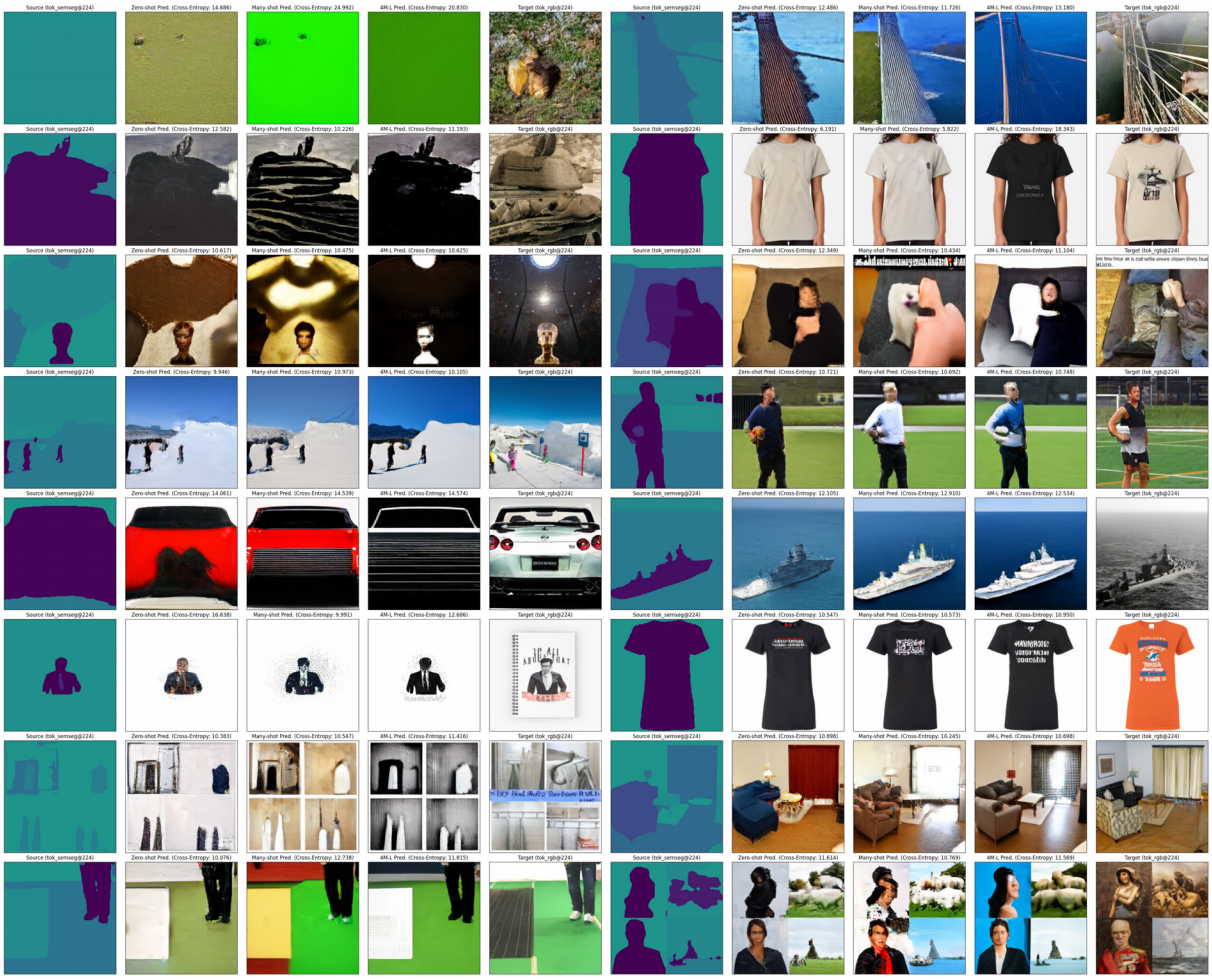

Figure 8. **Sem. Seg. → RGB**. For each sample, we show from left to right: decoded tokenized Sem. Seg. input, zero-shot RGB prediction, many-shot RGB prediction, and the decoded ground truth.

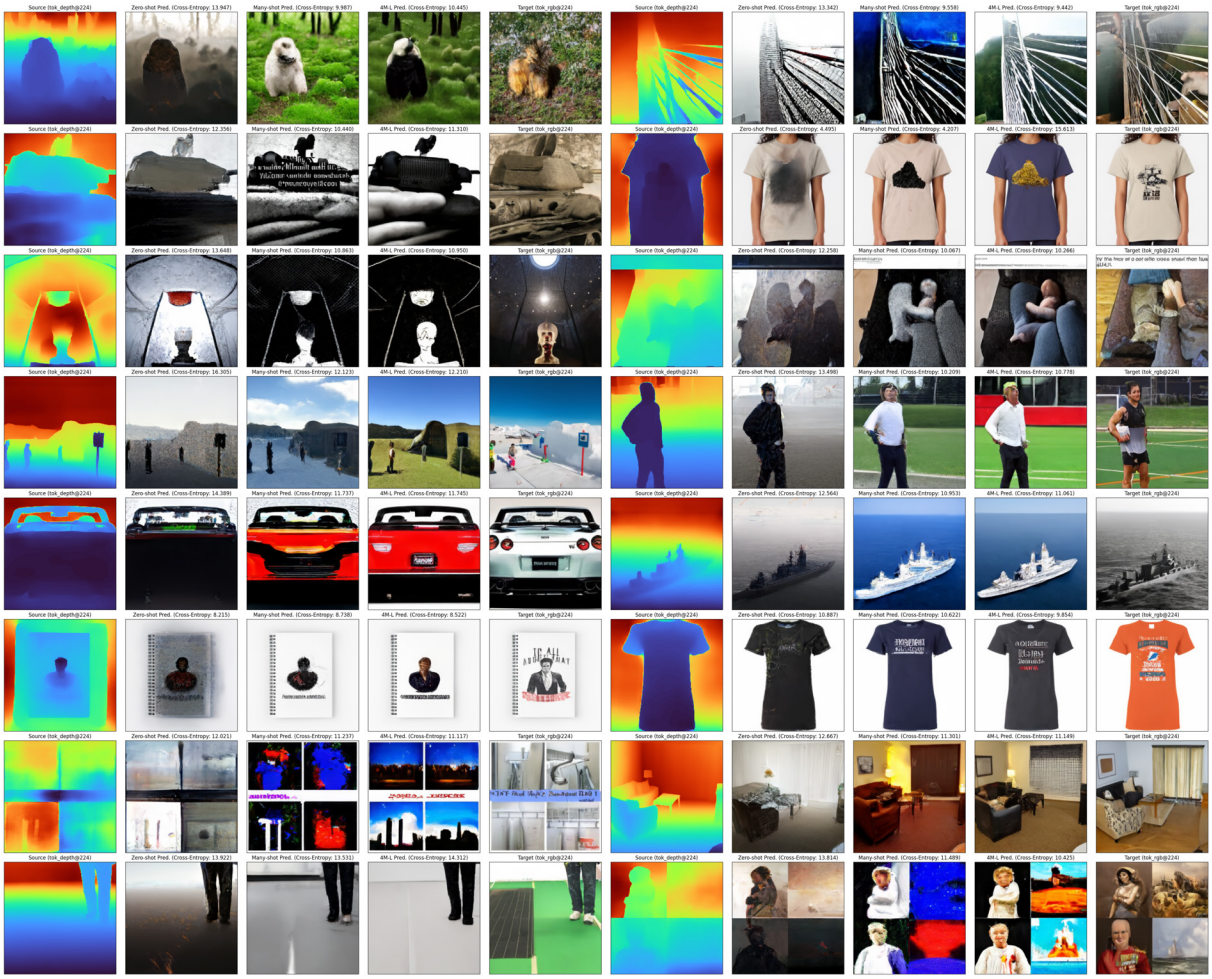

Figure 9. **Depth → RGB**. For each sample, we show from left to right: decoded tokenized Depth input, zero-shot RGB prediction, many-shot RGB prediction, and the decoded ground truth.

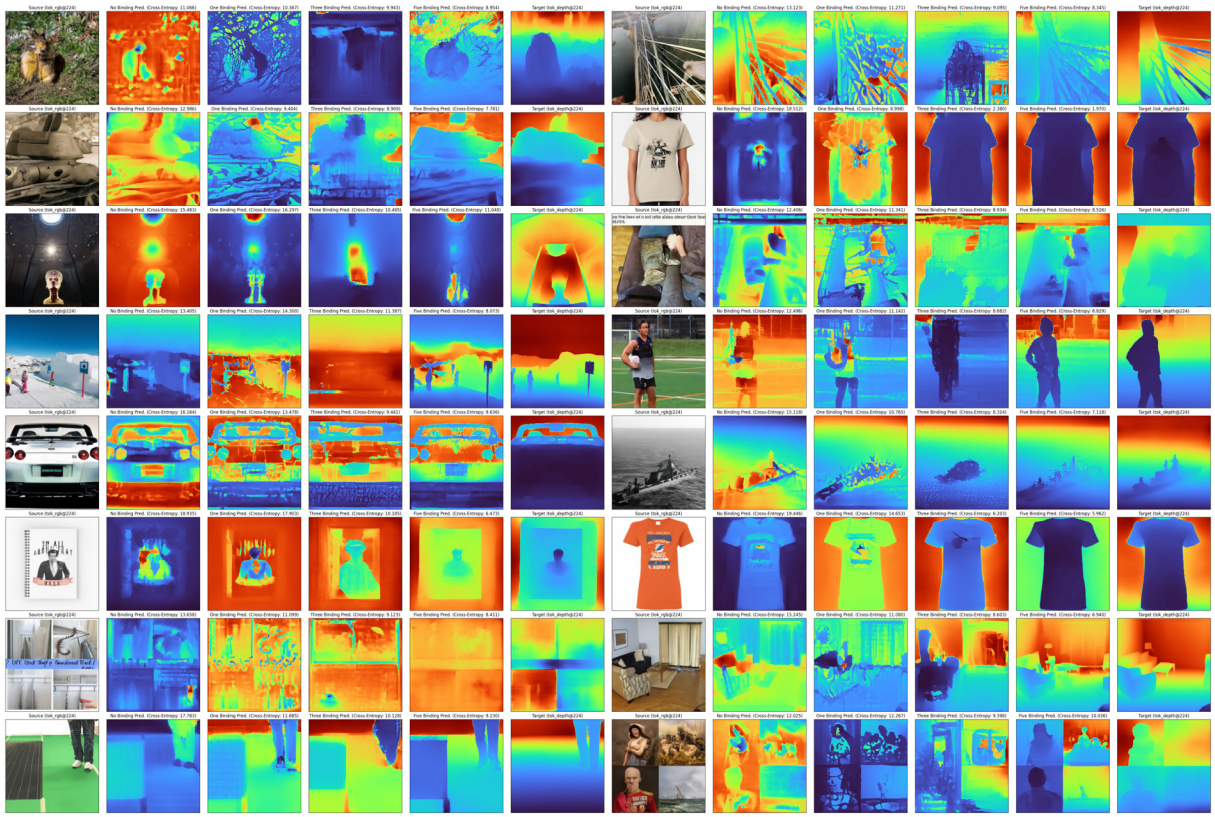

Figure 10. **RGB → Depth**. For each sample, we show from left to right the zero-shot prediction for zero, one, three and five binding modalities seen during training.

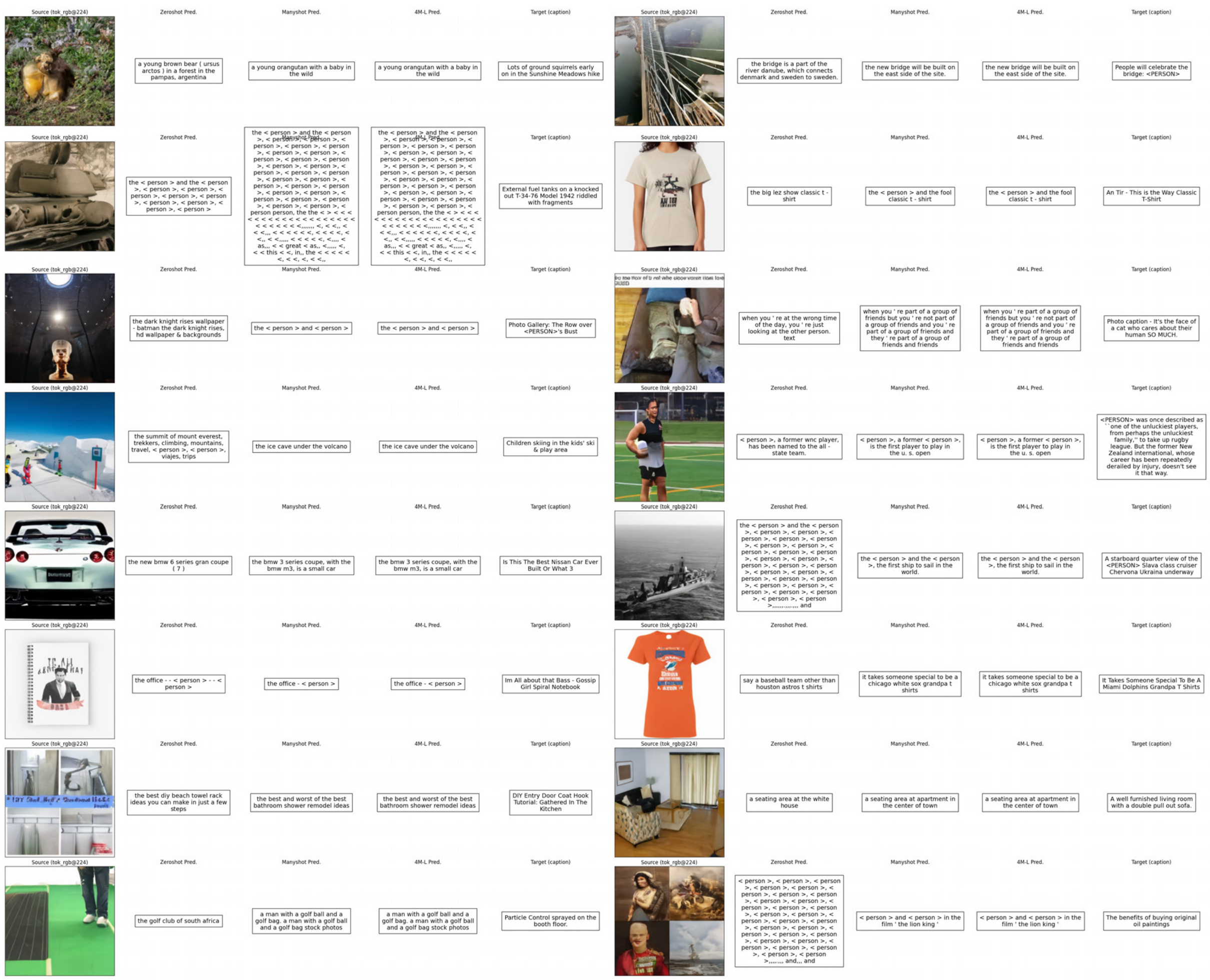

Figure 11. **RGB → Caption**. For each sample, we show from left to right: decoded tokenized RGB input, zero-shot Caption prediction, many-shot Caption prediction, and the decoded ground truth.

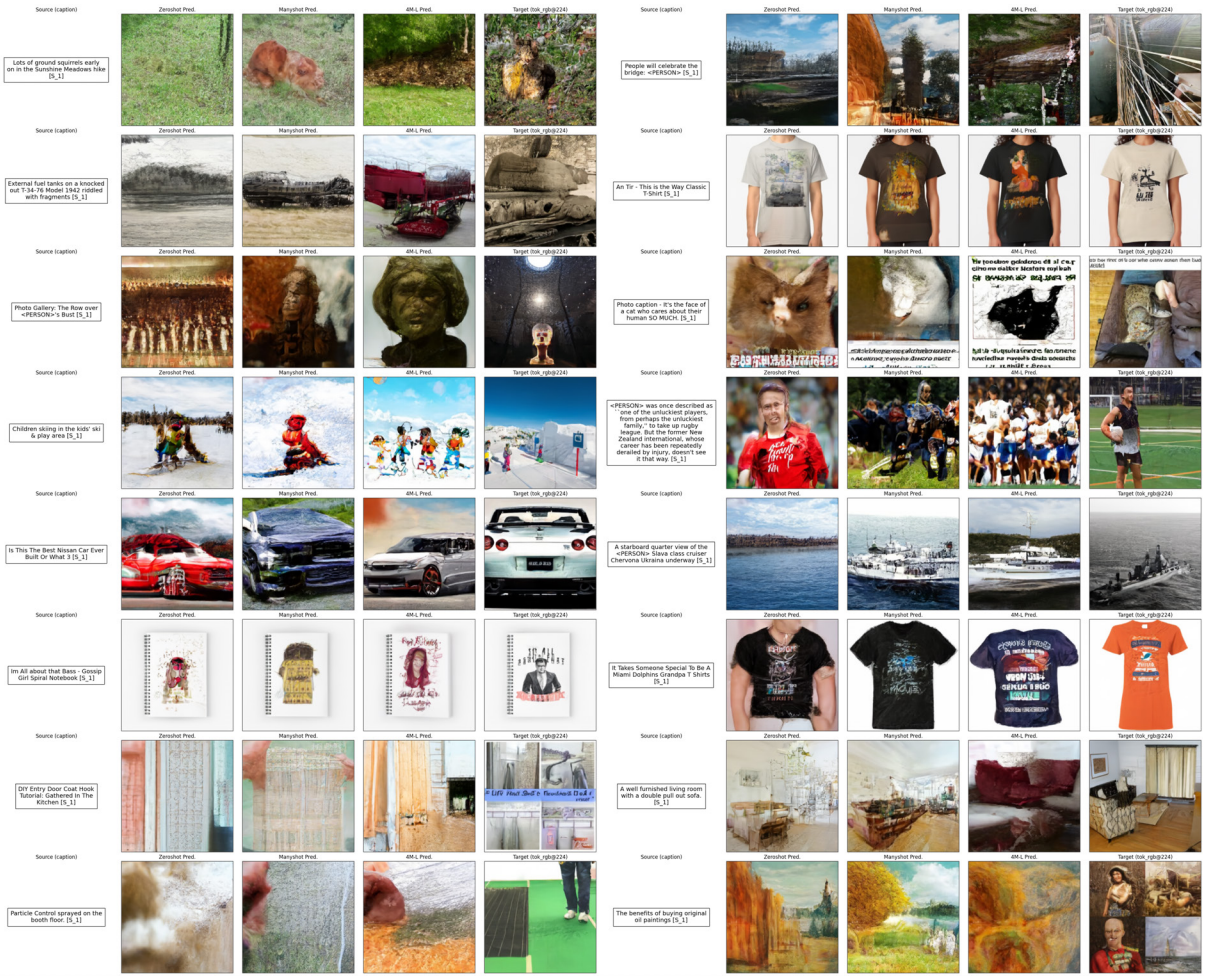

Figure 12. **Caption → RGB**. For each sample, we show from left to right: Caption input, zero-shot RGB prediction, many-shot RGB prediction, and the decoded ground truth.