

Advancing Semantic Future Prediction through Multimodal Visual Sequence Transformers

Supplementary Material

Efstathios Karypidis^{1,3} Ioannis Kakogeorgiou¹ Spyros Gidaris² Nikos Komodakis^{1,4,5}

¹Archimedes, Athena Research Center, Greece ²valeo.ai

³National Technical University of Athens ⁴University of Crete ⁵IACM-Forth

Contents

6. Extended Related Work	1
7. Additional Results	1
7.1. Comparison with VISTA	1
7.2. Tokenization: Comparing Our VAE-Free Approach to VQ-VAE	1
7.3. Impact of Our Multi-Modal Fusion Strategy .	2
7.4. Additional Results on SYNTHIA-Seq	2
8. Additional Qualitative Results	3
9. Limitations and Future Work	4

6. Extended Related Work

We provide an extended related work section on studies employing multimodal transformers.

Multimodal Learning with Transformers To effectively process multiple dense visual modalities and enable robust cross-modal interactions, several multimodal transformer approaches have been proposed. Some approaches employ flexible architectures that handle arbitrary input and output modalities. Notably, Perceiver IO [7] and ImageBind [5] integrate embeddings from multiple modalities into a shared representation space, enabling unified processing of diverse data types within a single framework.

Other works extend masked autoencoders to multimodal and multitask settings, such as MultiMAE [1], which introduces a masked transformer for joint pretraining across modalities. Building upon this, models for massively multimodal masked modeling and any-to-any vision tasks have been introduced [2, 9].

To further unify representations across various tasks, transformer architectures have been leveraged to integrate multiple modalities within a single framework. For instance, Xiao *et al.* [20] focused on combining vision and language data, while Lu *et al.* [8] have integrated vision, language, audio, and action data. Additionally, generative transformer-based multimodal models have demonstrated

the ability to perform in-context learning across modalities, improving adaptability and performance in diverse tasks [19]. Moreover, sequential modeling techniques have been applied to large multimodal transformers, enabling scalable training and enhancing performance in multimodal tasks [3]. Furthermore, UniT [6] introduces a unified transformer framework for multimodal multitask learning, effectively processing images, text, and videos across various tasks within a single model.

Our work proposes FUTURIST, a multimodal visual sequence transformer for semantic future prediction. Specifically, we introduce a transformer-based architecture tailored for multimodal future prediction, leveraging early modality fusion through the concatenation of per-modality tokens to enhance efficiency and cross-modal synergies. Unlike many existing approaches, our framework is VAE-free, which simplifies the training pipeline, reduces computational overhead and enables end-to-end optimization. To the best of our knowledge, our method is the first to employ a multimodal transformer for future prediction, effectively addressing the unique challenges of semantic future prediction with a unified architecture.

7. Additional Results

7.1. Comparison with VISTA

In Table 6, we compare our FUTURIST approach against the VISTA [4] baseline on both the segmentation and depth forecasting tasks. We conducted evaluations using a version fine-tuned on Cityscapes for 10 epochs to address potential performance issues arising from domain shift. LORA Fine-tuning required 8 GPUs, with a batch size of 1 per GPU, utilizing approximately $80 \times 8 = 640$ GB of VRAM. VISTA finetuned still fell far behind FUTURIST.

7.2. Tokenization: Comparing Our VAE-Free Approach to VQ-VAE

In Table 7, we compare our approach to variations that replace the proposed VAE-free tokenization process (described in Sec. 3.1) with opensource large-scale pretrained discrete tokenizers such as VQ-VAE model from DALL-

METHOD	SEGMENTATION				DEPTH			
	SHORT-TERM		MID-TERM		SHORT-TERM		MID-TERM	
	ALL \uparrow	MO \uparrow	ALL \uparrow	MO \uparrow	$\delta_1\uparrow$	ABSREL \downarrow	$\delta_1\uparrow$	ABSREL \downarrow
COPY LAST	55.5	52.7	40.5	32.2	90.5	10.780	82.2	18.345
VISTA FINE-TUNED	46.3	44.9	41.0	36.7	84.4	14.877	80.4	17.991
FUTURIST	73.9	74.9	62.7	61.2	96.0	5.384	91.9	9.111

Table 6. **Comparison with VISTA on semantic segmentation and depth forecasting.** Our FUTURIST model was trained for 3200 epochs. ABSREL is multiplied by 100 for readability.

TOKENIZER	SEGMENTATION				DEPTH			
	SHORT-TERM		MID-TERM		SHORT-TERM		MID-TERM	
	ALL \uparrow	MO \uparrow	ALL \uparrow	MO \uparrow	$\delta_1\uparrow$	ABSREL \downarrow	$\delta_1\uparrow$	ABSREL \downarrow
ORACLE [18]	78.6	80.8	78.6	80.6	-	-	-	-
ORACLE AFTER DALL-E’S VQ-VAE RECONSTRUCTION	71.5	72.2	71.5	72.2	98.6	2.263	98.6	2.263
ORACLE AFTER LDM’S VQ-VAE RECONSTRUCTION	72.1	71.9	72.1	71.9	97.3	4.665	97.3	4.665
ORACLE AFTER LDM’S VQ-VAE-FT RECONSTRUCTION	77.1	78.9	77.1	78.9	98.7	2.776	98.7	2.776
VQ-VAE FROM DALL-E [14]	65.0	62.8	54.4	48.4	94.5	6.643	88.3	10.855
VQ-VAE FROM LDM [16]	60.7	55.7	51.5	44.6	91.8	9.032	84.9	13.804
VQ-VAE-FT FROM LDM	69.2	68.9	57.2	53.1	90.7	8.912	82.2	14.068
Our VAE-free hierarchical tokenization	72.9	73.8	61.6	59.9	95.8	5.606	91.5	9.490

Table 7. **Tokenization: comparing our VAE-Free approach to VQ-VAE.** All models are trained for 800 epochs. ABSREL is multiplied by 100 for readability. We do not report results for the Oracle baseline in depth forecasting because, unlike segmentation where metrics are based on ground truth from the dataset, there is no ground truth for depth. Instead, we use the Oracle (DepthAnything [21]) to generate pseudo-ground truth for comparison. As a result, the Oracle is expected to achieve 100% accuracy and 0 error in depth prediction.

E [14] or LDM [16]. For this comparison, we rendered the segmentation or depth modalities as RGB images, by applying the cityscapes colormap for segmentation and replicating values across three channels for depth and fed them into the VQ-VAE encoder for tokenization. The results show that using VQ-VAE leads to significantly worse performance compared to our VAE-free approach.

The main reason for this gap is that the reconstruction process in VQ-VAE significantly degrades the oracle performance, particularly for the segmentation modality. In fact, segmentation results after VQ-VAE reconstruction are worse than our predicted segmentation results for short-term predictions. Moreover, we also fine-tuned LDM’s VQ-VAE [16] (since DALL-E does not offers training code and recipe) on segmentation and depth maps, improving both Oracle reconstruction (ORACLE AFTER LDM’S VQ-VAE-FT RECONSTRUCTION) and future segmentation prediction (VQ-VAE-FT FROM LDM), though depth prediction performance slightly declined. Still, our VAE-free method is more effective and efficient: fewer parameters (465M vs 558M total), faster training (8h vs 22h at 800 epochs), and lower inference time per sequence (52ms vs 274ms).

The key takeaway is that while VQ-VAE tokenizers are essential for generative models focused on image or video generation, they are likely unnecessary for semantic modal-

ities like those considered here. Our VAE-free approach not only simplifies the training pipeline but also achieves superior performance

7.3. Impact of Our Multi-Modal Fusion Strategy

In Table 8, we compare our approach to a variation that keeps tokens from the two modalities separate, instead of using our multi-modality early fusion strategy, which concatenates them along the embedding dimension. When trained for the same number of epochs, the separable tokens approach shows a slight improvement in segmentation performance, while our method performs better on most depth metrics. However, our approach is significantly more efficient, requiring half the training time and GPU memory. Furthermore, when trained for twice as many epochs—matching the total compute budget of the separable tokens approach—our method outperforms it in six out of eight metrics. This demonstrates that under the same compute budget, our approach delivers superior results.

7.4. Additional Results on SYNTHIA-Seq

In our work we focus on Cityscape which is the standard benchmark for future semantic prediction, used by most prior approaches. We also conducted experiments on the SYNTHIA-Seq [17] dataset to further validate our model’s

APPROACH	SEGMENTATION				DEPTH				TRAINING
	SHORT-TERM		MID-TERM		SHORT-TERM		MID-TERM		TIME
	ALL \uparrow	MO \uparrow	ALL \uparrow	MO \uparrow	$\delta_1\uparrow$	ABSREL \downarrow	$\delta_1\uparrow$	ABSREL \downarrow	HOURS \downarrow
SEPARATE TOKENS – 800 EPOCHS	73.3	74.3	62.2	60.6	95.8	5.622	91.5	9.442	18
OUR MULTI-MODAL FUSION – 800 EPOCHS	72.9	73.8	61.6	59.9	95.8	5.606	91.5	9.490	9
OUR MULTI-MODAL FUSION – 1600 EPOCHS	73.4	74.4	62.1	60.4	95.9	5.444	91.7	9.092	18

Table 8. **Impact of our multi-modal token fusion strategy.** ABSREL is multiplied by 100 for readability. Training time is computed in hours.

METHOD	SEGMENTATION		DEPTH			
	SHORT-TERM	MID-TERM	SHORT-TERM		MID-TERM	
	MIOU \uparrow	MIOU \uparrow	$\delta_1\uparrow$	ABSREL \downarrow	$\delta_1\uparrow$	ABSREL \downarrow
Oracle	87.9	87.9	76.8	20.195	76.8	20.195
Copy Last	54.3	44.4	65.7	27.120	61.8	30.499
Vista Fine-tuned	49.5	40.7	69.5	19.949	65.2	24.670
FUTURIST	63.8	53.2	73.7	23.720	69.4	25.355

Table 9. **Comparison on semantic segmentation and depth forecasting at SYNTHIA-Seq Dataset.** ABSREL is multiplied by 100 for readability.

effectiveness. SYNTHIA-Seq provides 5 synthetic urban driving sequences across diverse scenarios with multiple environmental variations (Spring, Summer, Fall, Winter, Rain, Soft-rain, Sunset, Fog, Night, and Dawn). Each sub-sequence contains approximately 8,000 frames at 5 fps with resolution of 1280×760 pixels. The dataset offers rich annotations including 8 camera views, semantic segmentation for 14 classes (misc, sky, building, road, sidewalk, fence, vegetation, pole, car, sign, pedestrian, cyclist, lane-marking, traffic-light), instance segmentation, global camera poses, depth maps, and calibration parameters. For our experimental setup, we utilized sequences 1, 2, 4, and 5 from the SYNTHIA-Seq dataset for training, while sequence 6 was reserved for validation. Given the diverse environmental variations per sequence, this configuration yielded 42 training sequences and 8 validation sequences. For evaluation, we selected 20 keyframes from each sequence in the validation set, resulting in 160 keyframes total for assessing model performance.

Our evaluation protocol followed both short-term and mid-term prediction paradigms. For semantic segmentation assessment, we computed mean Intersection over Union (mIoU) on a subset of classes present in the keyframes, specifically: misc, sky, building, road, sidewalk, fence, vegetation, pole, car, and lane-marking. While comparing with many prior approaches as in Table 1 is challenging due to lack of publicly available training code for the future segmentation task, we compared FUTURIST (fine-tuned on SYNTHIA) with VISTA (also fine-tuned) and the Copy-Last baseline. FUTURIST operates on predictions from

segmentation and depth models (Oracle), both comprising DINOv2 [10] as feature extractor and DPT [15] head, trained on SYNTHIA, while VISTA applies these same models to its predicted RGB outputs. Results in Tab. 9 show FUTURIST significantly outperforms both baselines.

8. Additional Qualitative Results

In Figures 8, 9, 10, 11 and 12, we present additional qualitative results using FUTURIST for forecasting semantic segmentation and depth maps over extended time horizons. These results were generated through autoregressive roll-outs. Starting with a sequence of four context frames (X_{t-9} to X_t), the model predicts up to 48 future frames, corresponding to 2.88 seconds, with a frame interval of 3.

The examples in Figs. 8 to 11 demonstrate that the model effectively preserves temporal coherence, maintaining consistent relationships between static and dynamic elements over time. The depth maps transition smoothly between frames, aligning well with changes in the scene. For instance, Figure 8 and Figure 10 show complex scenes with numerous static and moving objects. Here, the model captures the ego vehicle’s motion accurately, enabling precise predictions. In Figure 9, the model successfully predicts the motion of a car crossing perpendicularly to the ego vehicle, while Figure 11 highlights the model’s ability to anticipate the completion of a right turn.

However, as shown in Figure 12, the quality of predictions degrades toward the end of the rollout, particularly in the final four frames. The car masks become elongated instead of approaching the ego vehicle as expected. This

degradation likely stems from a mismatch between training and inference: during training, the model uses teacher forcing with oracle-provided context frames, whereas at inference, it relies on its own noisier predictions from previous steps. As discussed in [section 9](#), future work should aim to address this issue.

Videos of the predicted visual sequences shown in [Figs. 8 to 12](#) are included as .gif files in the supplementary material zip file.

9. Limitations and Future Work

While FUTURIST offers a simple and scalable approach to multi-modal semantic future prediction and demonstrates clear improvements over prior methods, several areas warrant further exploration to unlock its full potential.

Our work primarily focuses on short- and mid-term future predictions (0.18 seconds and 0.54 seconds, respectively), where our method excels. Although our model can be applied autoregressively for longer time horizons—generating mostly coherent predictions that capture both static and dynamic elements (as shown in [section 8](#))—further work is needed to improve robustness in these scenarios. This could involve addressing the challenges posed by noisy previous-step predictions (not currently considered due to teacher-forcing during training) or introducing stochasticity to better handle the inherent uncertainty of long-term predictions.

Another promising direction is extending FUTURIST to instance and panoptic segmentation tasks. Our approach could be adapted by encoding instance information through pixel-wise offsets to instance centers [\[11–13\]](#) rather than using instance IDs directly. This would allow each pixel within an instance mask to be represented by its x and y offsets to the corresponding instance center. These offsets could be processed as an additional modality with a dedicated embedding layer. During inference, instance masks could be recovered from these predicted offsets using a Hough transform-like approach and combined with semantic predictions to generate panoptic segmentation. This extension would enhance the applicability of our method to scenarios requiring instance-level understanding.

Currently, our method does not include action conditioning, which limits controllability. Incorporating sequences of control actions as an additional data modality would transform FUTURIST into a world model capable of predicting outcomes based on specific actions. This could increase its utility in autonomous driving and open up opportunities for applications in robotics and embodied AI.

Finally, we have not fully explored the scaling behavior of our approach, including model size, training data, and the number of modalities. However, even with training limited to Cityscapes, our approach delivers strong results, highlighting its significant potential for scalability and

broader generalization in future work. Additionally, our approach relies on pre-trained perception models to produce semantic modalities (e.g., segmentation and depth) from RGB images. While this dependency requires robust models for diverse scenes, the rapid advancements in foundational perception models and their increasing availability through open releases ensure that our method remains adaptable and continues to benefit from progress in the broader computer vision community. Moreover, operating on semantic modalities rather than RGB images makes it easier to exploit synthetic data generated by simulators, as these data do not need to be photorealistic.

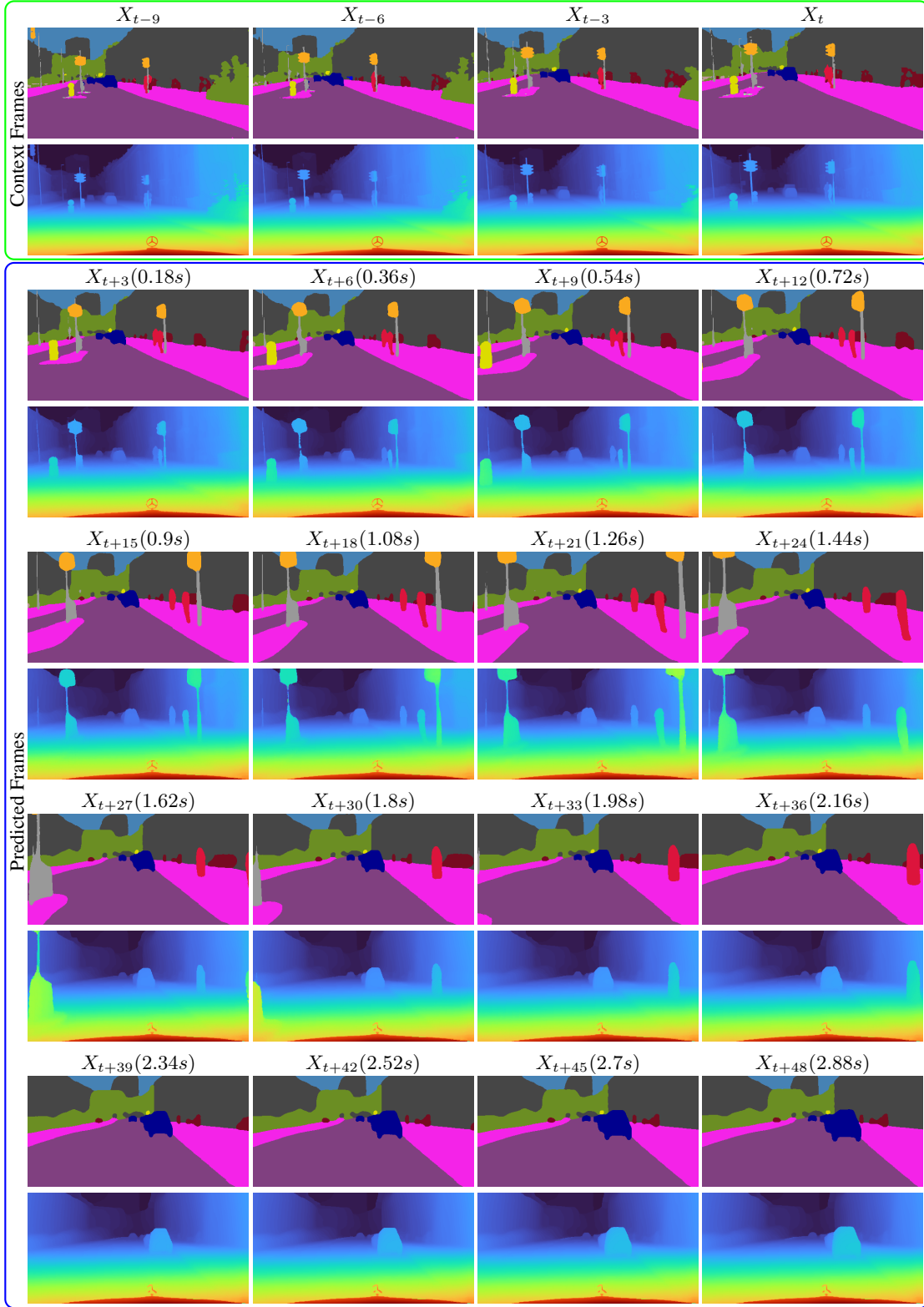


Figure 8. **Long-term semantic segmentation and depth predictions for Scene: Munster (23).** The model effectively captures temporal coherence in this complex scene with numerous static and moving objects.

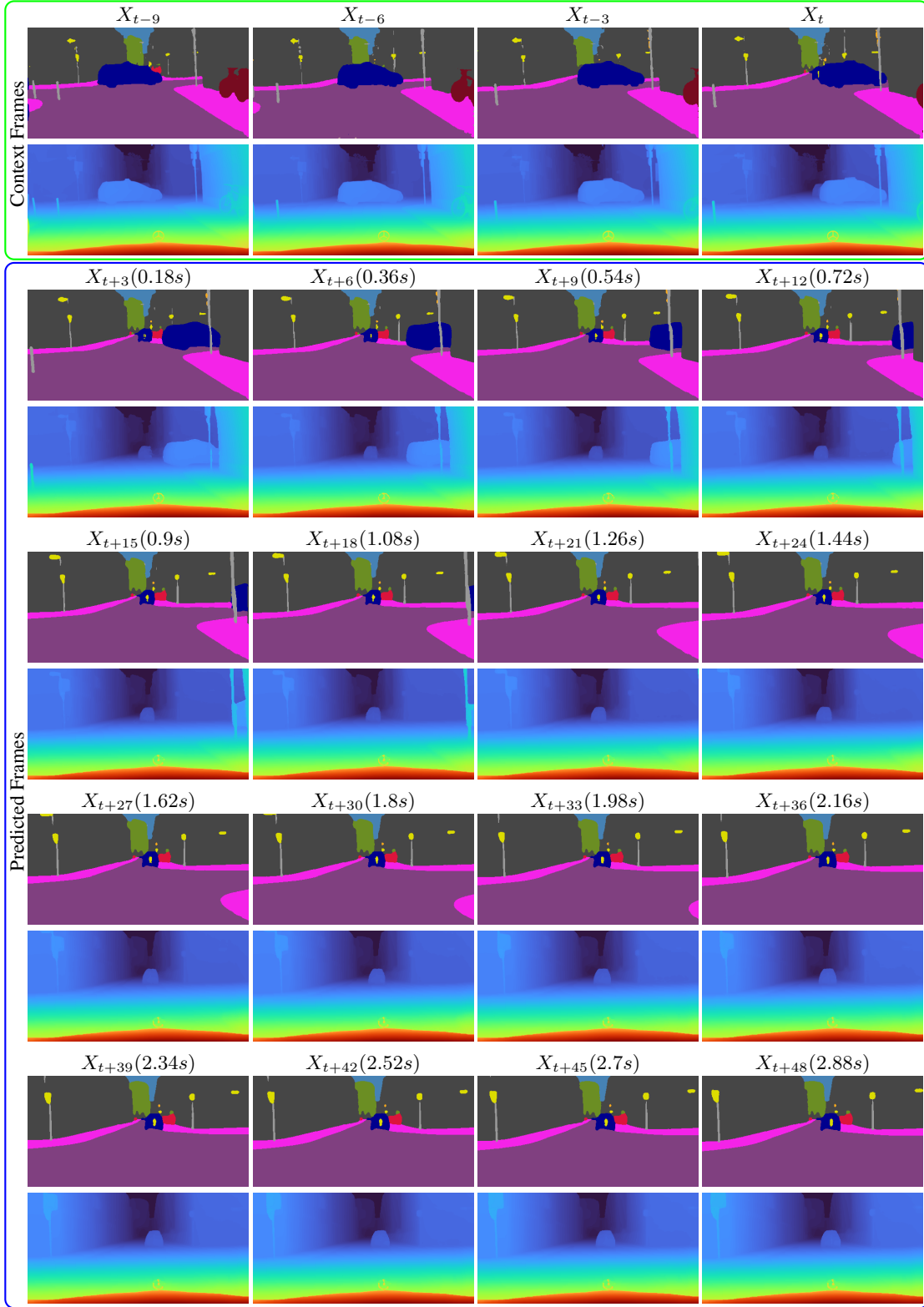


Figure 9. **Long-term semantic segmentation and depth predictions for Scene: Frankfurt (0_[275-304]).** Our approach accurately reflects the motion of static objects due to ego vehicle movement and predicts the motion of a car moving perpendicular to it.

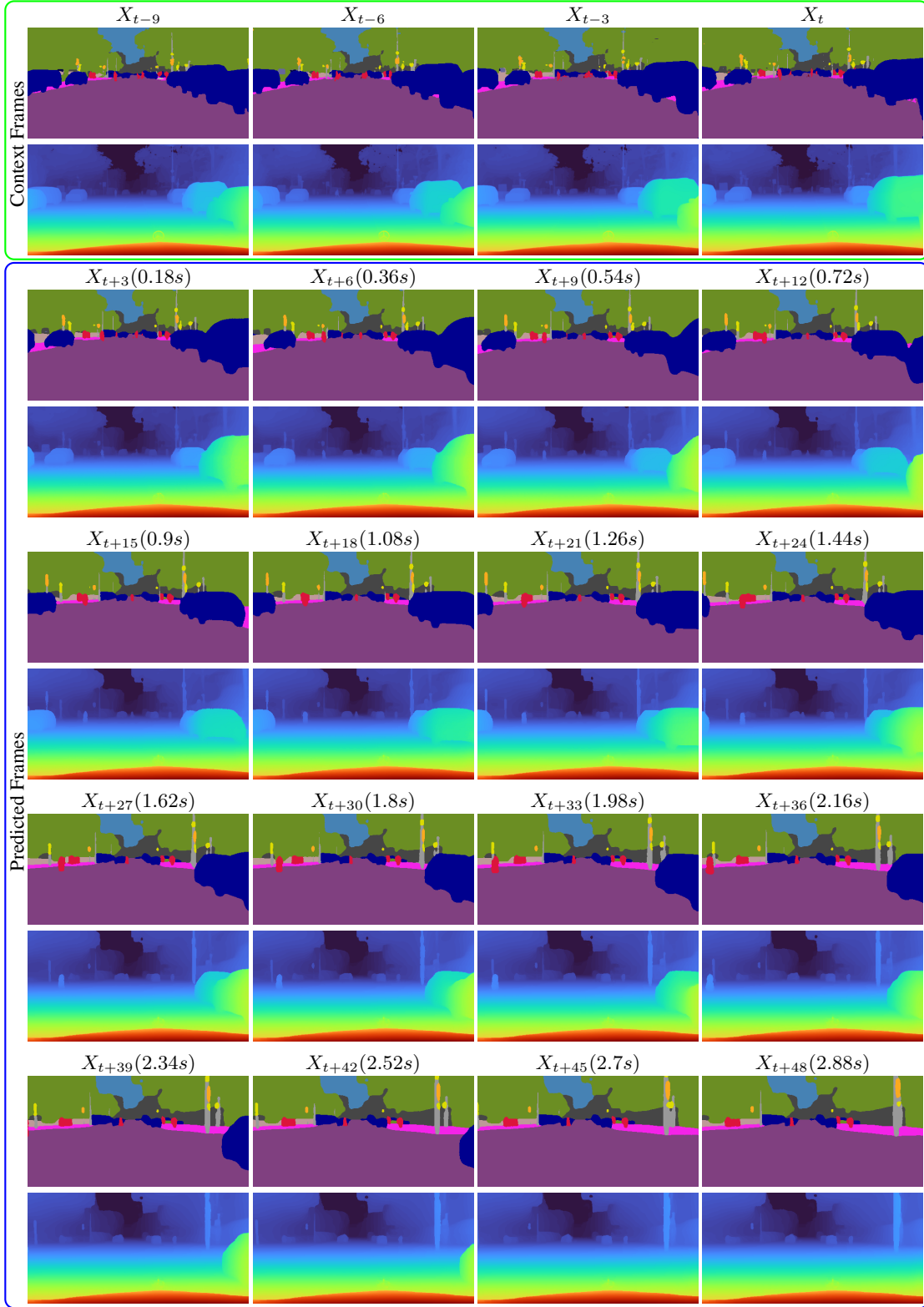


Figure 10. **Long-term semantic segmentation and depth predictions for Scene: Frankfurt (0_1217-1246).** Our model accurately preserves the relationships between static and dynamic elements and the motion of the ego vehicle.

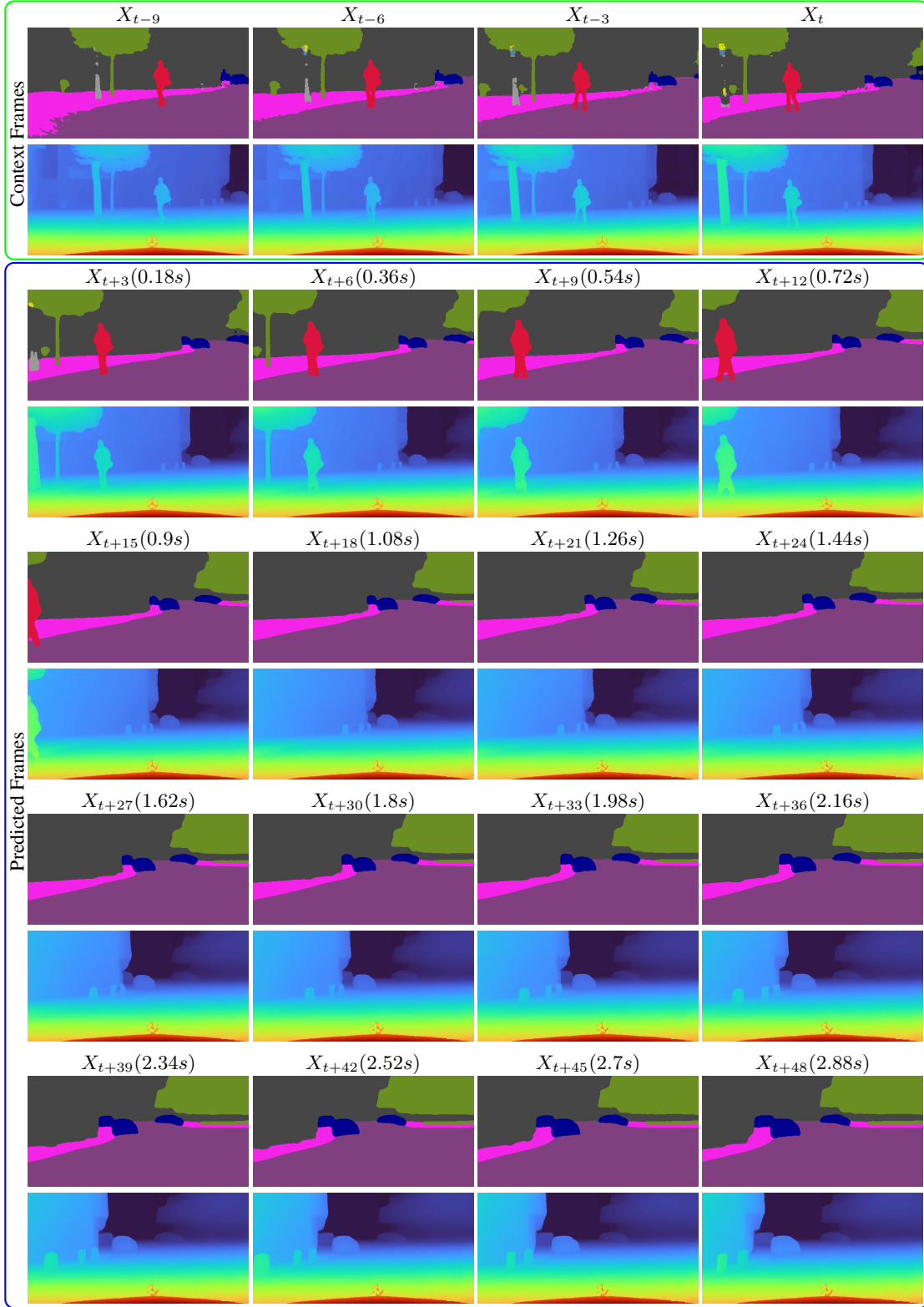


Figure 11. **Long-term semantic segmentation and depth predictions for Scene: Lindau (37).** This figure demonstrates our model's ability to anticipate the completion of a right turn.

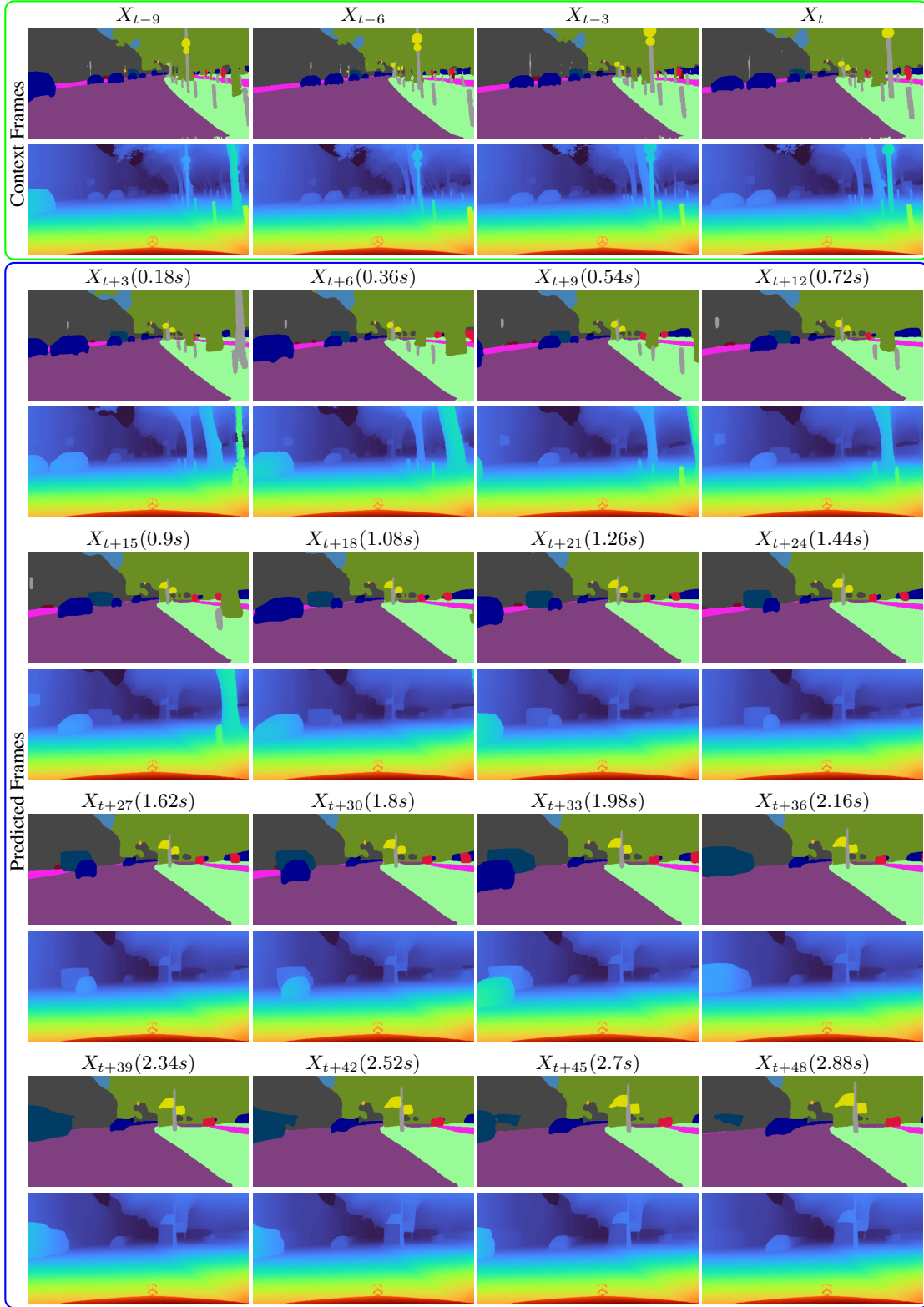


Figure 12. **Long-term semantic segmentation and depth predictions for Scene: Munster (160).** Despite precise short-term predictions, car masks become elongated towards the sequence’s end, indicating a need for future adjustments.

References

- [1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimaes: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 1
- [2] Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. In *NeurIPS*, 2024. 1
- [3] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, 2024. 1
- [4] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *NeurIPS*, 2024. 1
- [5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 1
- [6] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *CVPR*, 2021. 1
- [7] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*, 2022. 1
- [8] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, 2024. 1
- [9] David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. In *NeurIPS*, 2023. 1
- [10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3
- [11] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018. 4
- [12] Nedyalko Prisadnikov, Wouter Van Gansbeke, Danda Pani Paudel, and Luc Van Gool. A simple and generalist approach for panoptic segmentation. *arXiv preprint arXiv:2408.16504*, 2024.
- [13] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *CVPR*, 2021. 4
- [14] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2
- [15] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 3
- [16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2
- [17] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [18] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *CVPR*, 2021. 2
- [19] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiyang Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 1
- [20] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024. 1
- [21] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 2