ANY-TO-ANY VIDEO MODELING: MODELING THE WORLD BY TRAVERSING TIME AND MODALITIES

Anonymous authors
Paper under double-blind review

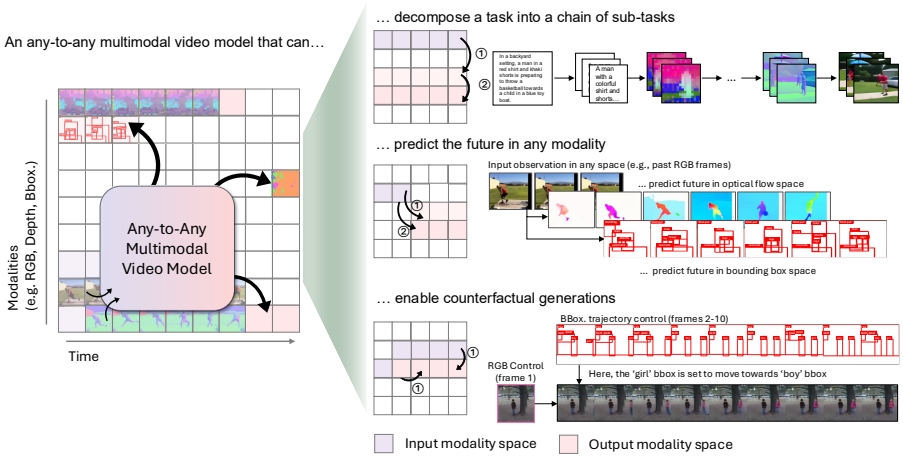

Figure 1: We develop an any-to-any multimodal model that can predict any modality, at any point in time, given an arbitrary combination of observations at arbitrary time steps. It is thus able to *traverse along both the modality and temporal axes*, enabling flexible inference capabilities: **chained generations**, where we decompose a task, e.g., caption-to-video, into a chain of potentially simpler sub-tasks of predicting intermediate modalities; **forecasting in any modality**, where given a sequence of past observations, the future frames can be rolled out in any of the modality; and **counterfactual generations**, where input conditioning modalities can act as *actions* that enables a diverse set of controls for steering generation, ranging from coarse (e.g., caption) to fine-grained (e.g., depth or bounding boxes).

ABSTRACT

Video modeling captures how the world is structured in space and how it evolves over time. Most approaches model this structure in a single representation space: RGB or a semantic representation (e.g., DINOv2, V-JEPA). However, different spaces capture different aspects of the underlying world and serve complementary roles, e.g., DINOv2 and V-JEPA excel at high-level semantic understanding, while RGB preserves fine-grained detail. In this work, we move beyond modeling the world in a single representation space and study video modeling jointly across time and modalities, each serving as a representation space (e.g., RGB, depth, optical flow, feature maps, and bounding boxes etc.). We develop an any-to-any multimodal model that can predict any modality, at any point in time, given an arbitrary combination of observations at arbitrary time steps. Flexibly traversing this time-modality grid gives rise to higher-level inference capabilities that are not supported by specialist video models. First, in chained generation, we decompose a task (e.g., caption-to-video) into a chain of potentially simpler sub-tasks of predicting intermediate modalities, and find that this yields higher-quality generations than direct generation. Second, we study forecasting in different modality spaces and find that no single space is optimal: the best prediction space varies across samples, and selecting it per sample yields substantially more accurate forecasts than committing to any single space. Finally, we show that the flexible modality-time conditioning enables counterfactual generation, where conditioning modalities act as actions that steer generation, from coarse (e.g., captions) to fine-grained (e.g., depth or bounding boxes) control.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Video models aim to model the spatiotemporal relationships of the underlying world. The majority of video models perform this modeling in RGB space, with some variance in input conditioning formats such as captions, class labels, depth, and Canny edges (Brooks et al., 2024a; Wang et al., 2024a; Yu et al., 2023; Wang et al., 2024d; Ho et al., 2022b; Chen et al., 2023). However, RGB is not always the best representation for modeling the world. It entangles both low-level appearance (e.g., pixel-level detail) with high-level semantics (e.g., object positions and scene composition). This is a difficult task for a video model, as it needs to learn these two aspects simultaneously. In contrast, several methods (Zhou et al., 2024; Luc et al., 2017; Karypidis et al., 2025b; Baldassarre et al., 2025) have moved towards modeling in feature spaces of pretrained vision encoders e.g., DINOv2 (Oquab et al., 2023), V-JEPA (Assran et al., 2025), etc. These approaches too, however, remain confined to a single representation space. We posit that different representations capture different aspects of the underlying scene and can serve complementary roles, e.g., DINOv2/V-JEPA for high-level semantic understanding, RGB for fine-grained detail, depth for geometry, etc. We therefore explore modeling the world jointly across multiple representations.

In this work, we move beyond modeling video in a single representation space, and develop Video-4M, an any-to-any multimodal model that models the world across different modalities (Fig. 1: left). Video-4M can predict any modality, at any point in time, given any combination of input modalities at any time-steps. It supports a diverse set of modalities: depth, normals, optical flow, bounding boxes, text, and feature maps, etc. Each modality captures a different aspect of an underlying scene often at different levels of abstraction: depth and surface normals encode more geometric detail than modalities such as captions or bounding boxes, which are more semantic. By traversing across time-modality axes, Video-4M does not fix a representation space and offers flexibility depending on the scene/task.

Video-4M also enables unique inference methods unavailable to specialist models (Fig. 1: right). In *chained generations* (§4.2), we show that decomposing a fixed task, e.g., text-to-RGB video generation, into a coarse-to-fine self-predicted chain of intermediate modalities (e.g., text to semantic to geometric to RGB) significantly improves results over both direct generation with the same model and a specialist text-to-video model trained in a controlled, matched-compute setup. Secondly, we study *forecasting in any modality* (§4.3) and find that the best representation (modality) for forecasting is highly scene-dependent, presumably since the required information (e.g. distance, object locations) differs by scene. Finally, we demonstrate that Video-4M enables counterfactual generations using modalities as actions (§4.4) at varying granularity, from coarse (e.g., a bounding box) to fine (e.g., a full depth map) conditioning, and across arbitrary points in time.

2 RELATED WORK

Multimodal video modeling. Many video models use different modalities as conditioning inputs for video generation: text in text-to-video approaches (Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a; Yu et al., 2023); depth, edges, and poses in ControlNet (Zhang et al., 2023) and ControlVideo (Chen et al., 2023); sketches, reference videos, and motion vectors in VideoComposer (Wang et al., 2023); and text, audio, optical flow, and depth in LLM-based VideoPoet (Kondratyuk et al., 2023). In most methods, other modalities mainly serve as inputs, while the world itself is modeled in RGB latent space (e.g., discrete tokens from quantized VAEs (Yu et al., 2023; Tang et al., 2024)). Other works, especially in robotics and driving, model the world in modalities beyond RGB, such as feature or segmentation maps (Yang et al., 2023; Zhou et al., 2024; Karypidis et al., 2025b; Pallotta et al., 2025; Karypidis et al., 2025c). However, the modality dimension remains largely underexplored, with most methods supporting only a limited set of modalities.

Any-to-any models. These models map any subset of supported modalities to any other. Notable examples include Unified-IO 2 (Lu et al., 2024) and 4M (Mizrahi et al., 2023; Bachmann et al., 2024), primarily developed for images. To our knowledge, any-to-any modeling has not been explored for video. Video-4M adopts the design choices of 4M (Mizrahi et al., 2023) and tailors them to video.

Hierarchical modeling. Another line of work decomposes generation into coarse-to-fine stages via intermediate representations. In the image domain, DALL-E 2 (Ramesh et al., 2022) first generates CLIP (Radford et al., 2021) image embeddings before decoding to pixels, and VAR (Tian et al., 2024)

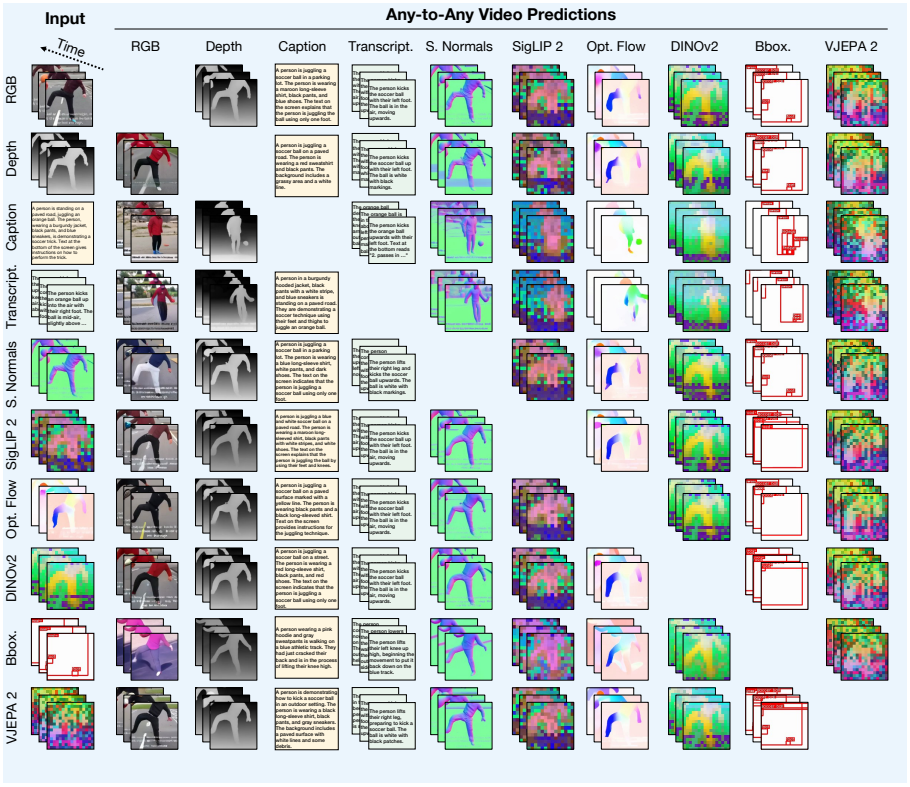

Figure 2: **Any-to-any capabilities.** Video-4M allows predictions from any input modality to any target modality using *a single model*. Each row shows predictions in a chained manner from different input modalities, which all correspond to the same video sample.

generates images progressively from coarse to fine token maps via next-scale prediction. For videos, MilliVid (Chandratreya et al., 2026) uses hierarchical autoencoder latents for long-range consistency, and Karypidis et al. (2025a) predict semantic latents before RGB for video prediction. Video-4M is related to these works in spirit, as it can also generate hierarchically through a chain of modalities. However, unlike these methods, where the hierarchy is fixed by design, both the chaining order and the output space in Video-4M remain flexible and can be chosen based on the task at hand.

forecasting in abstract spaces. DINO-WM (Zhou et al., 2024) and DINO-Foresight (Karypidis et al., 2025b) explore modeling the future in the feature spaces of vision models. Chiu et al. (2020) and Luc et al. (Luc et al., 2017) explore forecasting in segmentation space conditioned on RGB frames. However, these methods are limited to operating in a single or a few modalities. Video-4M supports multiple representations and enables joint traversal of modalities and the temporal axis, adjusting the choice of modalities based on the downstream settings or a particular sample.

3 METHOD

Video-4M extends the any-to-any multimodal framework of 4M (Mizrahi et al., 2023) to the video domain. It is built on three components: a multimodal video dataset constructed via pseudolabeling, a unified tokenization scheme, and a multimodal masked modeling objective.

3.1 MULTIMODAL PRETRAINING VIDEO DATASET

Training Video-4M requires video data annotated across a diverse set of modalities, where each modality represents a structured representation of the world, from geometric properties such as depth and surface normals to motion cues such as optical flow, to semantics such as bounding boxes and feature maps. Since no existing video dataset provides aligned annotations across many modalities at the scale we need, we generate them via pseudolabeling a base video dataset using specialist models. This is the established route for building large-scale multimodal datasets, as shown by prior work

such as 4M (Mizrahi et al., 2023), 4M-21 (Bachmann et al., 2024), and MODUS (Ye et al., 2026). To ensure a high quality bar for pretraining data, we use SOTA or near-SOTA models for each modality.

Starting from Kinetics-600 (Carreira et al., 2018) ($\sim$400K videos across 600 action categories), we generate 10 modalities across 4 types: **(i)** *geometric* including depth, normals, and 3D human poses, **(ii)** *motion* including optical flow, **(iii)** *semantic* including feature maps and bounding boxes, and **(iv)** *high-level* including captions and transcriptions. More details are provided in Appendix E.

3.2 TOKENIZATION

Different modalities have different structures (e.g., feature maps vs. RGB vs. bboxes) and are not directly compatible with a shared modeling architecture. We eliminate modality-specific design choices by mapping modalities into discrete tokens, which can be jointly modeled with a single pretraining objective.

Modalities with a dense 3D structure. For modalities with a 3D spatiotemporal structure such as RGB, depth, normals, and optical flow, we train FSQ (Mentzer et al., 2024) based VAE tokenizers (Tang et al., 2024) that compress inputs both spatially and temporally, resulting in 1280 discrete tokens per 4-second clip (from a $5 \times 16 \times 16$ token tensor) at 128 resolution and 4 FPS. For 3D modalities, we fine-tune on top of a pretrained RGB tokenizer. For feature maps, we replace the L2 reconstruction loss with a cosine similarity objective, and train dedicated tokenizers from scratch.

Modalities that are represented sequentially. For caption, transcriptions, and object names in bounding boxes are treated as text tokens and encoded using WordPiece (Devlin et al., 2019). Bounding box coordinates and human pose parameters are encoded using dedicated coordinate-bin tokens, with human poses further tokenized using VQ-VAE. Special frame markers [FRAME_1], [FRAME_2], etc., are included to delineate frame boundaries, which allows the model to be aware of the temporal structure. For transcriptions, we use four special temporal tokens [SEC_1], [SEC_2], [SEC_3], and [SEC_4] to mark the start of each one-second transcription segment aligned with the 4-second video clips. We show the tokenizer reconstruction results in the Appendix F. In total, the tokenized pretraining data amounts to 34B tokens. We denote this dataset as K600-MM.

3.3 MODELING

As shown in Fig. 3, Video-4M is an encoder-decoder transformer trained with a multimodal masked modeling objective (Mizrahi et al., 2023). During pretraining, we randomly select one subset of modality tokens (across both space and time) as encoder input and another as masked target tokens for the decoder to predict. For dense modalities, tokens are sampled randomly for input and target, while for sequential modalities, we perform span masking (Raffel et al., 2020). We train the model with cross-entropy loss to predict the target tokens. This pretraining objective enables cross-modal prediction across both space and time and scales to any number of modalities.

Adaptations for temporal modalities. We apply the following adaptations to the image-based 4M's architecture (Mizrahi et al., 2023; Bachmann et al., 2024) for video modeling.

(i) Spatio-temporal positional embeddings. For 3D modalities (RGB, depth, optical flow, etc.), we use factorized additive learnable spatio-temporal positional embeddings. For sequential modalities such as bounding boxes and human poses, the number of tokens per frame varies (e.g., depending on the number of objects present). We use learnable 1D positional embeddings combined with frame marker tokens (e.g., [FRAME_1]), which serve both as temporal anchors and as frame boundary delimiters. This allows the model to infer when all tokens for a given frame have been generated.

(ii) Time-aware masking. When selecting input and output tokens, we ensure that frame marker tokens (e.g., [FRAME_1], [SEC_1]) are never masked for sequential modalities and always remain part of encoder input. This keeps encoder context temporally grounded and allows the decoder to infer each token's temporal location when predicting masked tokens.

(iii) Masking modes for forecasting. 4M employs random and span masking (Raffel et al., 2020) for non-sequential and sequential modalities respectively. Random masking selects a subset of tokens to mask or keep visible that simulates a diverse set of video tasks such as inpainting, outpainting, and past or forecasting etc. We introduce two additional masking modes for non-sequential and one for sequential modalities to reinforce Video-4M forecasting capabilities.

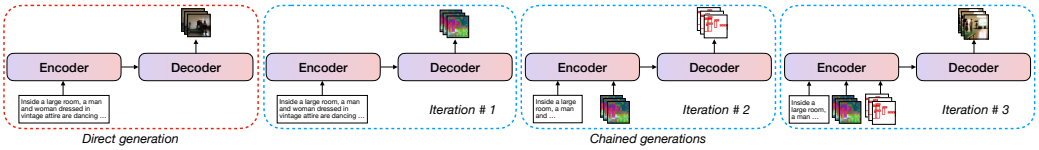

Figure 4: **Direct vs. Chained generation.** We show an example of decomposing a caption-to-RGB task into a *chain* of sub-tasks. We use Video-4M to sequentially predict intermediate modalities, e.g., SigLIP 2 and bounding boxes here, before generating the final RGB output.

While forecasting is implicitly present in random masking, it competes with other prediction modes due to the random token selection. These masking modes reinforce forecasting by ensuring the model always predicts from past context.

- **Frame-Conditioned Future Frame (FCFF):** Input tokens are sampled as a contiguous prefix from the initial frames, and target tokens are taken exclusively from subsequent future frames, also as a contiguous prefix. This enforces a temporal separation between input and output tokens.
- **Frame-Conditioned Any-Future (FCAF):** Input tokens are sampled as a contiguous prefix from initial frames similarly to FCFF, but the output tokens are sampled randomly in a non-causal manner from any of the subsequent future frames.
- **Causal Frame-level Prediction (CFP):** Designed for sequential modalities (e.g., bounding boxes, transcriptions), input tokens are selected as a contiguous prefix starting from the first token, and masked tokens begin immediately after the last input token. In our ablations, we found that incorporating CFP is essential for our inference strategy of feeding back predicted tokens to the encoder, as it simulates such cases during training. Without this masking mode, model generations significantly deteriorate when predicted tokens are fed to the encoder.

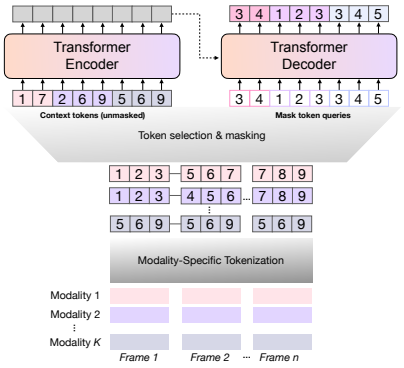

Figure 3: Video-4M **framework.** Context tokens and masked tokens are sampled from different modality tokens. Masking modes use a combination of random masking and forecasting masking modes tailored for both sequential and non-sequential modalities.

During training, non-sequential modalities use random, FCFF, or FCAF masking with equal probability, while sequential modalities use CFP with probability 35% alongside span masking. These masking modes improve the forecasting capabilities of Video-4M (Tab. 5 in Appendix C).

3.4 Chained generations

Unlike specialist video models that learn a fixed one-to-one mapping, Video-4M can traverse across modalities and generate any desired target modality from one or multiple input modalities. This higher-level flexibility gives rise to chained generations, where the model is iteratively tasked to generate one modality at a time, feeding each prediction back to the model as additional context before predicting the next target modality. We illustrate this in Fig. 4.

Chaining enables consistent predictions across modalities, e.g., depth that is in agreement with the predicted semantics. Crucially, chaining is not only useful for cross-modal consistency but also serves as a method to improve a desired task by decomposing a fixed one-to-one task into a self-predicted chain of potentially simpler sub-tasks. Below, we discuss one such instantiation for RGB generation.

Chaining for RGB video generation. Video generation can be viewed as two entangled sub-tasks: modeling high-level semantic changes across time (e.g., object positions, motion patterns, and scene composition), and modeling low-level pixel-level details. Jointly learning both aspects at once (e.g., directly generating in RGB), is challenging since the model must simultaneously reason about what is happening in a scene and how it should look. Video-4M instead generates a sequence of intermediate modalities: semantic ones (e.g., feature maps, bounding boxes etc.), followed by geometric ones (e.g., depth), before finally generating RGB.

Chain design. We use a coarse-to-fine strategy: the chain starts from high-level semantic abstractions towards modalities that represent low-level geometric structure. We empirically find that such chains result in improved RGB generation compared to randomly assigning the chain order (Tab. 3). To reduce the search space, we select a subset of Video-4M's modalities and categorize them into three groups: sparse (captions, transcriptions), semantic (SigLIP 2 (Tschannen et al., 2025), bounding boxes), and geometric (depth), and use Sparse → Semantic → Geometric chains across all tasks.

Optimal chain design. The ideal chain configuration is likely task-dependent. With n spaces, there exist $n!$ possible orderings, making an exhaustive search computationally prohibitive as n grows. In this work, we focus on RGB video generation and demonstrate that our coarse-to-fine chaining strategy consistently improves generation quality. We leave a systematic search for optimal chain configurations across all tasks to future work.

3.5 FORECASTING IN ANY MODALITY

Video models can generate rollouts into the future given past frames. This *forecasting* ability has sparked interest in their capacity as world models (Brooks et al., 2024b; Ali et al., 2025). While useful forecasting models with applications in robotics and driving are already emerging, existing models roll out the future in one or two spaces (Yang et al., 2023; Zhou et al., 2024; Karypidis et al., 2025b; Pallotta et al., 2025; Karypidis et al., 2025c). Video-4M adds an axis of modalities and allows the prediction space to be chosen based on the information of interest. For example, Video-4M can be used to forecast object-level motion directly in bounding-box space, rather than generating a future video from which motion must then be inferred.

Setup: we experiment forecasting through different modalities by giving as input some observed RGB frames and using Video-4M to generate the future frames of a chosen target modality. We test both direct prediction of the target modality (e.g. RGB → future Depth) or a single-hop chained prediction (e.g., RGB → future DINOv2 → future Depth). We evaluate the generated future frames.

Evaluation: There are many valid rollouts into the future given a single past sequence. This poses a problem for evaluating different generated rollouts because each video in the test set provides only one realized future. We utilize Gneiting & Raftery's energy score (Gneiting & Raftery, 2007), which provide a quantification of the quality of a predictive distribution against a single realized outcome. Given sampled rollouts $\hat{x}^{(1)}, \ldots, \hat{x}^{(K)}$ and the observed rollout x, the energy score is calculated as follows: $\text{ES} = \frac{1}{K}\sum_{i=1}^{K} d(\hat{x}^{(i)}, x) - \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} d(\hat{x}^{(i)}, \hat{x}^{(j)})$,where d is a task-appropriate distance metric. Explanation and intuition behind this score is explained in Appendix D.2. For the distance function, we use square root of the scale-invariant logarithmic loss (Sqrt-SiLog) (Eigen et al., 2014), L2 in occupancy maps, L2 in DINOv2 feature space, and L2 distance in Inception V3 feature space for depth, bounding boxes, DINOv2, and RGB respectively.

Table 1: **Comparison with specialist models on video tasks.** Video-4M is a single model evaluated across all 7 task groups, while baselines are specialists each covering a subset of tasks. Gray rows show their results at native resolution, without downsampling and tokenization offset.

		Video Generation				Future Prediction		Captioning		Depth				Normal			
		Cls→Vid		Cap→Vid		1 frame	5 frame	MSVD		DIODE		NYUv2		DIODE		NYUv2	
Method	**Tasks**	FVD ↓	Acc. ↑	FVD ↓	Text Align. ↑	FVD ↓	FVD ↓	MET. ↑	BERTSc. ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	Mean° ↓	<11.25° ↑	Mean° ↓	<11.25° ↑
Pseudolabelers																	
Video Depth Anything (Chen et al., 2025)	1/7	–	–	–	–	–	–	–	–	**0.571**	**0.673**	**0.194**	**0.855**	–	–	–	–
Video Depth Anything (Chen et al., 2025) (orig. res.)	1/7	–	–	–	–	–	–	–	–	0.564	0.681	0.192	0.859	–	–	–	–
NormalCrafter (Bin et al., 2025)	1/7	–	–	–	–	–	–	–	–	–	–	–	–	25.16	21.01	18.76	47.40
NormalCrafter (Bin et al., 2025) (orig. res.)	1/7	–	–	–	–	–	–	–	–	–	–	–	–	16.83	42.50	15.98	60.52
LARP (Wang et al., 2024a)	1/7	**166.8**	43.4	–	–	**78.8**	**72.0**	–	–	–	–	–	–	–	–	–	–
NBP (Ren et al., 2025)	2/7	897.1	13.7	–	–	255.9	141.6	–	–	–	–	–	–	–	–	–	–
OmniTokenizer (Wang et al., 2024b)	2/7	583.3	13.3	–	–	392.9	292.7	–	–	–	–	–	–	–	–	–	–
Video-LLaVA (Lin et al., 2024)	1/7	–	–	–	–	–	–	0.158	0.874	–	–	–	–	–	–	–	–
AuroraCap (Chai et al., 2025)	1/7	–	–	–	–	–	–	0.180	0.874	–	–	–	–	–	–	–	–
Video-4M (ours)	**7/7**	225.2	**70.1**	**180.2**	**20.8**	129.5	100.0	**0.221**	**0.895**	0.580	0.655	0.206	0.834	**20.61**	**38.73**	**18.70**	**50.86**

4 EXPERIMENTS

4.1 ANY-TO-ANY VIDEO PREDICTIONS

Video-4M supports diverse video tasks within a *single* neural network. As shown in Fig. 5, given RGB video as input, its predictions across modalities closely match those of the specialized pseu-

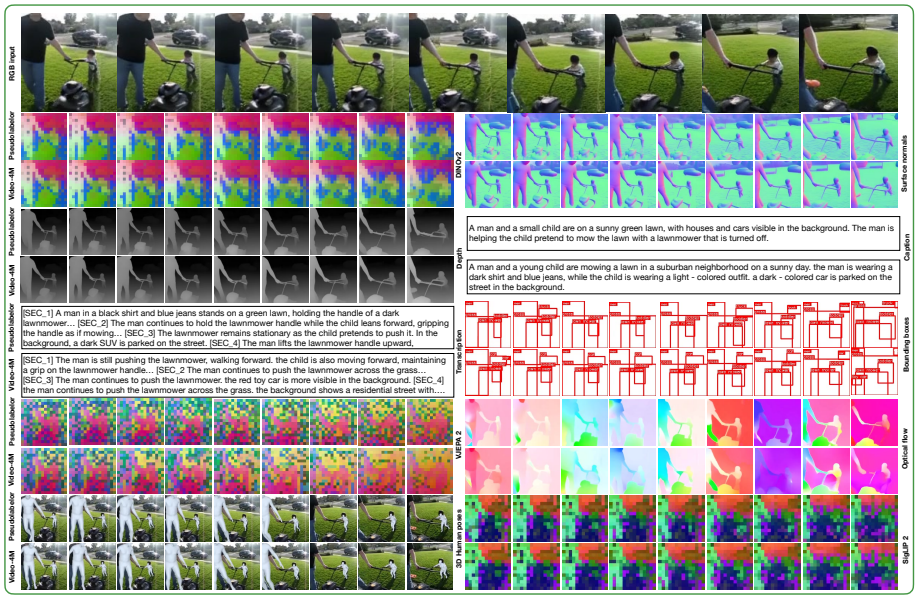

Figure 5: **Cross-modal predictions by Video-4M:** While being a *single* model, Video-4M generates diverse target modalities using RGB video as input. We compare its predictions with those of its pseudolabelers used to construct its pretraining data. Note that Video-4M uses tokenized 128-resolution inputs, while pseudolabelers use their native higher-resolution, pre-tokenized input videos.

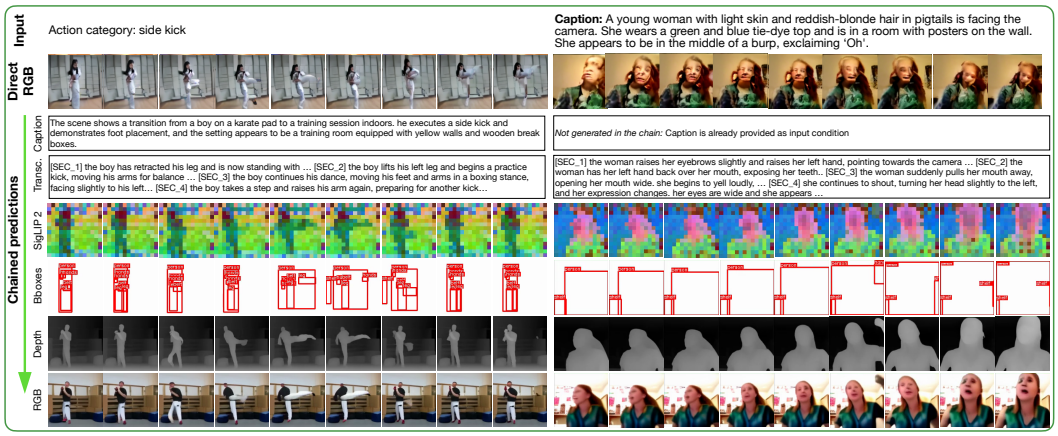

Figure 6: **Chained RGB video generations.** (*left:* class-to-video, and *right:* caption-to-video. Generating intermediate modalities such as transcriptions, SigLIP 2 feature maps, bounding boxes, and depth as middle modalities leads to improved RGB generations.

dolabelers used to create its pretraining data. In Tab. 1, we compare Video-4M against specialists on different video tasks. Video-4M shows overall competitive performance against specialist models. We provide additional discussion and evaluation details in Appendix C.

Beyond RGB-conditioned predictions, Video-4M maps between any pair of modalities representing the same spatiotemporal scene of the world. In Fig. 2, we show one-to-all generation, where Video-4M sequentially generates each modality from a single input. In the following sections, we demonstrate the higher-level inference capabilities of Video-4M.

4.2 Chained video generations

We now present experimental results for chained RGB generations. As our main baseline, we use Video-4M to generate RGB tokens conditioned only on the input, denoted Video-4M-direct. Both direct and chained generations use the same trained model. We also train a variant of Video-4M

Table 2: **Chained vs. direct modeling for RGB video generation.** Overall, chained predictions improve over directly modeling in RGB space.

Model	Video Generation				Video Prediction	
	Class→RGB		Caption→RGB		Frame conditional	
	FVD ↓	Top 1 Acc. ↑	FVD ↓	Text Align. ↑	1-frame ↓	5-frame ↓
RGB Baseline†	490.6	40.9	304.3	19.2	318.6	172.6
Video-4M-B Direct	325.0	**53.3**	418.6	17.9	230.0	138.0
Video-4M-B Chained	**263.7**	52.9	**250.0**	**19.2**	**155.0**	**119.0**
Video-4M-L Direct	330.6	68.8	220.8	**21.6**	188.1	117.1
Video-4M-L Chained	**225.2**	**70.1**	**180.2**	20.8	**129.5**	**100.0**

Table 3: **Which modality group matters most in chain?** Semantic ones contribute most among groups, while coarse-to-fine chain performs best overall.

Configuration	Video Generation		Video Prediction	
	Class→RGB	Caption→RGB	Frame conditional	
	FVD ↓	FVD ↓	1-frame ↓	5-frame ↓
Sparse (Caption → Transc. → RGB)	425.6	337.6	162.8	111.0
Semantic (SigLIP 2 → BBox → RGB)	<u>249.0</u>	**167.1**	<u>138.6</u>	<u>107.1</u>
Geometric (Depth → RGB)	446.3	282.3	173.3	109.3
Coarse-to-fine chain (default)	**225.2**	<u>180.2</u>	**129.5**	**100.0**

using RGB, caption, and class modalities, with the same pretraining schedule as Video-4M. This serves as our RGB-only baseline, and allows us to disentangle the effect of training exposure from the benefit of chaining. Results are shown in Tab. 2. Across model sizes, Video-4M with chaining outperforms the direct baseline. The trend is consistent across class-to-RGB, caption-to-RGB, and frame-conditional RGB generation tasks.

Qualitative results. We show generation results in Fig. 6. Videos generated using the direct generation baseline exhibit structural artifacts. While the overall semantics are largely preserved (e.g., a person attempting to perform a side kick when conditioned on this action), direct prediction struggles to model low-level details such as object boundaries, particularly in frames where sudden motion occurs. In chained generations, intermediate modalities such as semantic features and bounding boxes outline the semantics and motion trajectories, while geometric modalities such as depth help distinguish objects from each other. Additional qualitative results are shown in Appendix C.3.

Analysis on chain design. We study two questions: which modality group contributes most, and how the ordering of groups affects generation quality. In Tab. 3, we perform generation using individual modality groups and compare it with the full chain. Semantic modalities contribute most to generation quality, compared to the geometric and sparse groups. Conditioning on high-level semantics leads to better RGB generations than depth or other low-level modalities that mainly capture structural detail.

Tab. 4, shows the effect of chain order on generation quality. We reverse the chain order generating geometric modalities first, followed by sparse and semantic modalities. Reverse ordering performs worse than the semantic-first chain order. From qualitative comparisons, we note that when geometric modalities are generated first, subsequent modality predictions are constrained by the geometric modality. This limits semantic modalities to flexibly guide the video trajectory (except for details not captured by geometric modalities, such as color information), even when semantic modalities are part of the chain. Refer to Appendix C.3 for additional analysis.

Table 4: **Order of modality chain matters.**

Chain ordering	FVD ↓	Text Align. ↑
Caption → Geometric → Semantic → Sparse	236.2	20.6
Caption → Sparse → Semantic → Geometric (default)	**180.2**	**20.8**

4.3 FORECASTING IN ANY MODALITY

Fig. 7 shows the results of forecasting evaluated with energy scores (Gneiting & Raftery, 2007) – lower is better. We first note that, as expected, energy scores increase with prediction horizon, since farther future are more difficult to predict. The static baseline, which is simply a repetition of the last given frame used as the 'forecast', performs well in the initial few frames but rapidly degrades over time. Random and retrieval baselines, which are random and nearest-neighbor samples from the ground truth test set respectively, show static performance over the entire horizon. These baselines serve as reference sticks for interpreting the difference in energy scores.

On average, the performance gap between different modality chains is small. However, the best per-sample chain performs significantly better than any single fixed chain, showing that while different modality chains perform similarly on average the optimal chain is highly sample-dependent. Choosing the right representation space on a per-sample basis can substantially improve forecasting quality.

We note that the best-per-sample chain here is selected via an oracle, i.e., knowledge of optimal modality to forecast in, and is therefore not directly usable at inference time. As a step towards an oracle-free alternative, we show in Appendix D.1 that combining rollouts from multiple modalities into a single mixed-modality forecast can partially recover the benefits of per-sample modality selection, without requiring a best-chain oracle or predictor. Finding the best chain per sample itself,

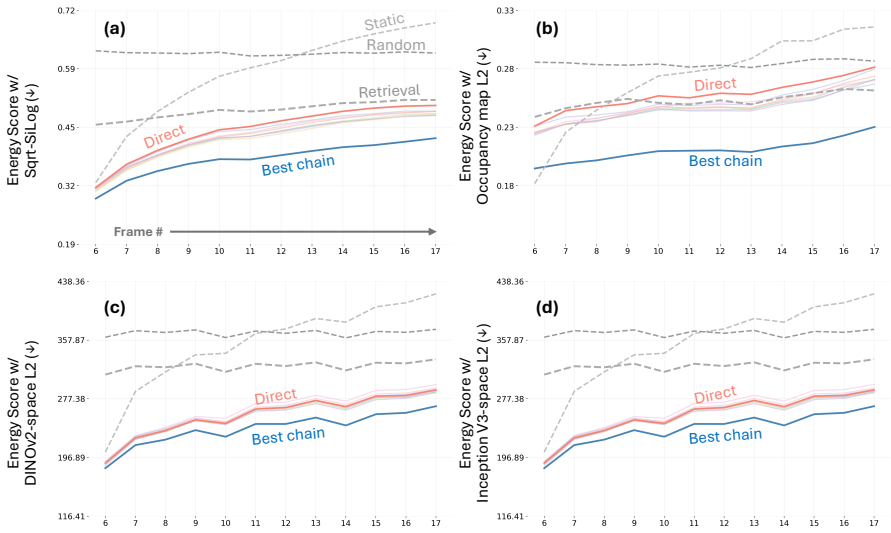

Figure 7: **Multimodal Forecasting** Future frames of each target modality (a) Depth, (b) Bounding Boxes, (c) DINOv2, and (d) RGB are predicted given 5 starting frames of RGB. Direct: target frames are predicted directly from the 5 RGB frames, e.g. `RGB[:5]` → `Depth[5:17]`. Static baseline: repetition of the last input frame as forecast. Random baseline: 16 random samples from the ground-truth test set as forecasts. Retrieval baseline: 16 closest neighbours from the ground-truth test set (closeness determined by modality-specific distance functions on the conditioning 5 frames) as forecasts. Best chain: result when the best modality chain for each sample is used – *cf.* unlabeled solid lines represent forecasts made through one particular modality for every sample.

however, is a non-trivial problem. In this work, we limit our contribution to showing the potential and leave the optimization problem itself for future work.

4.4 Counterfactual generations

Different modalities capture complementary spatio-temporal information of the world. Video-4M leverages this to enable a diverse set of modalities as *actions* for steering generation (in any output space), ranging from coarse (e.g., caption) to fine-grained (e.g., depth) control. We demonstrate a subset of these capabilities in Fig. 8 (Appendix B). Given the first RGB frame, we generate its future frames by conditioning using caption as a coarse steering modality. By varying the caption, we can generate plausible future that all start from the same image. For finer-grained control, we can condition on modalities such as depth or bounding boxes. Video-4M treats all of these tasks as instances of the same problem: completing the missing parts of a partially observed modality-time grid. Additionally, Video-4M can generate observations directly in modality spaces of interest instead of RGB. As shown in Fig. 9 (Appendix B),we use different input modalities to generate in other modalities such as optical flow and surface normals. This decoupling of input and output modalities highlights Video-4M's generality beyond RGB-centric generation.

5 Conclusion

We study video modeling across a diverse set of representation spaces (modalities), moving beyond the conventional RGB-centric video modeling paradigm, and develop an any-to-any multimodal video model. Our model traverses both the modality and time axes simultaneously, giving rise to higher-level inference capabilities that are not supported by specialist video models. First, we investigate that decomposing a downstream task (e.g., text-to-video) into a chain of sub-tasks improves the target task compared to direct generation. forecasting benefits similarly: different modality spaces allow for unique rollouts, such that the sample-dependent optimal chain enables significantly higher-quality predictions compared to predicting directly in a single modality space. Finally, we show that our model supports flexible input-output conditioning across modalities and time points, from coarse to fine granularity, which enables flexible counterfactual generation possibilities.

AI USE STATEMENT

In this work, we used generative AI tools primarily to polish the manuscript text, including grammar and coherence. To meet the submission page limit, we also used generative AI tools to shorten a few paragraphs. We did not use generative AI tools to formulate the research idea. We did, however, use them to develop scripts for running experiments after initial foundations were written by humans and for generating visualization examples. We also used generative AI tools to assist with formatting the manuscript in the conference template and to verify bibliographic entries. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work does not involve new data collection from human subjects or crowd workers. Video-4M is pretrained on K600-MM, which we build by pseudolabeling the Kinetics-600 dataset, a publicly available collection of videos sourced from YouTube. As a result, Video-4M inherits the limitations of this data. First, any biases present in Kinetics-600, such as the uneven representation of demographics, geographic regions, activities, or cultural contexts, may be absorbed into the model's representations and reflected in its predictions and generations. Second, because K600-MM extends Kinetics-600 with additional modalities, any privacy concerns in the source videos, such as the presence of identifiable individuals, carry over to our work. Third, our additional modalities are generated by off-the-shelf pseudolabelers (e.g., Gemini for captions and transcriptions, Grounding DINO for bounding boxes), so their biases can also potentially propagate into our training data and, in turn, into the model. Finally, as with other generative video models, Video-4M could be misused to produce misleading or deceptive visual content. Its fine-grained control, where a rollout can be steered through conditioning modalities such as bounding boxes or captions, may make such misuse easier. We plan to publicly release our code, model weights, and the K600-MM dataset to support reproducibility and future research. In doing so, we will follow the terms of use of Kinetics-600 and the pseudolabeling models, document the known limitations and biases described above, and release the model under a license that prohibits harmful uses such as creating deceptive content or surveillance. Our work is primarily methodological and is not intended for deployment in sensitive applications such as surveillance or behavioral inference. We discuss the broader impact of this work further in Appendix I.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce our results throughout the paper and the appendix. The construction of our pretraining dataset, K600-MM, including the pseudolabeling pipeline and the pseudolabeler used for each modality, is described in Sec. 3.1 and Appendix E. The tokenizers for each modality, both for 3D modalities and for sequentially represented ones (1D), are described in Appendix F. The model architecture, masking strategies, and chained generation procedure are described in Sec. 3.3. For training, we report the model scales, batch sizes, compute budget, and all pretraining hyperparameters in Appendix G and Tab. 10. For evaluation, we describe the metrics, the test sets, and the number of seeds used for each experiment in Appendix G, and explain the energy score used for forecasting in Appendix D.2. Finally, we plan to publicly release our code, pretrained model weights, and the K600-MM dataset to facilitate reproducibility and future research.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985, 2025.

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 Advances in Neural Information Processing Systems, 2024.

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. arXiv preprint arXiv:2507.19468, 2025.

Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. Normalcrafter: Learning temporally consistent normals from video diffusion priors. arXiv preprint arXiv:2504.11427, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024a. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. OpenAI Blog, 1(8):1, 2024b.

Neil Burgess, Jelena Milanovic, Nigel Stephens, Konstantinos Monachopoulos, and David Mansell. Bfloat16 processing for neural networks. In 2019 IEEE 26th Symposium on Computer Arithmetic (ARITH), pp. 88–91, 2019. doi: 10.1109/ARITH.2019.00022.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308, 2017.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. arXiv preprint arXiv:1808.01340, 2018.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In International Conference on Learning Representations, volume 2025, pp. 16845–16889, 2025.

Ishaan Preetam Chandratreya, David Charatan, Basile Van Hoorick, Sergey Zakharov, Vitor Guizilini, Phillip Isola, and Vincent Sitzmann. Millivid: Hierarchical latents for long-range consistency in video generation. arXiv preprint arXiv:2606.09056, 2026.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. arXiv preprint arXiv:2501.12375, 2025.

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852, 2021.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840, 2023.

Hsu-kuang Chiu, Ehsan Adeli, and Juan Carlos Niebles. Segmenting the future. IEEE Robotics and Automation Letters, 5(3):4202–4209, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics, 2019.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1323–1333, 2024.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Advances in Neural Information Processing Systems, 2014.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems, 35:35946–35958, 2022.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Google. Gemini, 2023. URL https://blog.google/technology/ai/google-gemini-ai/.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. ArXiv, abs/1706.02677, 2017.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. Advances in neural information processing systems, 35:8633–8646, 2022b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21807–21818, 2024.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: Unified real-time end-to-end vision models. arXiv preprint arXiv:2606.03748, 2026.

Efstathios Karypidis, Spyros Gidaris, and Nikos Komodakis. Representations before pixels: Semantics-guided hierarchical video prediction. 2025a.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. DINO-foresight: Looking into the future with DINO. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025b. URL https://openreview.net/forum?id=gimtybo07H.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Advancing semantic future prediction through multimodal visual sequence transformers. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 3793–3803, June 2025c.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125, 2023.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 19948–19960, 2023.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Proceedings of the 2024 conference on empirical methods in natural language processing, pp. 5971–5984, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision, pp. 38–55. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26439–26455, 2024.

Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In Proceedings of the IEEE international conference on computer vision, pp. 648–657, 2017.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In International Conference on Learning Representations, volume 2024, pp. 51772–51783, 2024.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In Advances in Neural Information Processing Systems, 2023.

Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 17810–17820, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

Enrico Pallotta, Sina Mokhtarzadeh Azar, Shuai Li, Olga Zatsarynna, and Juergen Gall. Syncvp: Joint diffusion for synchronous multi-modal video prediction. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 13787–13797, June 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 2021.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022.

Shuhuai Ren, Shuming Ma, Xu Sun, and Furu Wei. Next block prediction: Video generation via semi-autoregressive modeling. arXiv preprint arXiv:2502.07737, 2025.

Noam Shazeer. GLU variants improve transformer. ArXiv, abs/2002.05202, 2020.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022.

Anni Tang, Tianyu He, Junliang Guo, Xinle Cheng, Li Song, and Jiang Bian. Vidtok: A versatile and open-source video tokenizer. arXiv preprint arXiv:2412.13061, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems, 37:84839–84865, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. arXiv preprint arXiv:2410.21264, 2024a.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. Advances in Neural Information Processing Systems, 37:28281–28295, 2024b.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems, 36:7594–7611, 2023.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In International Conference on Learning Representations, volume 2024, pp. 42055–42079, 2024c.

Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models, 2024d. URL https://arxiv.org/abs/2410.02757.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114, 2023.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In European Conference on Computer Vision, pp. 521–539. Springer, 2022.

Mingqiao Ye, Zhaochong An, Zhitong Gao, Xian Liu, Oğuzhan Fatih Kar, Jesse Allardice, Roman Bachmann, David Mizrahi, François Fleuret, Chuan Li, et al. Modus: Decoder-only any-to-any modeling of diverse modalities. In Forty-third International Conference on Machine Learning, 2026.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David C. Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. 2023. URL https://api.semanticscholar.org/CorpusID:263830733.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 3836–3847, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2024. URL https://arxiv.org/abs/2411.04983.

APPENDIX

A PROJECT SUMMARY VIDEOS

To provide a visual summary of our work, we include two videos in the supplementary material, and we kindly invite reviewers to view them. `presentation.mp4` gives a full overview of our research, including generation examples from Video-4M. `project_rundown.mp4` walks through the model main capabilities, which shows qualitative examples of (i) any-to-any generation, (ii) chained generation, (iii) forecasting in different modality spaces, and (iv) counterfactual generation. We hope these videos complement the main paper.

B QUALITATIVE RESULTS FOR COUNTERFACTUAL GENERATIONS

In Sec. 4.4, we showed different use cases of Video-4M for video editing and control using a diverse set of input controls. Those examples used chained generation (i.e., coarse-to-fine-grained chain design, as discussed in Sec. 3.4). Here, we provide additional qualitative results and show comparisons with the direct-generation inference strategy. Fig. 10 shows generation results conditioned on the first RGB frame and caption. Videos from direct generation do not strongly adhere to the input conditions defined by captions, while in chained generation, conditioning using captions is more strongly reflected in the generated frames. These observations are consistent with the chained-generation results for standard RGB video generation tasks, which we discussed in Sec. 4.2.

C ADDITIONAL EXPERIMENTAL RESULTS

In this section, we provide additional experiments related to Video-4M pretraining ablations, chained generation, and additional discussion on comparisons with external specialist models.

C.1 MASKING STRATEGIES FOR PRETRAINING Video-4M

For pretraining Video-4M, we use a combination of different masking modes. As discussed in Sec. 3.3, in addition to random masking, we introduce causal masking modes that incentivize forecasting across both sequential (e.g., bounding boxes, transcriptions, etc.) and non-sequential modalities (e.g., RGB, Depth, etc.). The frame-conditioned modes (i.e., FCFF/FCAF for dense modalities and CFP for sequential ones) mask all future context tokens and are therefore harder than random masking. Under random masking, the model can interpolate a masked token from context on both sides in time, whereas causal masking forces it to predict from past context alone. This shows up in the training loss, which stays relatively higher when these modes are included, but overall optimization remains smooth across model scales (both Video-4M Base and Video-4M Large). Their effect appears in downstream evaluations.

Tab. 5 shows ablations of these design choices using the Video-4M Base model. Removing causal masking modes for non-sequential modalities improves Caption→RGB generation in isolation, but significantly degrades video prediction performance. This trend is expected, as future-prediction masking modes are more aligned to the frame-conditioned generation task we perform at inference. Our final pretraining uses a combination of random and future-prediction masking, which provides the best performance trade-off across tasks.

Table 5: **Ablation on future prediction masking modes.** Incorporating masking modes that target future prediction during pretraining improves performance across generation and prediction tasks.

Configuration	Video Generation	Video Prediction
	Caption→RGB FVD ↓	1-frame FVD ↓
Video-4M (default)	<u>310.0</u>	**238.0**
Video-4M - future pred. modes (non-seq. modalities)	**289.0**	<u>331.0</u>
Video-4M - future pred. modes (seq. modalities)	316.0	240.0

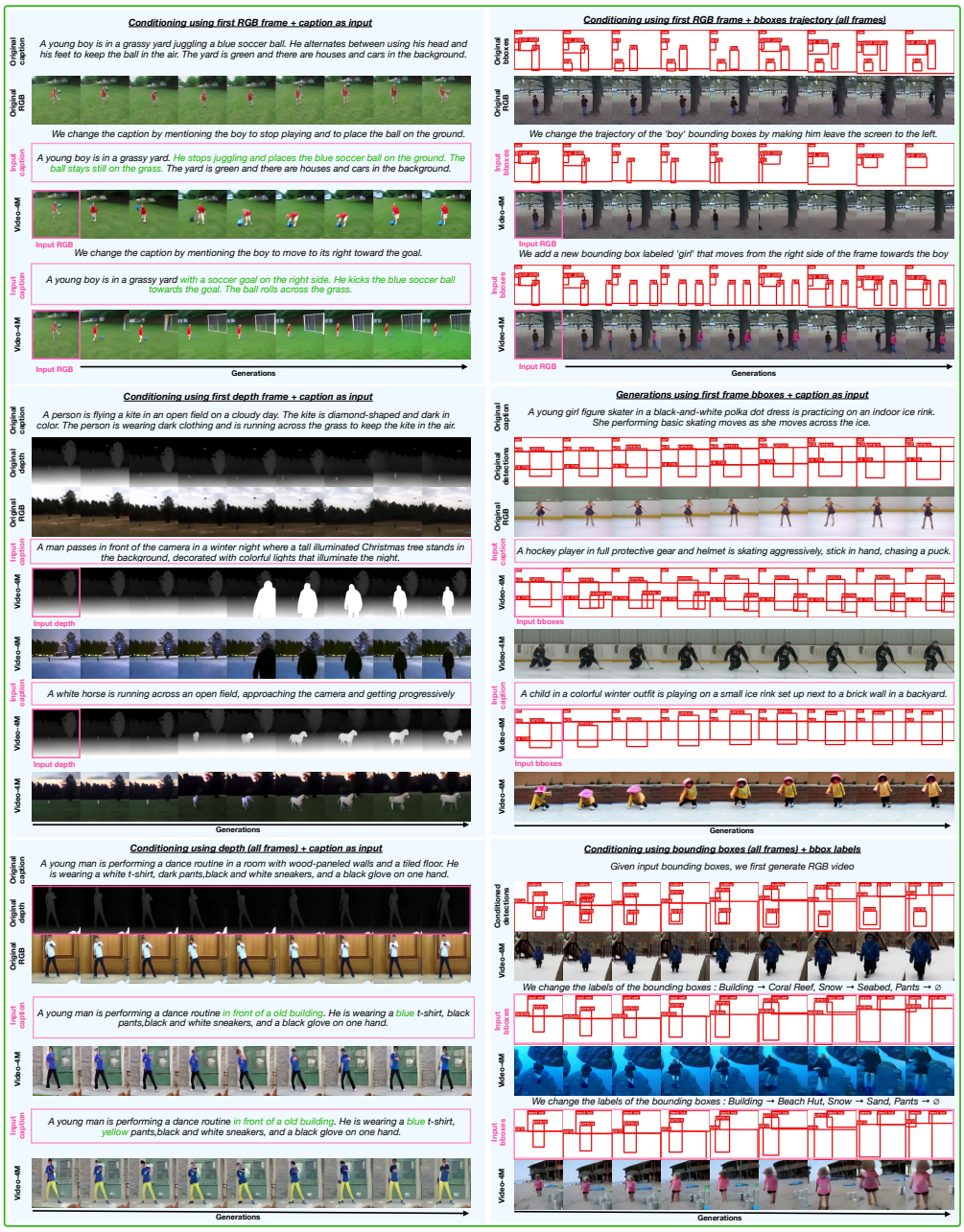

Figure 8: **Fine-grained and controllable video generation using multimodal input conditioning.** We showcase several instantiations of video generation through multimodal input probing. Video-4M allows any combination of modalities as input to generate output videos. In most scenarios, we provide one conditioning modality (e.g., first frame of depth/RGB/bounding boxes), and vary the output generation using a probing modality (e.g., caption/bounding box labels or coordinates).

C.2 Comparison with external specialist models.

In this section, we provide zero-shot evaluation comparison results with specialist models on standard video tasks (i.e., RGB to others). We consider 7 different video tasks including video generation, prediction, captioning, depth, and normal estimation.

Evaluation settings. As external specialist models (Wang et al., 2024a;b; Ren et al., 2025) for video generation and prediction use a different training setup (i.e., training on 17 frame videos that cover a

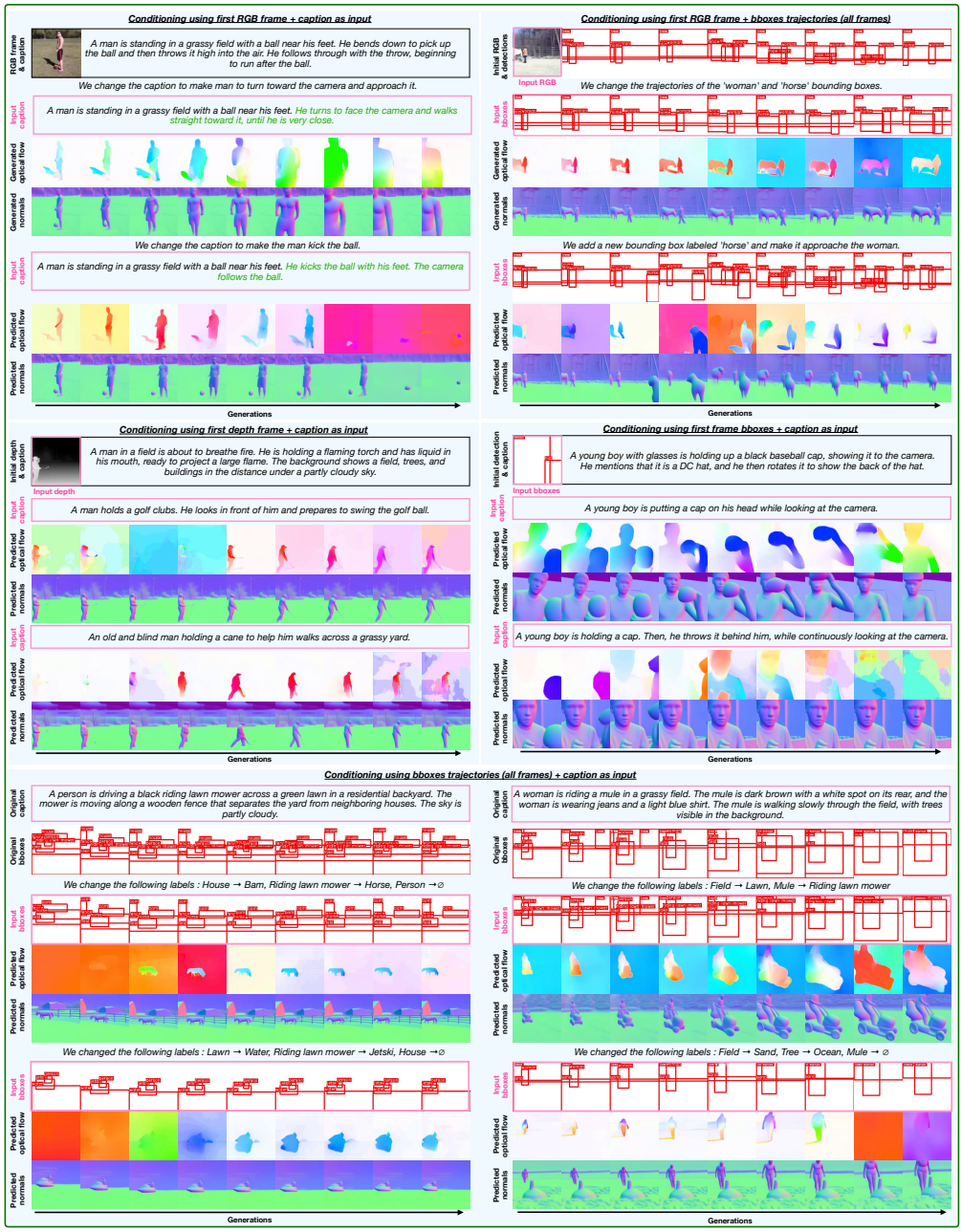

Figure 9: **Multimodal input probing for controllable generations in non-RGB modality spaces.** We extend our probing analysis to non-RGB modality spaces as outputs. For instance, we can control how surface normals in the future will look by varying the probing modality, such as captions, the first RGB frame, or bounding boxes.

0.8s time horizon sampled at 30 FPS), we re-trained those models in our setting (i.e., 17 frames with a 4-second video sampled at 4 FPS), using their official training code repositories. For depth and normal estimation, models such as Video Depth Anything (Chen et al., 2025) and NormalCrafter (Bin et al., 2025) are trained with high-resolution videos and images, whereas Video-4M operates on tokenized inputs of 128^2 resolution. We therefore compare with two pseudolabeler variants: one evaluated at their native resolution and one that processes high-resolution inputs through downsampling followed by upsampling with a tokenizer reconstruction offset, to better reflect the constraints under which



Figure 10: **Caption-conditioned with direct and chaining generations.** Given a single RGB frame and a caption text as inputs, we use Video-4M to generate different RGB trajectories by changing the caption. Notice how the chaining improves the generates.

Video-4M operates. For the video captioning task, we simply resize input videos to 128x128, followed by upsampling to 224x224 resolution videos for the specialist models.

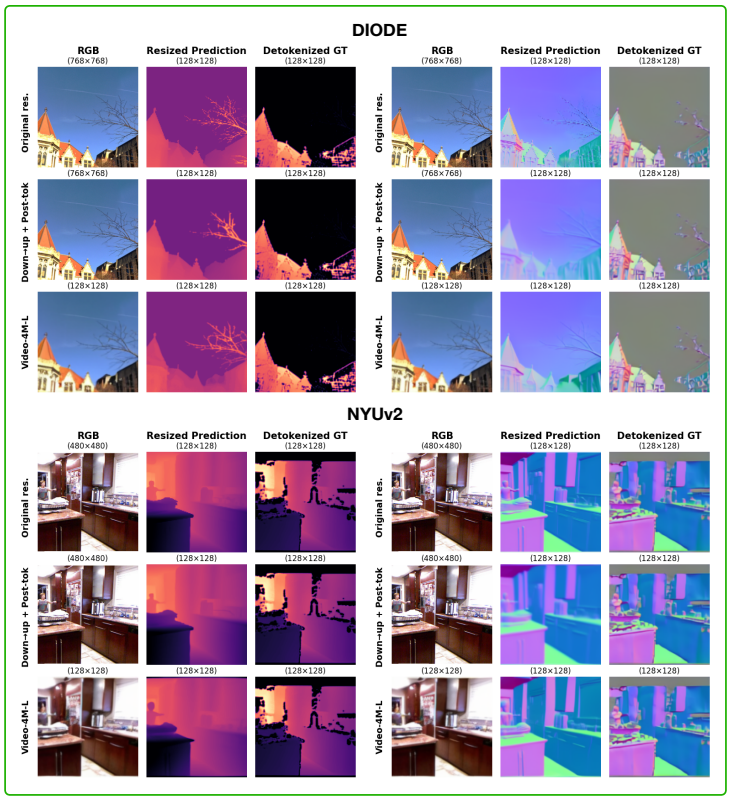

Figure 11: **Zero-shot qualitative examples for depth (left) and surface normal (right) estimation tasks.** For pseudo-labelers, we use the following configuration for comparable evaluation settings: *downsampling of input video to 128×128, resampling to the original resolution before prediction, and passing the prediction through the tokenizer with a reconstruction offset*. The tokenized GT represents the reconstruction upper bound after tokenization.

Results are shown in Tab. 1. On video generation and prediction tasks, LARP (Wang et al., 2024a) performs best in terms of FVD, while Video-4M shows the best performance on text similarity and class-alignment metrics. As LARP tokenizes videos into 1D sequences, a single second stage generation model cannot be trained for arbitrary frame-conditioned forecasting tasks. We therefore train a separate LARP model for both the 1-frame and 5-frame conditional benchmarks. In contrast, Video-4M uses a single network that supports all tasks. On depth and normal estimation, Video-4M performs competitively with external methods when evaluated under matched evaluation setups. Qualitative results are shown in Fig. 11. Overall, Video-4M, pretrained across 12 modalities, stays competitive on established mappings while allowing higher-order video modeling capabilities (e.g., chained generations, multimodal forecasting etc.) along that modality time grid that are not applicable to specialized models.

C.3 Additional experiments on chained generations

Here we present additional ablation studies, analysis, and qualitative results for chained generations with Video-4M.

Reversing the chain order. In Sec. 3.4, we showed that changing the order of modalities in the chain affects the generation results. Fig. 14 shows generation results for the coarse-to-fine chain and its reverse-order variant. We observe two consistent trends. First, when geometric modalities are placed at the start of the chain, the generated videos struggle to follow the semantics of the input condition. Second, the generated videos exhibit relatively less motion compared to the default coarse-to-fine chain design. We hypothesize that inferring semantic content from an input condition (e.g., a caption) is harder when conditioning first on low-level modalities like depth than when conditioning on

semantic modalities such as SigLIP 2. This is because geometric modalities emphasize low-level details such as edges and boundaries instead of encoding the high-level semantics carried by the input condition.

Despite the presence of semantic modalities later in the chain, their ability to guide generation is limited since the generated depth map has already defined the overall video trajectory. During training, modalities from the same training sample are always mutually consistent; the model, therefore, learns that subsequent modalities in the chain should be consistent with the earlier ones. This results in limited flexibility to steer the video content once the video scene layout has been determined by the earlier modalities in the chain.

Progressively adding modality groups in chain design. In Tab. 7, we study the effect of progressively adding modality groups in our coarse-to-fine chain design. Using only the sparse modality group, which includes caption and transcription, leads to the lowest overall performance. Including the semantic group alongside the sparse group significantly improves performance and leads to the best FVD. Further adding the geometric modality improves the top-1 classification score while slightly reducing the FVD.

Table 6: **Geometric modalities in chaining improves edge-based metrics.**

Chain	FVD ↓	Top-1 Acc. ↑	RGB Edge Sharp. ↑	Depth Edge Sharp ↑
Sparse → Semantic → Geometric (default)	225.2	**70.1**	**70.81**	**18.47**
Sparse → Semantic	**215.6**	69.3	68.39	17.81

We next study how the geometric modality in the chain design affects generation beyond what FVD captures. As observed from our qualitative analysis, adding a geometric modality to the chain improves low-level structural details in the generated videos. We show generation comparisons in Fig. 12 (top). Using identical hyperparameters and seed settings, we generate two RGB videos, one using depth in the chain and the other without. Adding depth improves low-level frame details such as background clarity, clearer inter-object boundaries, and object sharpness; however, this improvement is not reflected in FVD. We hypothesize that this is because FVD relies on I3D features (Carreira & Zisserman, 2017) and is mainly tuned for action recognition, and is therefore less sensitive to fine-grained spatial improvements such as object sharpness and boundary clarity.

Table 7: **Coarse-to-fine chain design ablation.**

Chain	FVD ↓	Top-1 Acc. ↑
Sparse only	425.6	41.4
Sparse → Semantic	**215.6**	69.3
Sparse → Semantic → Geometric	225.2	**70.1**

In Fig. 12 (middle), we generate depth maps using Video-4M on videos that were generated without depth in the chain, and compare them with depth maps from the chained generation. Fine-grained differences are more noticeable in the depth space than in RGB. Depth maps produced by the chain that uses depth contain more fine-grained and low-level structural details. To quantify this, we use an edge sharpness metric that detects edges in the generated videos with a Sobel filter and averages the resulting edge magnitude. We compute this metric on both depth maps and RGB videos and report the results in Tab. 6. As expected, the edge sharpness metric is higher for the chain that uses depth than for the no-depth chain design. We visualize the edge detections in Fig. 12 (bottom). For example, in the hammer throw action category case, the chain design using depth generates fine-grained boundaries of the protective net commonly seen in this sport, which are absent from generations that did not use depth in the chain.

Table 8: **Ablating modality groups in the chain design.**

Chain	FVD ↓	Top-1 Acc. ↑
Sparse → Semantic → Geometric (default)	<u>225.2</u>	<u>70.1</u>
Leave-one-out		
Sparse → Geometric	496.0	47.9
Sparse → Semantic	**215.6**	69.3
Semantic → Geometric	242.0	**79.2**

Ablating on the chain design: Tab. 8 shows results for ablating modality groups in our coarse-to-fine chain design. Consistent with our earlier analysis, removing semantic modalities from the chain leads

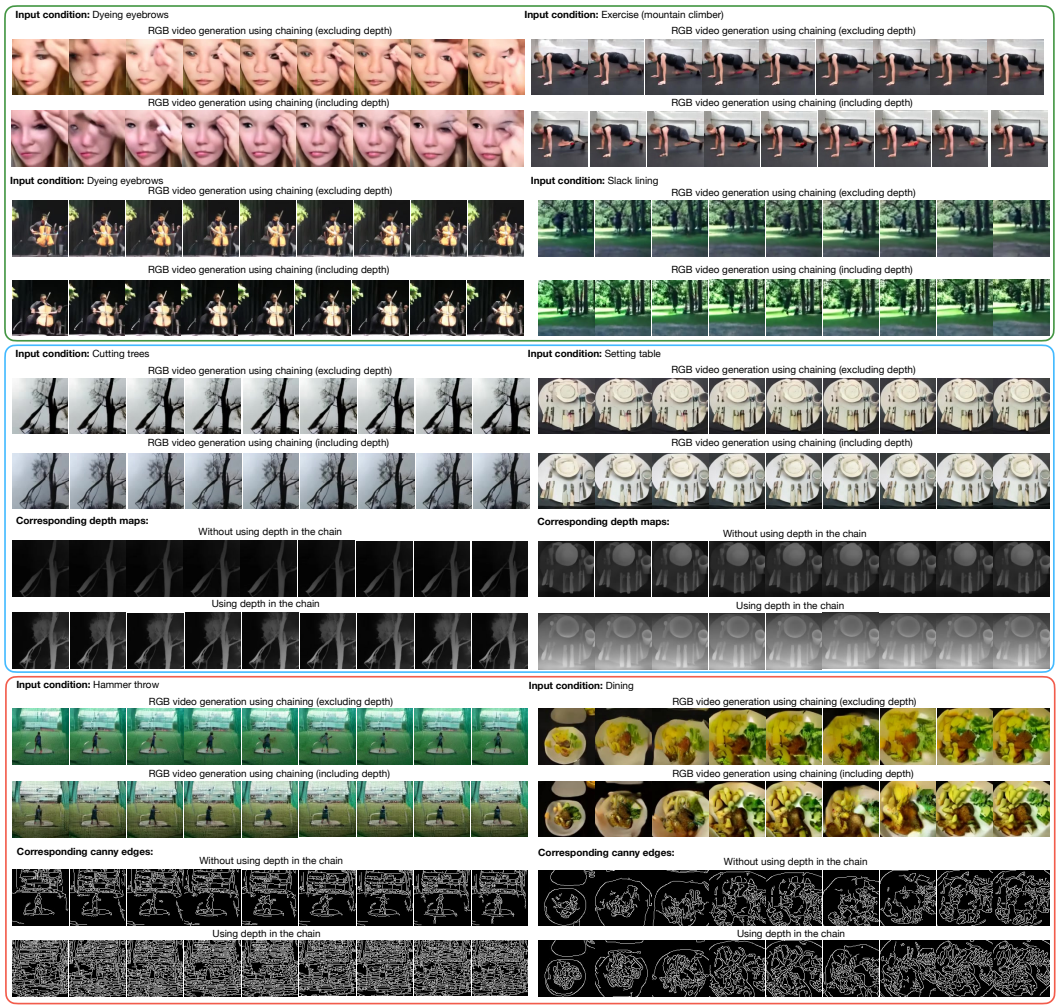

Figure 12: **Effect of including depth in the chain.** Top: RGB generations for different input conditions using the coarse-to-fine chain with and without depth in the chain, under identical hyperparameter and seed settings. Middle: depth maps corresponding to the two chain variants. Bottom: Canny edge maps of the generated videos. Including depth improves low-level structural details such as object boundaries, background clarity, and sharpness.

to lowest generation results. Overall, using all modality groups results in the best trade-off between FVD and top-1 accuracy.

Qualitative results. In Sec. 4.2, we presented qualitative results for class and caption conditional RGB generation tasks. In Fig. 13, we show results for 5-frame conditional generations for both chained and direct generations using Video-4M. Overall, chained generation leads to higher-quality video generations, with high consistency among predictions across different modalities in the chain.

D Additional results on multimodal forecasting

D.1 Oracle-independent Mixture of Modalities forecast

Motivated by the observation that the best forecasting modality varies across samples, we conduct an additional experiment using a modality-mixture forecast. Because the best modality for a given sample is not known at inference time, we form one forecast distribution by combining rollouts from multiple modalities. The composition is selected based on a validation set disjoint from the test set. The resulting mixed-modality distribution achieves a better energy score than any direct forecast with

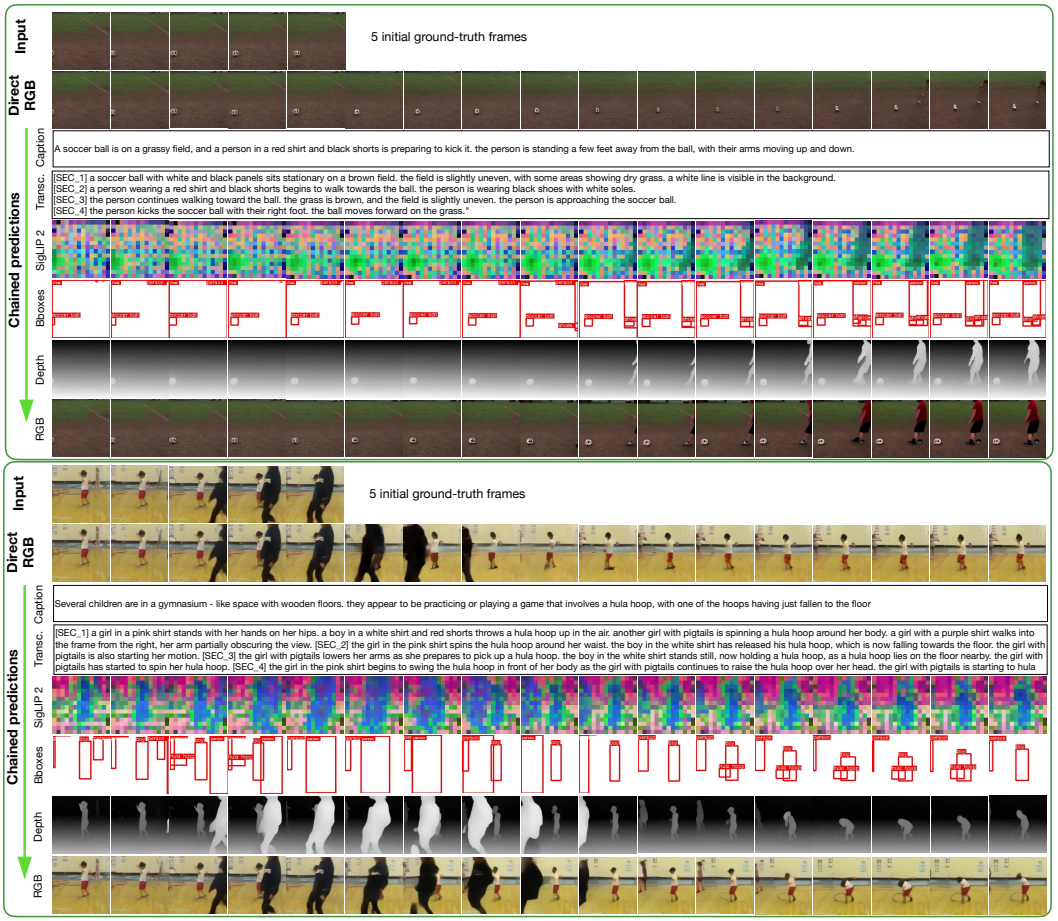

Figure 13: **Chained vs. direct predictions for 5-frame conditional RGB generation.** Given 5 initial ground-truth RGB frames as input, we compare direct RGB generation with chained generation, where Video-4M sequentially predicts caption, transcription, SigLIP 2 features, bounding boxes, and depth as intermediate modalities before generating RGB. The high-level textual modalities establish the global context of the future frames, semantic modalities outline object layout and motion, and the geometric modality refines low-level structure, leading to more coherent generations than the direct baseline.

matched validation search budget (Fig. 15). Note that this modality-mixture forecast is only enabled by a model such as Video-4M.

D.2 Energy scores

Gneiting & Raftery (2007) propose proper scoring rules for evaluating probabilistic forecasts, i.e., functionals that assign a score to a predictive distribution against the (single) real outcome that is actually observed. A scoring rule is called strictly proper if in expectation it is uniquely optimized when the predicted distribution matches the true distribution. They introduce the energy score, a strictly proper scoring rule for multivariate outcomes that generalizes the continuous ranked probability score (CRPS). Crucially, the energy score allows evaluation of a predictive distribution from a single observed realization. Thus, rather than requiring access to multiple ground-truth futures for the same past, which would be impossible to obtain, it compares the single observed rollout to samples drawn from the model's predictive distribution, while also accounting for the internal dispersion of those samples. This makes it well-suited for forecasting settings, where each past sequence is associated with only one realized future, but the model is expected to represent a distribution over many plausible futures.

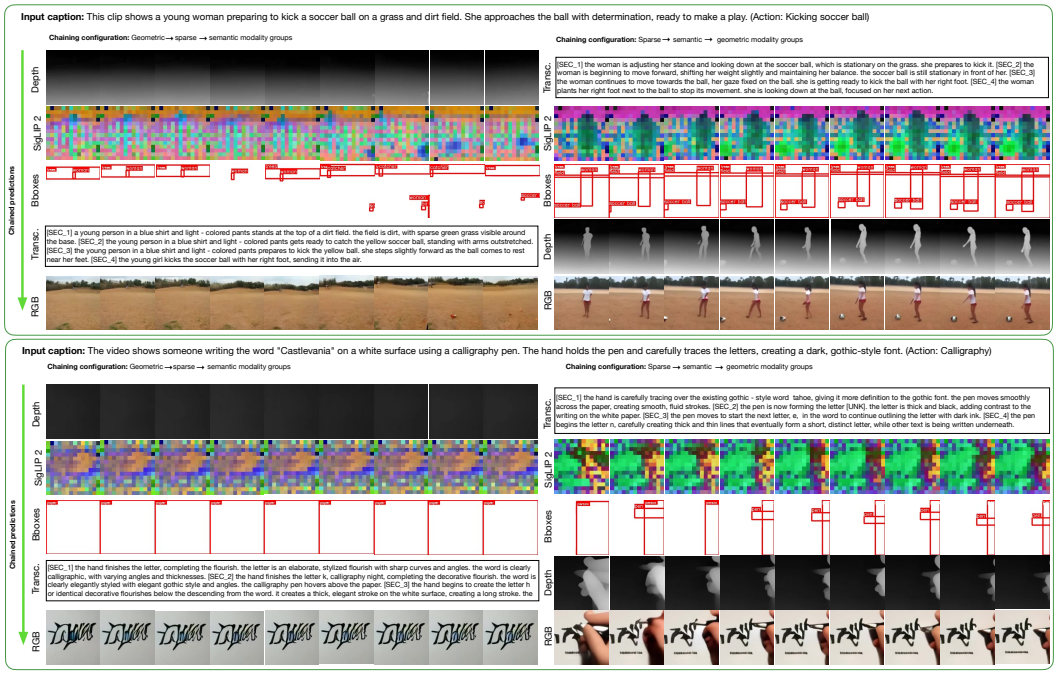

Figure 14: **Effect of reversing the chain order.** For the same input caption, we compare generations using the reversed chain (left: Geometric → Semantic → Sparse) and the default coarse-to-fine chain (right: Sparse → Semantic → Geometric). When the geometric modality is predicted first, the early depth prediction constrains the subsequent modalities, and the resulting videos struggle to follow the semantics of the input condition and exhibit less motion compared to the coarse-to-fine ordering.

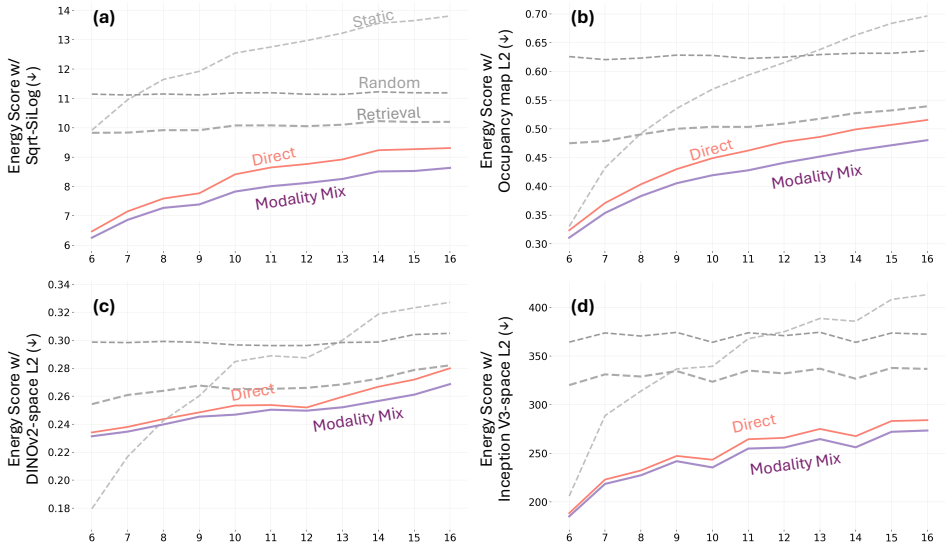

Figure 15: **Forecasting with a Mixture of Modalities.** Future frames of each target modality (a) Depth, (b) Bounding Boxes, (c) DINOv2, and (d) RGB are predicted given 5 starting frames of RGB. Modality Mix curves show energy scores achieved by a predicting the future frames using a mixture of modalities e.g. bounding boxes + transcripts + SigLIP 2 + etc. The mixture is determined by a validation set. Other curves are defined the same as in Fig. 7. While the energy scores are not as low as that of the per-sample best chain, Modality Mix does not require the use of a best-chain oracle or a predictor.

Given sampled prediction rollouts $\hat{x}^{(1)}, \ldots, \hat{x}^{(K)}$ and the single real ("observed") rollout x, we estimate the energy score as

$$\mathrm{ES} = \frac{1}{K}\sum_{i=1}^{K} d(\hat{x}^{(i)}, x) - \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} d(\hat{x}^{(i)}, \hat{x}^{(j)}),$$

where d is a task-appropriate distance between rollouts. For the energy score to define a proper scoring rule, the distance metric d must be of negative type (Gneiting & Raftery, 2007). The first term measures how close the predicted futures are to the realized future, while the second term corrects for the spread of the predictive distribution. Intuitively, the score does not require observing all plausible futures for the same past; instead, it evaluates the distance from the realized future to the model's sampled predictive distribution, with a correction for the distribution's own sample-to-sample dispersion so that, in expectation, the score favors calibrated predictive distributions over collapsed or artificially overdispersed ones.

D.3 Qualitative results

Figure 16 shows forecasts made in different modality spaces: SigLIP 2, Bbox., and Normal using four seeds each. To interpret the results, the forecasts are rendered back into RGB space using `Video-4M`.

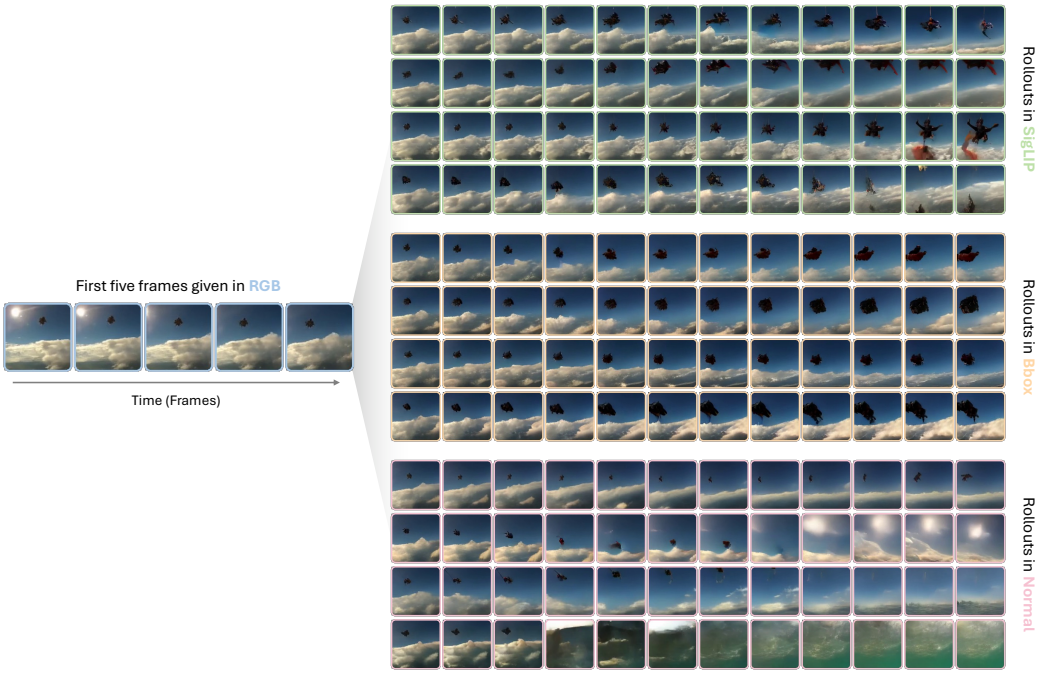

Figure 16: **Rollouts of Future Frames (Forecasting) in Different Modalities.** Future frames are rolled out in SigLIP 2, Bbox., and Normal modality spaces, from the same five initial RGB frames, using 4 different seeds each. Forecasts in SigLIP 2 tend to preserve the semantics of the scene (skydiving and parachute), Bbox. the object locations, Normal the surface characteristics (sometimes at the cost of semantic change from clouds to water surface). Note, this is an example selected to intuitively demonstrate the effect that the forecasting modality space can have; for other cases, the effect of forecasting modality is not always obvious or interpretable. However, this visualization is intended to strengthen the intuition behind how multimodal forecasts allow for exploration of a wider space of possible futures and add support to the finding in Fig. **??** that the best forecasting modality depends on the sample and the task at hand.

E Pseudolabeling

No existing video dataset provides aligned ground truth across multiple modalities at the scale on which our pretraining is based (approximately 400K videos across 12 modalities, which amounts to

34B tokens). As discussed in Sec. 3.1, we use pseudo-labeling to construct our K600-MM pretraining dataset. Starting from the Kinetics-600 dataset (Carreira et al., 2018), we obtain additional modalities through pseudolabeling using task-specific models. Tab. 9 provides an overview of all the modalities supported by Video-4M. Below we provide details on how different modalities are pseudo-labeled.

Modalities with a dense 2D structure. For modalities like feature maps, depth, normals, and optical flow, we directly use off-the-shelf models to generate them. During pseudolabeling, we generate modalities using the native resolution of the videos and later perform spatial resizing in the modality space to obtain 128×128 resolution samples. For feature maps, we use the default resolution of the neural network (224 for SigLIP 2 (Tschannen et al., 2025) and DINOv2 (Oquab et al., 2023), and 256 for VJEPA 2 (Assran et al., 2025)), as inference at 128×128 resolution would result in out-of-distribution inputs for these models.

Modalities that are represented sequentially. For generating captions and transcriptions, we use Gemini-2.0 Flash (Google, 2023) by providing the input video alongside a prompt instruction. Additionally, we obtain metadata and object tags for each video in our dataset.

Object tags consist of category names of active (e.g., objects involved in the main actions) and passive (e.g., static or background objects) entities. For the bounding box modality, we use object tags obtained from Gemini as detection labels for Grounding DINO, which generates the bounding boxes. This diverse nature of object tags effectively enables open-vocabulary detection capabilities of Video-4M.

For human poses, we first generate 3D human poses using TokenHMR (Dwivedi et al., 2024). We noticed that the pseudolabelers also generate complete human poses for partially visible human instances in the videos (e.g., where only hands or fingers are visible). Additionally, there were false positive predictions for out-of-distribution videos. We filter such bounding boxes by first estimating 2D poses (Jocher et al., 2026) on the predicted humans and retaining only those bounding box instances for which the 2D keypoint prediction confidence for the nose, shoulders, and hips is above 0.85. This significantly reduced the number of human pose predictions for partially visible human instances, for which complete pose prediction was deemed invalid.

Table 9: **Overview of Modalities**. We group different modalities into four categories based on the type of information they capture. In total, Video-4M supports 12 multimodal abstractions. For class category, we learn a token embedding layer.

Modality	Pseudo Labeler	Tokenizer
RGB	N/A	ST FSQ
Geometric modalities		
Depth	VDA-v2 (Chen et al., 2025)	FSQ
Surface Normal	NormalCrafter (Bin et al., 2025)	FSQ
3D Human Pose	TokenHMR (Dwivedi et al., 2024)	VQ-VAE
Motion modality		
Optical Flow	DPFlow (Morimitsu et al., 2025)	FSQ
Semantic modalities		
SigLIP 2	SigLIP 2 (Tschannen et al., 2025)	FSQ
DINOv2	DINOv2 (Oquab et al., 2023)	FSQ
V-JEPA 2	V-JEPA 2 (Assran et al., 2025)	FSQ
Detection	Grounding DINO (Liu et al., 2024)	WordPiece
1D/High-Level Modalities		
Caption	Gemini-2.0 Flash (Team et al., 2023)	WordPiece
Class category	N/A	Embed.
Transcriptions	Gemini-2.0 Flash (Team et al., 2023)	WordPiece

F TOKENIZATION

Video-4M is pretrained on the tokenized K600-MM dataset. As discussed in Sec. 3.2, we train modality-specific discrete tokenizers. In Fig. 17, we show the reconstruction quality of trained tokenizers for different modalities. As we sample videos at 4 FPS with a time horizon of 4 seconds, most of our training samples contain abrupt motion changes (e.g., a person in a standby position in one frame and in a kicking position in the very next frame), which are more challenging for the tokenizer to reconstruct, compared to video samples where the motion change is more gradual (e.g., a person using a sewing machine). We qualitatively show reconstructions using these two representative cases. For modalities such as feature maps, depth, or optical flow, reconstructions are highly consistent with the raw modalities. For the RGB modality, reconstructed videos contain small artifacts when fast or abrupt motion changes occur. Note that each tokenizer is fine-tuned on the K600-MM dataset using its respective modality-specific data.

F.1 TOKENIZATION OF SEQUENTIALLY REPRESENTED MODALITIES

Unlike modalities that have a dense 2D structure, modalities like human poses and bounding boxes are represented sequentially and use different tokenization strategies. We provide their tokenization details below.

Bounding boxes. Each bounding box coordinate is quantized into one of 1000 uniform bins per x and y axis, and every bin is assigned its own dedicated special token, added to the vocabulary alongside the WordPiece subword tokens. WordPiece encoding is applied only to the object class names within bounding box sequences. This enables open-vocabulary detection capabilities rather than predicting over a fixed category list. A detection sequence therefore consists of coordinate-bin tokens, class-name subword tokens, and frame-marker special tokens such as [FRAME_1], all drawn from one shared vocabulary. We found this approach to be scalable to any arbitrary number of detections as well as frames. This tokenization strategy is also widely adopted in the literature, including prior sequence-based detection work such as Pix2Seq (Chen et al., 2021) and UniTAB (Yang et al., 2022), as well as image-based any-to-any models (Mizrahi et al., 2023; Bachmann et al., 2024).

Human poses. For human poses, we use a learned VQ-VAE tokenizer from 4M-21 (Bachmann et al., 2024), which tokenizes each pose into 8 tokens, as detailed in Tab. 9. We apply coordinate encoding to the global orientation, body shape, and bounding box of each person using 1000 discrete bins, which are likewise added to the vocabulary as special tokens. The final sequence is formed by prepending identifiers (bbox, pose, shape) to the corresponding sub-sequences within each frame, with frame-wise pose sequences separated by frame-marker special tokens such as [FRAME_1].

G Experimental setup details

G.1 Evaluation details.

For RGB generation tasks, we use Fréchet Video Distance (FVD) (Unterthiner et al., 2019) as the primary metric, computed on 2,000 videos from the test set. Results are averaged over three seeds, with each seed randomly sampling 2,000 videos from the test set; the sampled videos for each seed remain consistent across experiments. For multimodal forecasting experiments, we similarly evaluate on 2,000 videos, but average results over 16 random seeds. In this setup, we use the same set of 2,000 videos for all 16 seeds. We additionally report classification accuracy and a text similarity score as alignment metrics with respect to the input condition. Following VBench (Huang et al., 2024), we use a UMT-L (Li et al., 2023) model finetuned on Kinetics-600 classification to measure class-video alignment (using the first 16 of 17 frames), and ViCLIP-InternVid-10M-FLT (Wang et al., 2024c) to measure text-video alignment (subsampling 8 of 17 frames with a temporal stride of 2). In both cases, videos are interpolated to 224×224 resolution.

G.2 Pre-training details

Implementation details. Video-4M is trained at three model scales: Video-4M-B with 198M parameters, Video-4M-L with 705M parameters, and Video-4M-XL with 2.8B parameters. For Video-4M-B, we use a batch size of 11,200 (70 per GPU $\times$ 40 GPUs $\times$ 4 accumulation steps) and train for 1,500B tokens, taking approximately 1 day on H200 GPUs. For Video-4M-L, we use a batch size of 11,520 (36 per GPU $\times$ 80 GPUs $\times$ 4 accumulation steps) and train for 1,500B tokens over approximately 2 days. For Video-4M-XL, we use a batch size of 36 per GPU across 20 nodes with 4 gradient accumulation steps, resulting in an effective batch size of approximately 6,400, and train for approximately 10 days (covering 1,500B training tokens). Our implementation is based on PyTorch, using Distributed Data Parallel (DDP) for Video-4M-B and Fully Sharded Data Parallel (FSDP) for Video-4M-L and Video-4M-XL. We pretrain all variants of Video-4M from scratch. Pretraining hyperparameters are shown in Tab. 10.

H Limitations

While Video-4M demonstrates several advantages of multimodal video modeling, several limitations remain.

- **Error propagation in chained generations.** Our experiments in the main paper show that chaining through intermediate modalities improves over directly predicting the target modality. However, chains are susceptible to error propagation, and generation errors in early steps can compound and negatively affect subsequent modalities and the final output. Investigating strategies for

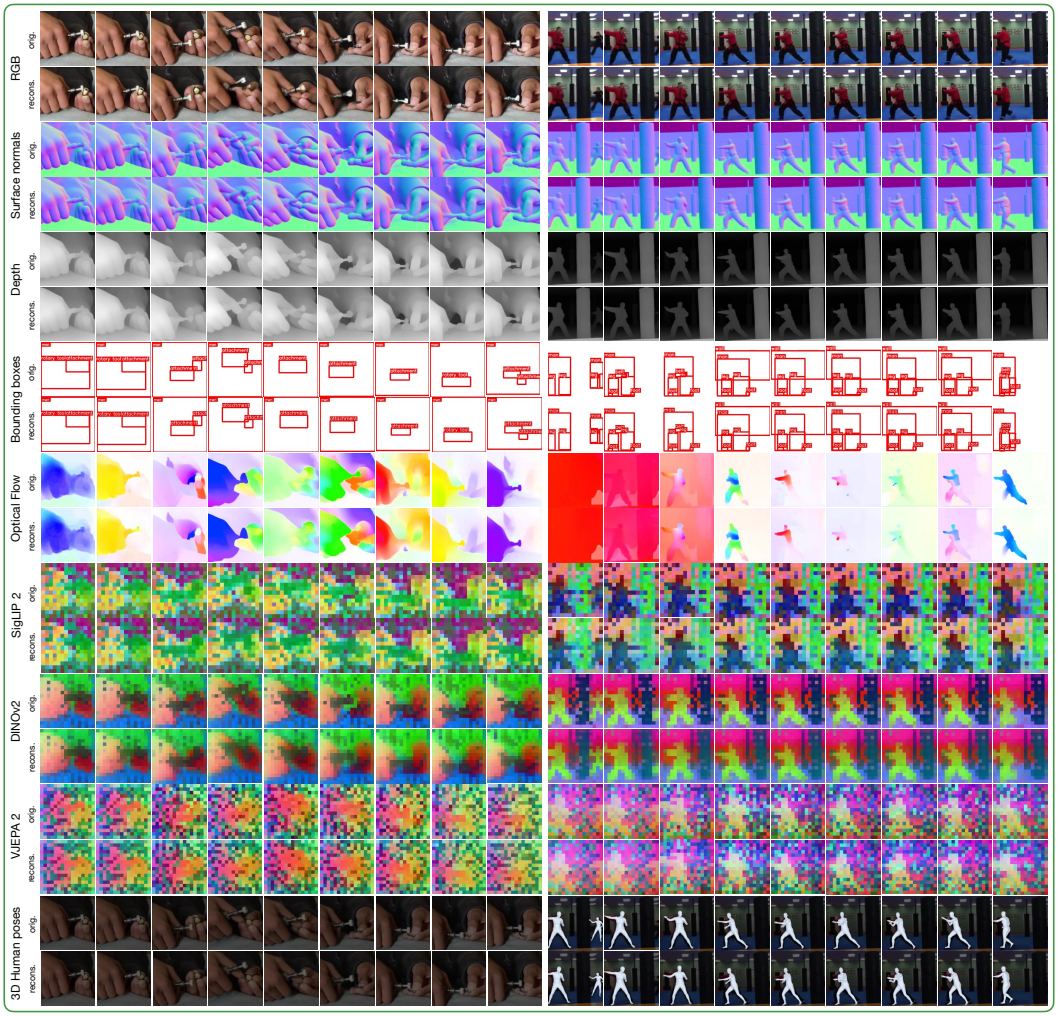

Figure 17: **Tokenizer reconstruction quality across modalities.** For each of the tokenized modalities, we show the original pseudolabeled input (orig.) alongside its reconstruction after tokenization (recons.), for two representative K600-MM clips: one with gradual motion (a person using a sewing machine, left) and one with abrupt motion changes (a person performing a kick, right). Reconstructions of semantic, geometric and motion modalities are highly consistent with the raw modalities, while RGB reconstructions contain notable artifacts when fast or abrupt motion occurs.

detecting and correcting such errors during chained inference is a promising direction for further strengthening the benefits of chaining.

- **Dependence on pseudolabeler and tokenizer quality.** We use SOTA or near-SOTA models for each modality for pseudolabeling and tokenization (Appendix E, F), but their quality is nonetheless bounded by these models. This also applies to tokenizers, as they are not lossless, and predictions are therefore bounded by their reconstruction quality. As pseudolabelers and tokenizers continue to improve, we expect Video-4M to directly benefit, since our framework is agnostic to the specific choice of pseudolabeler or tokenizer used for each modality.
- **Evaluation of video generation and forecasting.** Evaluating video generation and forecasting remains an open challenge. Standard video generation metrics such as FVD are not always fully representative of structural or fine-grained improvements: we show in Appendix C.3 that including depth in the generation chain measurably improves low-level structural details (e.g., object boundaries, edge sharpness), despite this improvement not being fully reflected in FVD. Similarly, current forecasting metrics assume a single ground-truth future, whereas multiple plausible futures

Table 10: **Pre-training settings for scratch initialization.**

Configuration	Base	Large	XL
Training length (n tokens)	1500B	1500B	1500B
Warmup length (n tokens)	10B		
Optimizer	AdamW (Loshchilov & Hutter, 2019)		
Opt. momentum	$\beta_1, \beta_2 = 0.9, 0.95$		
Base learning rate (Goyal et al., 2017)	2e-5		
Batch size	16.3K	14.3K	6.4K
Weight decay	0.05		
Learning rate schedule	Cosine decay		
Feedforward activation	SwiGLU (Shazeer, 2020)		
Input token budget	512	512	512
Target token budget	512	512	512
Input and target α	Mixture (Bachmann et al., 2024)		
Masking strategy	Mixture (Bachmann et al., 2024)		
Image resolution	128^2		
Augmentation	Center Crop		
Repeated sampling (Feichtenhofer et al., 2022)	4		
Data type	`bfloat16` (Burgess et al., 2019)		

may exist for a given observation. We believe developing better evaluation protocols for both video generation and forecasting is an important direction for future work.

I Broader Impact

As demonstrated in our paper, a `Video-4M` enables higher-order capabilities such as chained generation, multimodal forecasting, and fine-grained generation and editing via arbitrary modality-time conditioning, which specialist models do not support. We also believe the broader impact of our work extends beyond introducing any-to-any modeling to videos:

Relevance to other domains. We propose a thoroughly evaluated design to handle an additional sequential structure in data, which is directly relevant to other domains such as audio, time-series, or interleaved data. Prior any-to-any modeling methods were mainly developed for static images and do not offer clear design guidance for these domains. `Video-4M` addresses this gap, which required non-trivial design changes to handle the temporal axis.

A conceptual contribution to world modeling. Furthermore, we see this work as an important building block for world models operating in multimodal spaces. It is an open debate whether world modeling should happen in raw sensory space, such as RGB video, in a latent space, as in JEPA-style world models, or in other abstractions, such as semantics. This work proposes joint multimodal modeling as a principled approach that unifies the strengths of all, rather than an either-or answer. A joint multimodal model models one underlying physical reality through various modalities simultaneously, and enables traversing across them without hard-committing to one, so raw sensory observations, semantics, and latents can all be present at once. This shifts the question from which space to model in, to how to effectively model a rich and diverse hierarchy of modalities that accommodates all ranks of abstractions.