

A2A-Video: Modeling the World Across Time and Modalities

Muhammad Uzair Khattak¹ Won Jun Kim¹ Reza Abbassi¹ Michael Murphy¹
 Albias Havolli¹ Amir Zadeh³ Chuan Li³ Oğuzhan Fatih Kar¹
 Muhammad Ferjad Naeem² Andrei Atanov¹ Roman Bachmann¹
 Federico Tombari² Amir Zamir¹

¹EPFL ²Google ³Lambda

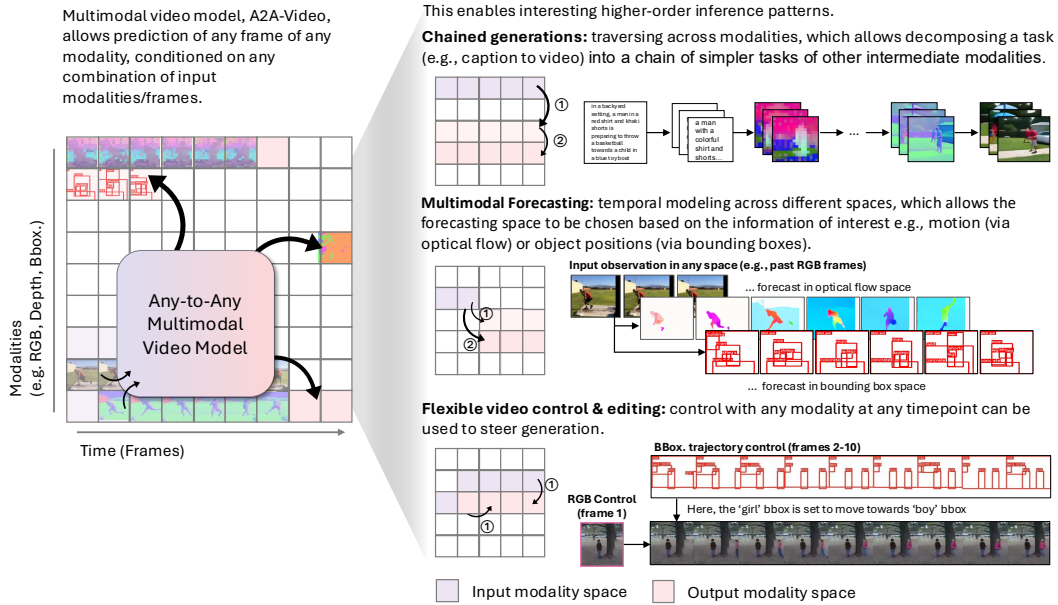


Figure 1: We introduce A2A-Video, an any-to-any multimodal video model that is able to predict any modality at any point in time, given any combination of observations from any other modalities. It is thus able to *traverse along both the modality and temporal axes*, enabling flexible, diverse inference schemes: **chained generations**, where we decompose a task, e.g., caption-to-video, into a chain of sub-tasks of other intermediate modalities; **Multimodal forecasting**, where given a sequence of past observations, the future can be forecasted in any of the modality spaces; and **Flexible video control & editing** where a generation rollout can be controlled by conditioning on any of the supported modalities specified at any point in time.

Abstract

Video modeling aims to capture the spatiotemporal relationships of an observation. The de facto approach is to represent and model these relationships in RGB space. While this can happen in RGB, doing so is challenging, as it requires simultaneously learning high-level semantics (e.g., what is happening and where) and low-level appearance (e.g., pixel-level details). It is also often unnecessary; for certain tasks, estimating where an object will be in the future is often sufficient for decision-making compared to generating its image in that future. While certain models have gone beyond RGB towards modeling higher-level spaces such as DINOv2, they

operate in a single representation space. We hypothesize that no single space is universally optimal, as different downstream use-cases require modeling different aspects of the future.

In this work, we investigate the role of multimodal modeling for video and develop A2A-Video, an any-to-any video model that supports a diverse set of modalities, including depth, surface normals, optical flow, bounding boxes, text, and neural network feature maps. A2A-Video allows traversal across both the modality and time axes simultaneously, establishing direct input-output mappings between all supported modalities and giving rise to higher-order inference paths that are not supported by specialist models. We explore three instantiations of this framework. First, in chained generation, we decompose a task (e.g., caption-to-video) into a chain of simpler tasks by predicting intermediate modalities; we find that this chained inference yields higher-quality generations than direct generation. Second, we study forecasting in multimodal abstract spaces and find that it can be more accurate for information such as distances (via depth) and object motion (via bounding boxes) than forecasting directly in the RGB modality. Finally, we demonstrate the flexible video control and editing capabilities of A2A-Video.

1 Introduction

Understanding and predicting how a scene evolves over time is a central goal of video modeling, which requires capturing the spatiotemporal relationships of an observation. Most video models operate in RGB space, with some variance in input conditioning formats such as caption, class labels, depth, and canny edges [48, 56, 52, 23, 12]. While this can happen directly in RGB space, we posit that RGB is not always the most suitable space to operate in - it entangles both low-level appearance (pixel details) with high-level semantic changes across time. This is a difficult task for a video model, as it aims to learn these two aspects simultaneously. Moreover, it is often unnecessary: for certain predictive tasks, estimating where an object will be in the future is often sufficient for decision-making and more actionable than generating its image in that future [3, 58]. Jointly formulating multiple spaces for video modeling and studying the consequences for the relationships between them is, therefore, of fundamental interest.

In this work, we investigate video modeling using abstract spaces beyond RGB, defined through multimodality. We first develop an any-to-any video model, A2A-Video (Fig. 1: left), pretrained with a multimodal masked modeling objective [36, 17]. A2A-Video supports a diverse set of modalities: depth, normals, optical flow, bounding boxes, text, and neural network feature maps, etc. These modalities capture different aspects of an underlying scene and naturally form a hierarchy of abstractions depending on the type and complexity of the information each captures: depth and normals, for instance, encode more geometric detail about a video than modalities such as captions or bounding boxes, which are more semantic. By introducing any-to-any modeling in videos, A2A-Video allows traversal across both the modality and time axes simultaneously and gives rise to higher-order, diverse inference paths that are not supported by specialist video models. This means inference itself, and the representation space, no longer need to be fixed in advance.

We study video-specific implications of A2A-Video in three settings (Fig. 1: right). In *chained generations* (§4.2), we show that decomposing a fixed one-to-one task, text-to-RGB video generation in our case, into a coarse-to-fine self-predicted chain of intermediate modalities (e.g., semantic to geometric to RGB) significantly improves results over both direct generation and a specialist text-to-video model trained in a controlled, matched-compute setup. Secondly, we explore *temporal modeling in abstract modalities* (§4.3) and find that it can lead to more accurate forecasting of information such as distances (via depth) and object motion (via bounding boxes), rather than generating a future video from the input observation and subsequently extracting the desired information from it. Finally, we demonstrate that A2A-Video allows control over video generation and editing (§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.

In summary, we make the following contributions: (1) we investigate any-to-any modeling in the video domain and develop A2A-Video. A2A-Video traverses both the modality and temporal axes simultaneously within a single model, and give rise to higher-order video-specific capabilities. We demonstrate these capabilities in three video task settings: (2) we study the benefit of chained

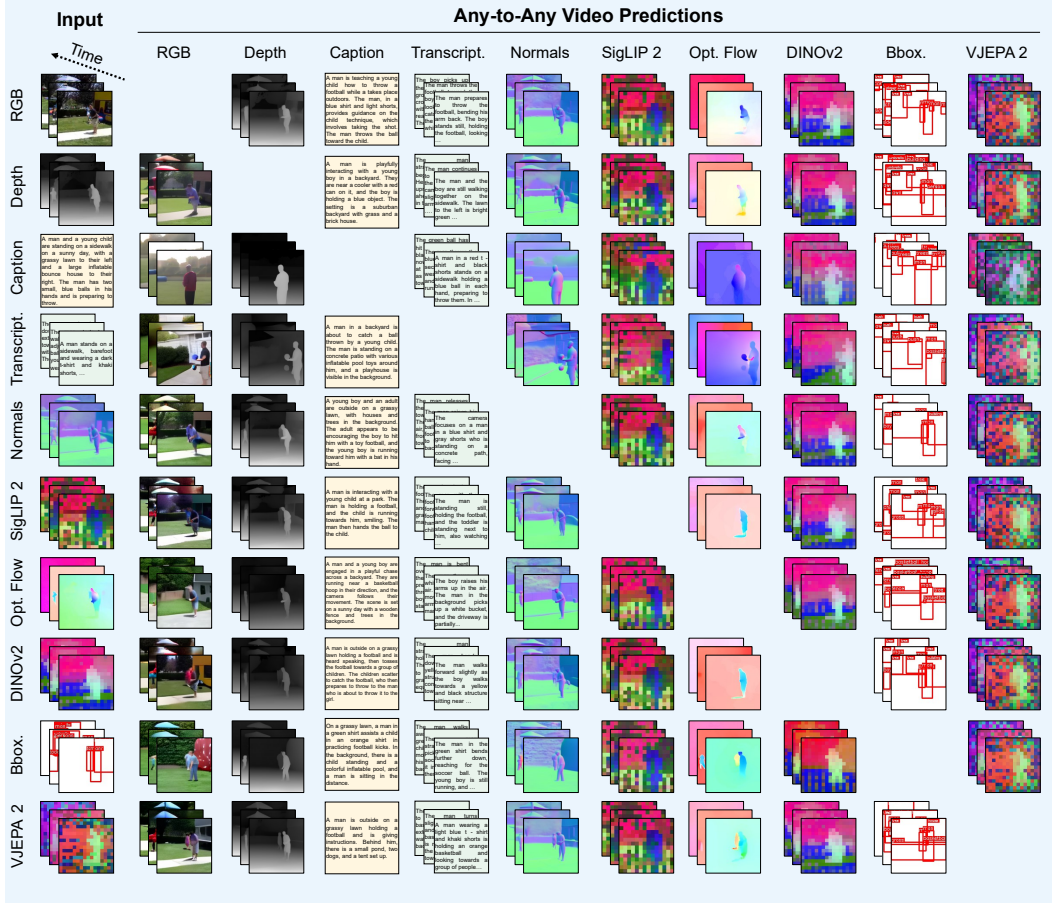


Figure 2: **Any-to-any capabilities.** A2A-Video allows predictions from any input modality to any target modality using a single model. Each row shows predictions in a chained manner from different input modalities, which all correspond to the same video sample. High consistency is observed across predictions, even from sparse modalities such as text $\rightarrow$ other and transcription $\rightarrow$ other modalities.

generations through intermediate modalities for video generation tasks; (3) we investigate forecasting in different modality spaces and show its advantages over modeling directly in RGB; and (4) we show that A2A-Video allows flexible control and editing of video generations by conditioning on an arbitrary set of modalities at arbitrary time points. The diverse set of modalities allows conditioning to range from coarse (e.g., a caption) to fine-grained (e.g., a depth map).

2 Related Work

Multimodality in video modeling. Many video models incorporate different modalities as conditioning inputs to guide video generation: text-to-video approaches [23, 43, 22, 56]; depth, edges, and human pose with ControlNet [57] and ControlVideo [12]; sketch sequences, reference videos, and motion vectors with VideoComposer [50]; and text, audio, optical flow, and depth with the LLM-based VideoPoet [29]. In these methods, most modalities serve only as conditioning inputs, while videos themselves are modeled in RGB space. Multimodal video models have also been explored, especially for robotics or driving, where videos are modeled in other modalities such as segmentation maps under specific settings [53, 58, 28, 39, 27]. However, there is comparably little investigation into unified and general multimodal settings that enable exploration of the modality axis in the video modeling domain.

Any-to-any models. These are class of models that can accept any modality from a supported set as input and generate any modality as output. These models aim to address the long-standing goals of model unification and multi-task learning. UnifiedIO [34] and 4M [36, 17] are powerful examples

in the image domain. To our knowledge, any-to-any modeling has not been explored in the video domain. A2A-Video adopts the design choices of 4M [36] and tailors it for the video domain.

Forecasting in abstract spaces. DINO-WM [58] and DINO-Foresight [26] explore modeling in feature spaces of pre-trained vision models. Chiu et al. [13] and Luc et al. [35] explore predicting the future in semantic segmentation space conditioned on input RGB frames. However, these methods are limited to operating in a single or a few modality spaces. A2A-Video supports multiple abstractions, and enables joint analysis of the modality and the temporal axis and adjusting the choice of modalities based on the downstream setting or a particular sample.

3 Method

Developing A2A-Video requires curating a multimodal video dataset, tokenizing different modalities, and designing a model architecture and pretraining framework for any-to-any video modeling. We discuss our approach to each in the following sections.

3.1 Multimodal pretraining video dataset

Multimodal video data with alignment across more than just video-caption pairs is scarce: no existing video dataset provides aligned ground truth across multiple modalities at the scale we use for pretraining (approximately 400K videos across 12 modalities). We overcome this by 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 [36], 4M-21 [17], and MODUS [55]. To ensure a high quality bar for pretraining data, we use SOTA or near-SOTA models for each modality.

Starting from video-class pairs from the Kinetics-600 [7] (600 action categories across $\sim 400k$ videos of ~ 10 seconds each), we generate 10 different modalities that can be categorized into 4 types: **(i) *geometric modalities*** including depth maps, surface normals and 3D human poses, **(ii) *motion modality*** including optical flow, **(iii) *semantic modalities*** including neural network feature maps from pretrained vision models and bounding boxes, and **(iv) *1D/high-level modalities*** including captions and transcriptions. Additional details on pseudolabelers and modalities are provided in Appendix D.

3.2 Tokenization

We use discrete tokenizers that transform each modality input into a discrete tokens. This allows us to train a unified model without introducing modality-specific architectural changes, since all modalities are in discrete token space. Below we discuss tokenization details for different modality types.

Modalities with a dense 2D structure. We use video tokenizers that perform compression both spatially and temporally. For modalities with a 2D spatial structure such as RGB, depth, normals, and optical flow, we use the VidTok tokenizer [44]. This results in 1280 (flattened from a $5 \times 16 \times 16$ token tensor) tokens per 4-second 128 resolution video clip sampled at 4 FPS. For feature maps, we modify the input and output projection layers of VidTok’s architecture to be compatible with the feature map input dimensions, and replace the L2 based reconstruction objective with a cosine similarity loss. For feature-map modalities, we train tokenizers from scratch; for others, we fine-tune the pretrained RGB VidTok tokenizer.

Modalities that are represented sequentially. For modalities such as captions, transcriptions, and object class names within bounding boxes, we treat them as text tokens and encode using WordPiece tokenization [14]. Bounding box coordinates and human pose parameters are encoded using dedicated coordinate-bin tokens, with human poses further tokenized using a learned VQ-VAE. For bounding boxes and human poses, we introduce special frame markers [FRAME_1], [FRAME_2], etc., 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 qualitatively show the pseudo-labeling and tokenizer reconstruction results in the Appendix E. In total, the tokenized pretraining data amounts to 34B tokens, which are used for pretraining A2A-Video. We denote this dataset as K600-MM.

3.3 Developing A2A-Video

As shown in Fig. 3, A2A-Video is an encoder-decoder transformer model and uses a multimodal masked modeling objective [36] for pretraining. During pretraining, one subset is randomly selected from all modality tokens (across both space and time) as input to the encoder and another random subset as masked (or target) tokens which is predicted by the decoder. For dense modalities, tokens are sampled randomly for input and target, while for sequential modalities, we perform span masking [40]. The model is trained with a cross-entropy loss to predict the output tokens.

Adaptations for temporal modalities. We apply the following adaptations to the image-based 4M’s architecture [36, 17] for video modeling.

(i) *Spatio-temporal positional embeddings.* For modalities with explicit spatial and temporal structure (RGB, depth, optical flow, etc.), we use factorized additive learnable spatio-temporal positional embeddings. For sequential modalities such as bounding boxes and human poses, which do not have a spatial grid structure, we use learnable 1D positional embeddings and frame marker tokens (e.g., [FRAME_1]) to define the frame boundaries.

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

(iii) *Masking modes for future prediction.* 4M employs random and span masking [40] for non-sequential and sequential modalities, respectively. Random masking selects a subset of tokens to mask or keep visible from a given modality that simulates a diverse set of video tasks such as inpainting, outpainting, and past or future prediction etc. We introduce two additional masking strategies for non-sequential and one for sequential modalities to reinforce A2A-Video future prediction capabilities. In contrast to random masking where the model can interpolate a masked token from context on both sides in time, frame-conditioned masking forces it to predict from past context alone.

- **Frame-Conditioned Future Frame (FCFF):** The input tokens are sampled causally from the initial frames, and target tokens are taken exclusively from subsequent future frames, also in a causal manner. This enforces a temporal separation between input and output tokens.
- **Frame-Conditioned Any-Future (FCAF):** Input tokens are sampled causally from the 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):** Instead of random span masking, input tokens are selected causally starting from the first token, and masked tokens begin immediately after the last input token. This mirrors FCFF but is designed for sequential modalities. In our ablations, we found that incorporating CFP is essential for our inference strategy of feeding back predicted tokens to encoder, as it simulates such cases during training. Without this masking mode, model generations significantly deteriorate when predicted tokens are fed to the encoder.

During training, one of the three masking modes (random, FCFF, FCAF) is selected with equal probability for non-sequential modalities. For sequential modalities, CFP is sampled alongside the span masking at a probability of 35%. Incorporating these frame-level causal masking modes improves performance on future prediction tasks, as shown later in our ablation studies.

3.4 Chained generations

Unlike specialist video models that learn a fixed one-to-one mapping, A2A-Video can traverse across modalities and generate any desired target modality from one or multiple input modalities.

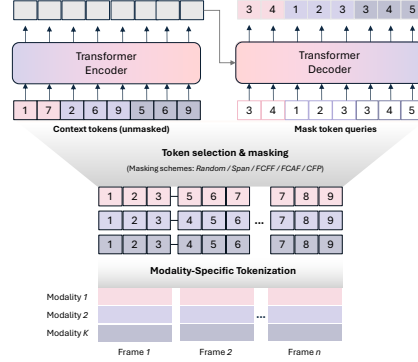


Figure 3: A2A-Video **framework**. Context tokens (*i.e.*, input to the transformer encoder) and masked tokens (*i.e.*, target tokens to be predicted by the transformer decoder) are sampled from different modality tokens. The masking strategy is a combination of random masking and future prediction masking modes tailored for both sequential and non-sequential modalities.

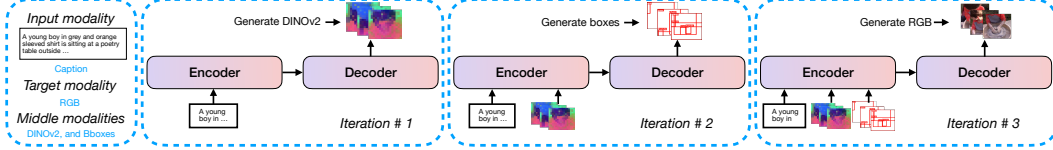


Figure 4: **Chained generations with A2A-Video.** We show the chaining capability of A2A-Video for the task of generating RGB video given caption. Using DINOv2 [38] and bounding boxes as the middle modality, A2A-Video predicts one modality at a time sequentially. Though each modality is visualized in raw form here, A2A-Video operates in shared discrete token space of all modalities.

This higher-order flexibility gives rise to chained generations,¹ 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.

Chained generation enables consistent predictions across modalities, e.g., depth that is in agreement with predicted semantics, and allows multimodal conditioning at test time without requiring annotated intermediate modalities that often require external models. Crucially, chaining is not only useful for this kind of cross-modal consistency but also serves as a methodological way to improve a single desired function. Using chaining, we can decompose a fixed one-to-one task into a self-predicted chain of intermediate, simpler sub-tasks. Below, we discuss one such instantiation for the RGB generation task.

Chaining in abstract spaces for RGB video generation. Video generation can be viewed as consisting of 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. We hypothesize that disentangling these two sub-tasks into a sequential process can instead lead to improved generation quality. We study this, and use chained generation in A2A-Video to decompose class-, caption-, and frame-conditional (1 and 5 frames) video generation into sequence of intermediate modalities: semantic modalities such as detections and feature maps, followed by geometric modalities such as 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. 2). To reduce the search space, we select a subset of A2A-Video’s modalities and categorize them into three groups: sparse (captions, transcriptions), semantic (SigLIP 2 [46], 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 abstract 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 through modality spaces

Video models can forecast rollouts into the future given past frames. This *future prediction* ability has sparked interest in their capacity as world models [5, 1]. While useful future prediction models with applications in robotics and driving are already emerging, existing models roll out the future in one or two spaces [53, 58, 28, 39, 27]. A2A-Video allows exploration of the effect of extending the prediction axis to many spaces (modalities) and allows the space to be chosen based on the information of interest. For example, A2A-Video 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. The latter is both less efficient and redundant since it requires modeling details (e.g., background pixels) in RGB that are unnecessary for motion prediction. For this investigation, we study forecasting in the RGB, depth, normal, bounding box, transcription, DINOv2, and SigLIP 2 spaces.

¹Chaining is broadly supported in any-to-any models, as demonstrated in 4M and 4M-21 [36, 17] for the image domain.

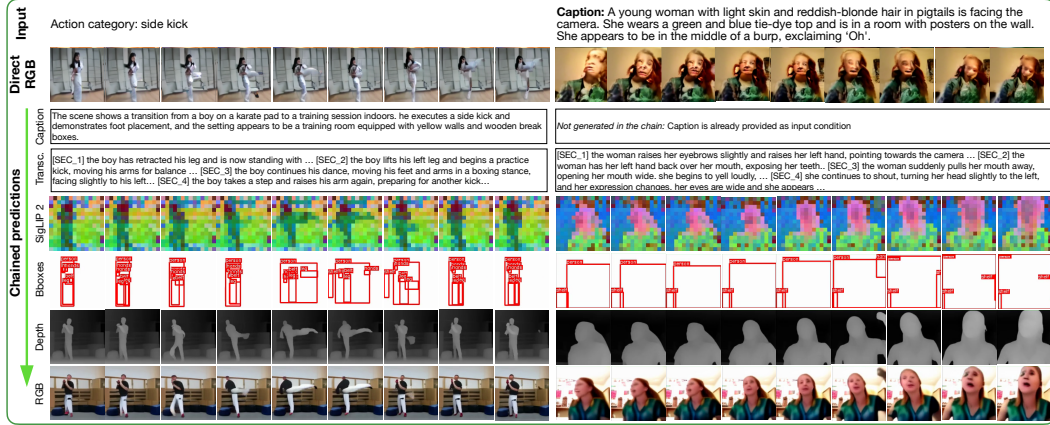


Figure 6: **RGB video generations with chained generations in abstract spaces.** For different RGB tasks (*left*: class-to-video, and *right*: caption-to-video). Generating intermediate abstract modalities such as transcriptions, SigLIP 2 feature maps, bounding boxes, and depth as middle modalities leads to improved RGB generations. In contrast, frames generated using direct RGB often show artifacts, especially for sudden motion changes, and lack geometric consistency.

modality from a single input with high consistency between predictions across modalities. In the following sections, we show the usefulness of A2A-Video’s video-specific implications, i.e., chaining and multimodal forecasting, for video generation and future prediction tasks respectively.

4.2 Video generation via chaining in abstract spaces

Using A2A-Video, we show that RGB video generations can benefit from chaining using the coarse-to-fine chain design discussed in Sec. 3.4. As our main baseline, we use A2A-Video to generate RGB tokens conditioned only on the input, denoted A2A-Video-direct. Both direct and chained generations use the same trained model. We also train a variant of A2A-Video using RGB, caption, and class modalities, with the same pretraining schedule as A2A-Video. This serves as our RGB-only baseline, and allows us to disentangle the effect of training exposure from the benefit of chaining.

We show the results in Tab. 1. Across different model sizes, A2A-Video with chaining in abstract spaces outperforms the direct prediction baseline. The trend is consistent across class-to-RGB, caption-to-RGB, and frame-conditional RGB generation tasks. For direct baselines, we optimize the generation hyperparameters, including CFG scale, using grid search.

Qualitative results. We show generation results in Fig. 6 for different tasks. 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. We hypothesize that this leads to RGB generations with coherent motion and fine-grained details. Additional qualitative results are shown in Appendix A.3.

Analysis on chain design. We study two questions regarding chain design: which abstract space group contributes most, and how the ordering of groups affects final generation quality. In Tab. 2, we perform generation using individual modality groups and

Table 1: **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
A2A-Video-B Direct	325.0	53.3	418.6	17.9	230.0	138.0
A2A-Video-B Chained	263.7	52.9	250.0	19.2	155.0	119.0
A2A-Video-L Direct	360.3	49.5	231.0	20.6	229.3	132.3
A2A-Video-L Chained	249.0	57.1	191.3	20.2	143.0	105.0

Table 2: **Which modality group matters most in chaining?**

Configuration	Video Generation		Video Prediction	
	Class→RGB	Caption→RGB	Frame conditional	
	FVD ↓	FVD ↓	1-frame ↓	5-frame ↓
Sparse (Caption → Transc. → RGB)	431.0	367.3	198.3	122.3
Semantic (SigLIP 2 → BBox → RGB)	266.3	191.7	150.7	107.7
Geometric (Depth → RGB)	439.7	309.3	209.0	122.0
Coarse-to-fine chain (default)	249.0	191.3	144.0	102.0

compare it with the full chain. Semantic modalities contribute most to generation quality, compared to the geometric and sparse groups. This suggests that conditioning on high-level semantics (e.g., via SigLIP 2) leads to better RGB generations than conditioning on depth or other low-level modalities that primarily capture structural detail.

In Tab. 3, we investigate the effect of chain order on the quality of the generation. We reverse the chain order such that geometric modalities are predicted first, followed by sparse and semantic modalities. This reverse ordering performs worse than the semantic-first chain order. From qualitative comparisons, we observe that when the geometric modalities are generated first, subsequent modality predictions are constrained by the geometric modality. This strictly limits the ability of semantic modalities to flexibly guide the generation trajectory (except for details not captured by geometric modalities, such as color information), even when semantic modalities are part of the chain. Therefore, not only the presence of semantics but also their order in the chain design significantly affects generation quality. Refer to Appendix A.3 for additional analysis.

Table 3: **Order of modality chain matters.**

Chain ordering	FVD ↓	Text Align. ↑
Caption → Geometric → Semantic → Sparse	239.0	19.9
Caption → Sparse → Semantic → Geometric (default)	191.6	20.2

4.3 Forecasting in abstract modalities

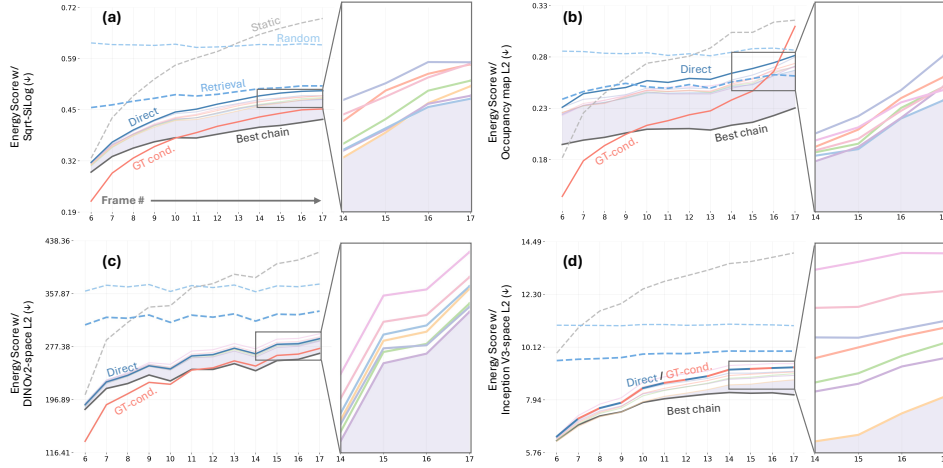


Figure 7: **Forecasting future (a) Depth, (b) Bounding Boxes, (c) DINOv2, and (d) RGB frames given 5 starting frames of RGB.** **Direct:** prediction of the future frames of the target modality directly from the 5 RGB frames, e.g. RGB[5] → Depth[5:17]. Chained predictions (e.g. RGB[5] → DINOv2[5:17] → Depth[5:17]) with RGB, optical flow, SigLIP 2, Bbox, DINOv2, transcription, depth, normal as middle modalities. **GT-cond:** prediction conditioned on the 5 ground-truth frames of the target modality, e.g. Depth[5] → Depth[5:17]. **Static** baseline: last input frame is used as forecast. **Random** baseline: 16 random ground-truth test samples are used as forecasts. **Retrieval** baseline: 16 closest neighbours selected from the ground-truth test set based on the conditioning frames, based on the modality-specific distance function, are used as forecasts. **Best chain:** result if the best modality chain for each sample was used. **Gap:** Gap between per-sample best chain result and the best single chain result.

Fig. 7 shows the results of forecasting evaluated with energy scores [19]. We first note the expected findings: (1) for most forecasts, the energy scores increase with longer prediction horizons, since predictions of the farther future are less likely to be close to the observed rollout; (2) 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, while the 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 – with model prediction-based forecasts getting closer to the retrieval baseline over time; and (3) the ground truth-conditioned prediction performs well, especially in the earlier frames, because past ground truth frames allow for closer predictions to the ground truth future frames.

One finding is that chained predictions generally outperform direct predictions for forecasting, reinforcing the results in Sec. 4.2. On average, the performance gap between different modality chains is small, especially relative to the energy score gap between the lower-bound baselines and the GT-conditioned prediction. However, the best per-sample chain performs significantly better than any single fixed chain — it even overtakes the GT-conditioned prediction in later frames and degrades much more slowly over time. This shows that while different modality chains perform similarly on average, the optimal chain is highly sample-dependent. It strengthens our hypothesis that choosing the right abstraction for prediction, on a per-sample basis, can substantially improve forecast quality. This is a capability directly enabled by an any-to-any video model like A2A-Video.

We note that the best-per-sample chain here is selected via an oracle, i.e., knowledge of the optimal forecast modality, and is therefore not directly usable at inference time. As a step towards an oracle-free alternative, we show in Appendix B.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 oracle access at test time. Finding the best chain per sample, however, is itself a non-trivial problem. In this work, we limit our contribution to demonstrating the potential of optimizing over modality chains and leave the chain-optimization problem itself to future work.

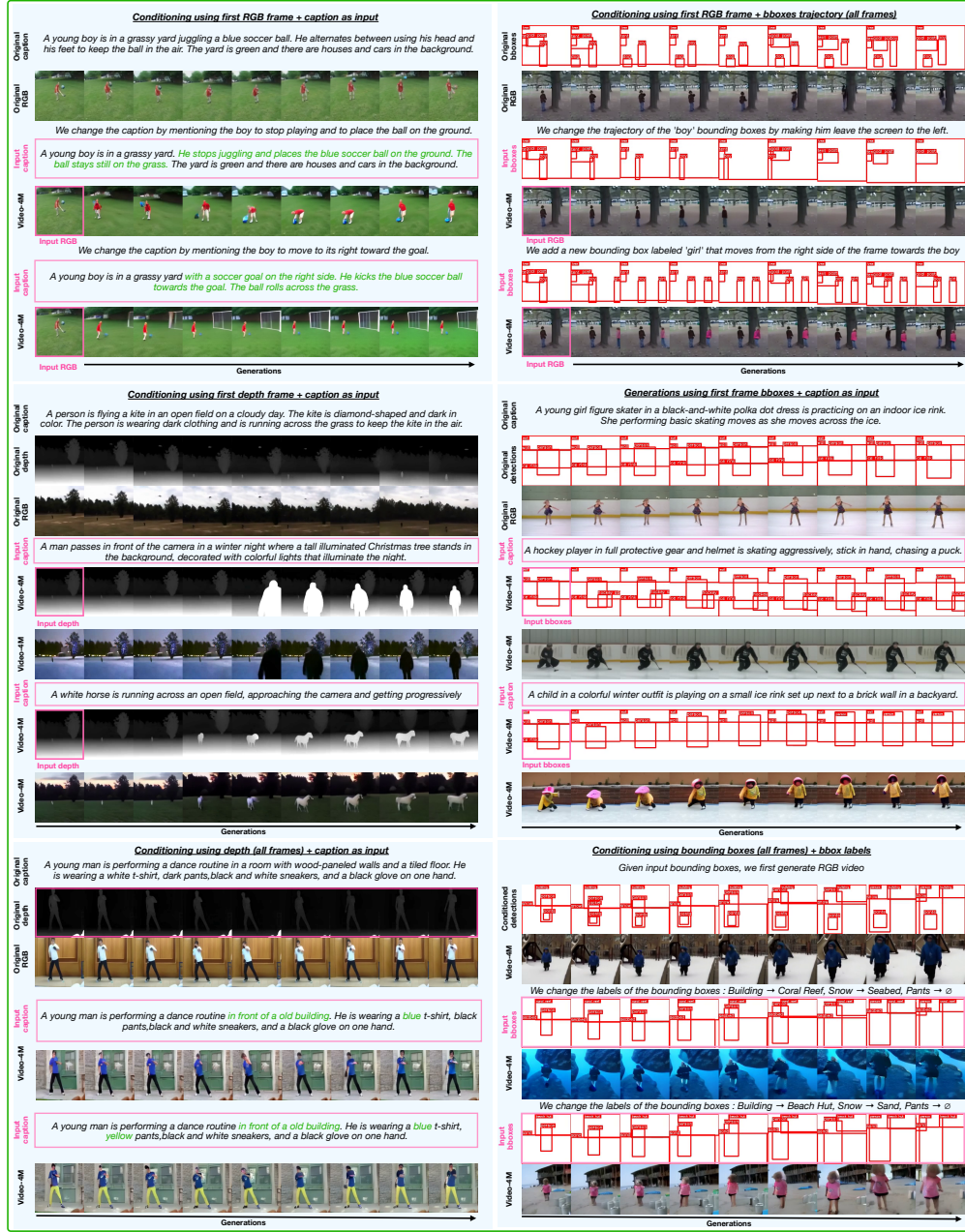


Figure 8: **Fine-grained and controllable video generation using multimodal input conditioning.** We showcase several instantiations of video generation through multimodal input probing. A2A-Video allows any combination of modalities as input to generate output videos. In most scenarios, we provide one 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).

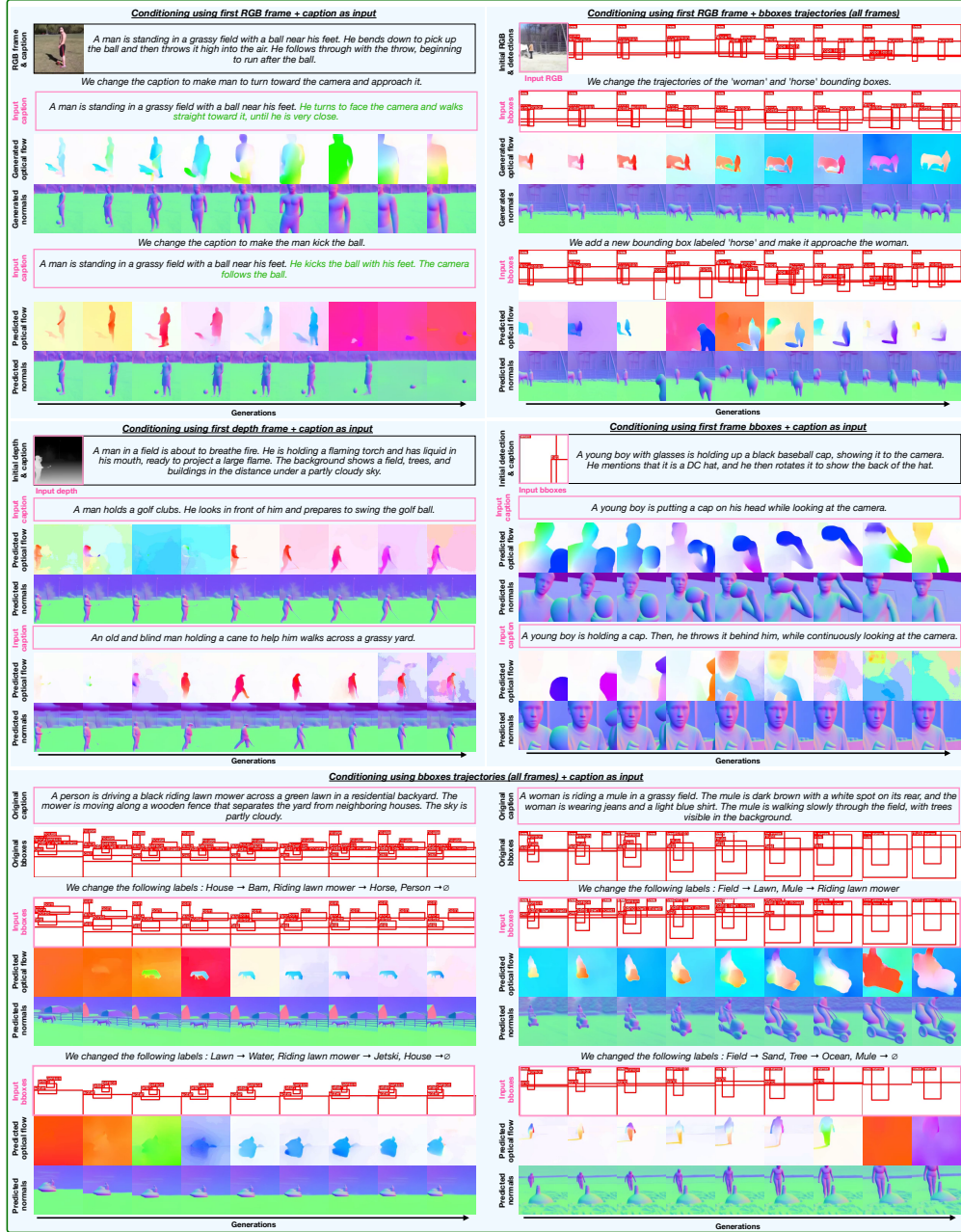


Figure 9: **Multimodal input probing for controllable generations in non-RGB modality spaces.** We extend our probing analysis to non-RGB modality spaces as outputs. For instance, we can control how surface normals in the future will look by varying the probing modality, such as captions, the first RGB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

4.4 Arbitrary modality-time conditioning for video editing and control

Different modalities capture complementary spatio-temporal information in videos. Using A2A-Video, this enables a diverse set of controls for video generation, ranging from coarse (e.g., caption) to fine-grained (e.g., depth or bounding boxes). We demonstrate a subset of these capabilities in Fig. 8. Given the first RGB frame, we generate its future frames by conditioning using caption as a coarse steering modality. By changing the caption (e.g., stopping the football juggling or kicking the soccer ball), we can generate plausible future frames that all start from the same input image. For finer-grained control, instead of an RGB frame as input, we can generate videos by providing the first frame in other modalities such as depth or bounding boxes. A2A-Video treats all of these tasks as instances of the same problem: completing the missing parts of a partially observed modality-time grid.

Most video models generate in RGB space; A2A-Video, in contrast, can generate and predict observations directly in modality spaces of interest, while still supporting multimodal input conditioning. As shown in Fig. 9, we use different input modalities to generate in the output space of other modalities such as optical flow and surface normals. For instance, using the first RGB frame, we can generate future surface normals and optical flow by probing with the caption modality. These are analogous to the RGB-output examples above, except the output space is non-RGB. This decoupling of input and output modalities highlights A2A-Video’s generality beyond RGB-centric video generation.

5 Conclusion

We develop A2A-Video, an any-to-any multimodal video model that supports a diverse set of abstract spaces, moving beyond the conventional RGB-centric video modeling paradigm. These abstract spaces are defined through different modalities, which provide complementary information about the underlying temporal observation. A2A-Video traverses both the modality and time axes simultaneously, giving rise to higher-order inference paths which are not supported by specialist video models. We study these implications in three settings. First, for generation, decomposing a downstream task (e.g., text-to-video) through intermediate representations improves the target task over direct generation. Forecasting benefits similarly: different modality spaces allow for unique rollouts, such that the sample-dependent optimal chain enables significantly higher-quality forecasts compared to forecasting directly in a single modality space. Finally, for control, we demonstrate that A2A-Video supports flexible input-output conditioning across arbitrary modalities and time points, from coarse to fine granularity, for video editing and generation.

References

- [1] 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.
- [2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. [arXiv preprint arXiv:2506.09985](#), 2025.
- [3] 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.
- [4] 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.
- [5] 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, 2024.

- [6] 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), pages 88–91, 2019.
- [7] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. arXiv preprint arXiv:1808.01340, 2018.
- [8] 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, pages 6299–6308, 2017.
- [9] 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, pages 16845–16889, 2025.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Hsu-kuang Chiu, Ehsan Adeli, and Juan Carlos Niebles. Segmenting the future. IEEE Robotics and Automation Letters, 5(3):4202–4209, 2020.
- [14] 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.
- [15] 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, pages 1323–1333, 2024.
- [16] 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.
- [17] Bachmann et al. 4M-21: An any-to-any vision model for tens of tasks and modalities. Neurips, 2024.
- [18] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems, 35:35946–35958, 2022.
- [19] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.
- [20] Google. Gemini, 2023.
- [21] 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.
- [22] 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, 2022.
- [23] 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, 2022.

- [24] 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, pages 21807–21818, 2024.
- [25] 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.
- [26] Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Dino-foresight: Looking into the future with dino. arXiv preprint arXiv:2412.11673, 2024.
- [27] 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), pages 3793–3803, June 2025.
- [28] 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, 2025.
- [29] 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.
- [30] 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, pages 19948–19960, 2023.
- [31] Bin Lin, Yang Ye, Bin Zhu, Jiayi 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, pages 5971–5984, 2024.
- [32] 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, pages 38–55. Springer, 2024.
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.
- [34] 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, pages 26439–26455, 2024.
- [35] 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, pages 648–657, 2017.
- [36] 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.
- [37] 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, pages 17810–17820, 2025.
- [38] 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.

- [39] 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), pages 13787–13797, June 2025.
- [40] 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.
- [41] Shuhuai Ren, Shuming Ma, Xu Sun, and Furu Wei. Next block prediction: Video generation via semi-autoregressive modeling. arXiv preprint arXiv:2502.07737, 2025.
- [42] Noam Shazeer. GLU variants improve transformer. ArXiv, abs/2002.05202, 2020.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] Michael Tschanen, 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.
- [47] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.
- [48] 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, 2024.
- [49] 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, 2024.
- [50] 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.
- [51] Yi Wang, Yanan 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 multi-modal understanding and generation. In International Conference on Learning Representations, volume 2024, pages 42055–42079, 2024.
- [52] 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, 2024.
- [53] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114, 2023.
- [54] 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, pages 521–539. Springer, 2022.
- [55] 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.

- [56] 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.
- [57] Lymin 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, pages 3836–3847, 2023.
- [58] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2024.

Appendix

A Additional experimental results

In this section, we provide additional experiments related to A2A-Video pretraining ablations, chained generation, and comparisons with external specialist models.

A.1 Masking strategies for pretraining A2A-Video

For pretraining A2A-Video, 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 future prediction 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 A2A-Video Base and A2A-Video Large). Their effect appears in downstream evaluations.

Tab. 4 shows ablations of these design choices using the A2A-Video 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 4: **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 ↓
A2A-Video (default)	<u>310.0</u>	238.0
A2A-Video - future pred. modes (non-seq. modalities)	289.0	331.0
A2A-Video - future pred. modes (seq. modalities)	316.0	240.0

A.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 [48, 49, 41] for video generation and prediction use a different training setup (i.e., training on 17 frame videos that cover a 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 [10] and NormalCrafter [4] are trained with high-resolution videos and images, whereas A2A-Video operates on tokenized inputs of 128² 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 A2A-Video operates. For the video captioning task, we simply resize input videos to 128x128, followed by upsampling to 224x224 resolution videos for the specialist models.

Results are shown in Tab. 5. On video generation and prediction tasks, LARP [48] performs best in terms of FVD, while A2A-Video 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 future prediction tasks. We therefore train a separate

Table 5: **Comparison with specialist models on standard video tasks.** A2A-Video is a single model evaluated across all 7 task groups, while baselines are specialist models each covering a subset of tasks (**Tasks** column). Under comparable evaluation settings, A2A-Video achieves overall competitive performance while being a single model **Gray rows** show specialist depth/normal models results at their native resolution, without the downsampling and tokenization offset.

Method	Tasks	Video Generation			Future Prediction		Captioning		Depth				Normal	
		Cls→Vid		Cap→Vid	1 frame	5 frame	MSVD		DIODE		NYUv2		DIODE	
		FVD ↓	Acc. ↑	FVD ↓ Text Align. ↑	FVD ↓	FVD ↓	MET. ↑	BERTSc. ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	Mean° ↓	Mean° ↑
RGB A2A-Video baseline	4/7	490.6	40.9	304.3	19.2	318.6	172.6	—	—	—	—	—	—	—
LARP [48]	1/7	166.8	43.4	—	—	78.8	72.0	—	—	—	—	—	—	—
NBP [41]	2/7	897.1	13.7	—	—	255.9	141.6	—	—	—	—	—	—	—
OmniTokenizer [49]	2/7	583.3	13.3	—	—	392.9	292.7	—	—	—	—	—	—	—
Video-LLaVA [31]	1/7	—	—	—	—	—	—	0.158	0.874	—	—	—	—	—
AuroraCap [9]	1/7	—	—	—	—	—	—	0.180	0.874	—	—	—	—	—
Video Depth Anything [10]	1/7	—	—	—	—	—	—	—	—	—	—	—	—	—
Video Depth Anything [10] (native res.)	1/7	—	—	—	—	—	—	—	0.571	0.673	0.194	0.855	—	—
NormalCrafter [4]	1/7	—	—	—	—	—	—	—	0.564	0.681	0.192	0.859	25.16	21.01
NormalCrafter [4] (native res.)	1/7	—	—	—	—	—	—	—	—	—	—	—	16.83	42.50
A2A-Video (ours)	7/7	249.0	57.1	191.7	20.2	143.0	105.0	0.218	0.892	0.590	0.631	0.218	0.823	23.94
													32.84	21.56
														47.40

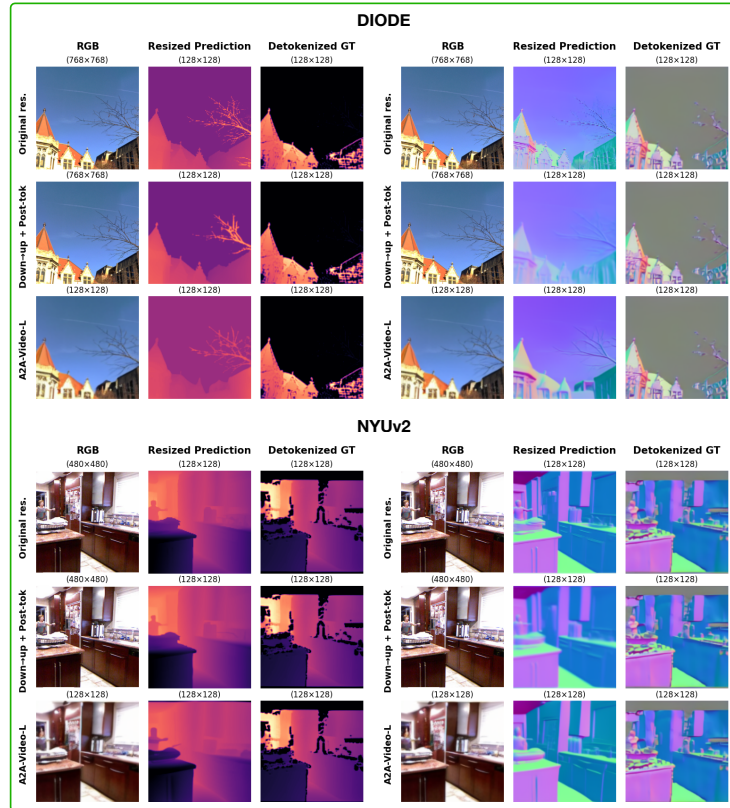


Figure 10: **Zero-shot qualitative examples for depth (left) and surface normal (right) estimation tasks.** For pseudo-labelers, we use the following configuration for comparable evaluation settings: *downsampling of input video to 128×128, resampling to the original resolution before prediction, and passing the prediction through the tokenizer with a reconstruction offset.* The tokenized GT represents the reconstruction upper bound after tokenization.

LARP model for both the 1-frame and 5-frame conditional benchmarks. In contrast, A2A-Video uses a single network that supports all tasks. On depth and normal estimation, A2A-Video performs competitively with external methods when evaluated under matched evaluation setups. Qualitative results are shown in Fig. 10. Overall, A2A-Video, 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.

A.3 Additional experiments on chained generations

Here we present additional ablation studies, analysis, and qualitative results for chained generations with A2A-Video.

Reversing the chain order. In Sec. 3.4, we showed that changing the order of modalities in the chain affects the generation results. Fig. 13 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. 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. 6, 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 7: Geometric modalities in chaining improves edge-based metrics.

Chain	FVD ↓	Top-1 Acc. ↑	RGB Edge Sharp. ↑	Depth Edge Sharp ↑
Sparse → Semantic → Geometric (default)	249.0	57.1	70.81	18.47
Sparse → Semantic	246.3	55.4	68.39	17.81

We next examine how the geometric modality in the chain design affects generation quality 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. 11 (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 [8] and is primarily tuned for action recognition, and is therefore potentially less sensitive to fine-grained spatial improvements such as object sharpness and boundary clarity.

In Fig. 11 (middle), we generate depth maps using A2A-Video 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. 7. 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. 11 (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 6: Coarse-to-fine chain design ablation.

Chain	FVD ↓	Top-1 Acc. ↑
Sparse only	439.0	30.1
Sparse → Semantic	246.3	55.4
Sparse → Semantic → Geometric	249.0	57.1

Table 8: Ablating modality groups in the chain design.

Chain	FVD ↓	Top-1 Acc. ↑
Sparse → Semantic → Geometric (default)	249.0	57.1
<i>Leave-one-out</i>		
Sparse → Geometric	433.0	35.4
Sparse → Semantic	246.3	55.4
Semantic → Geometric	273.9	62.9

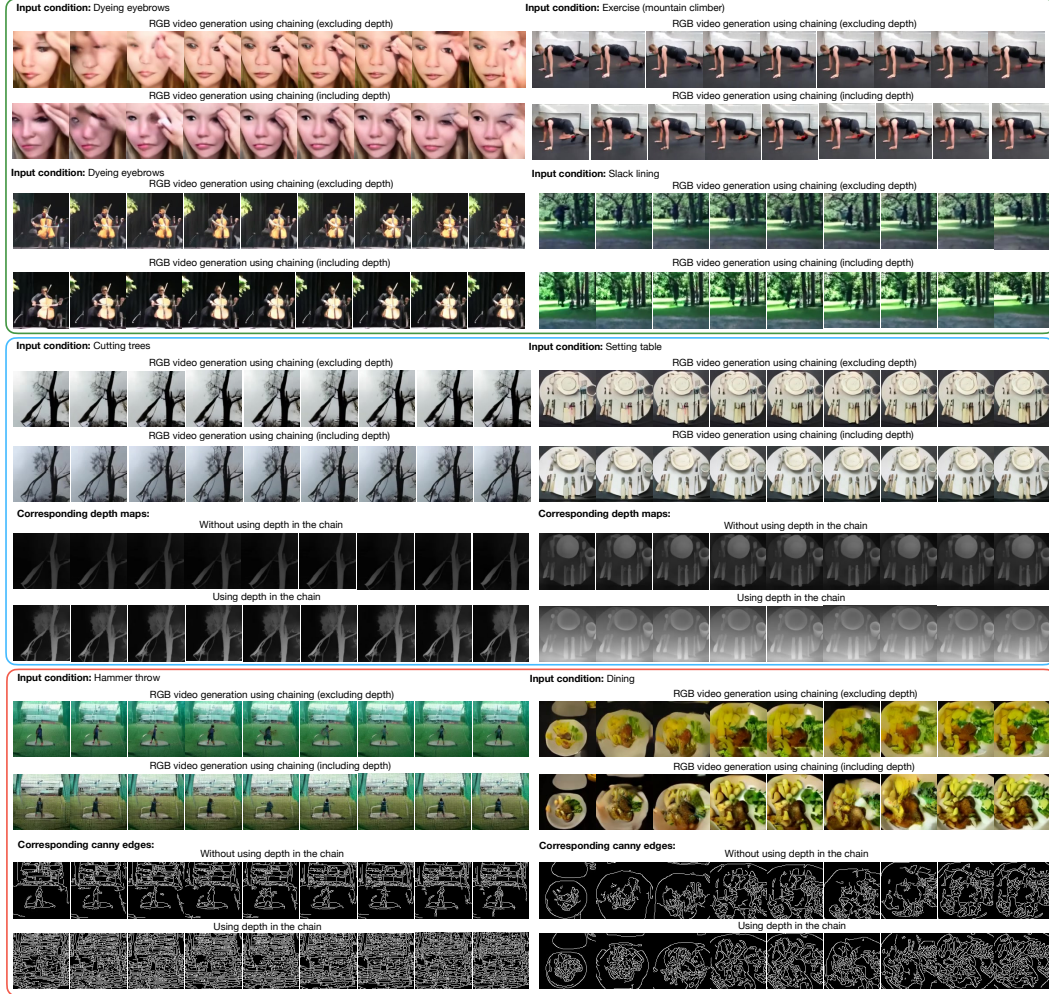


Figure 11: **Effect of including depth in the chain.** Top: RGB generations for different input conditions using the coarse-to-fine chain with and without depth in the chain, under identical hyperparameter and seed settings. Middle: depth maps corresponding to the two chain variants. Bottom: Canny edge maps of the generated videos. Including depth improves low-level structural details such as object boundaries, background clarity, and sharpness.

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 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. 12, we show results for 5-frame conditional generations for both

chained and direct generations using A2A-Video. Overall, chained generation leads to higher-quality video generations, with high consistency among predictions across different modalities in the chain.

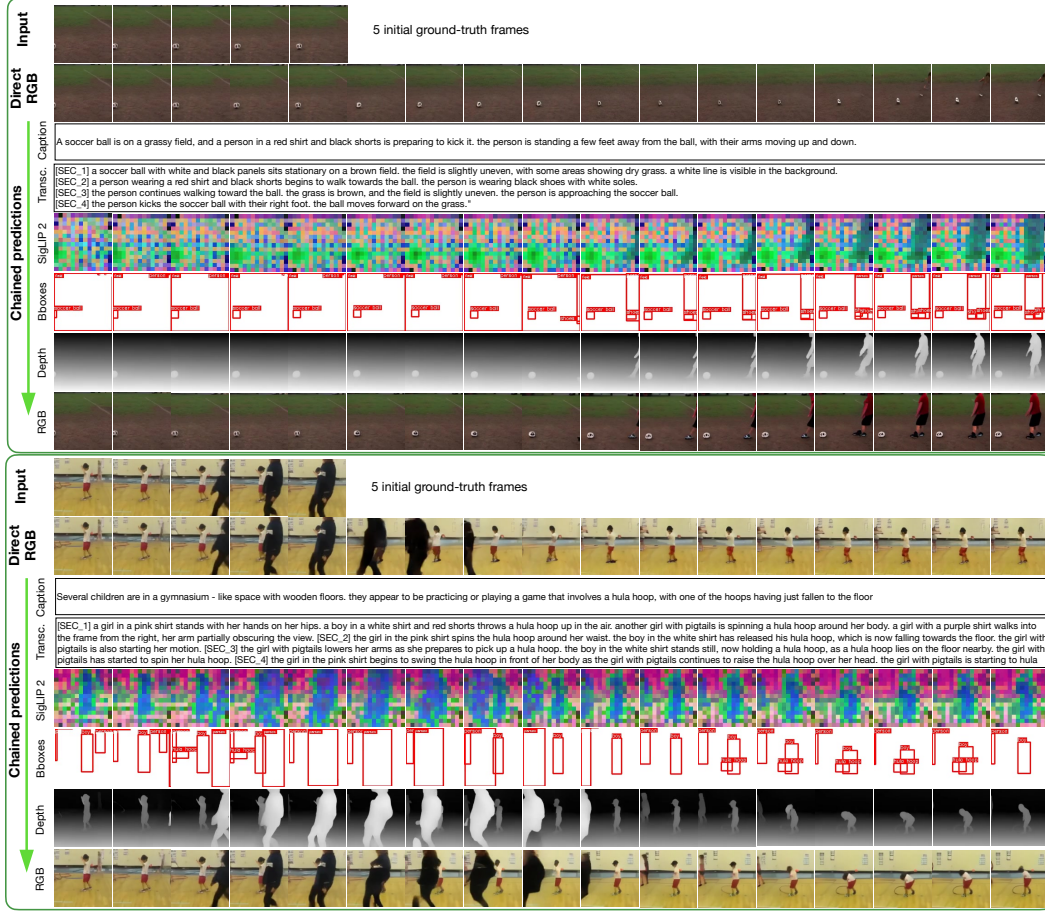


Figure 12: **Chained vs. direct predictions for 5-frame conditional RGB generation.** Given 5 initial ground-truth RGB frames as input, we compare direct RGB generation with chained generation, where Video-4M sequentially predicts caption, transcription, SigLIP 2 features, bounding boxes, and depth as 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.

B Additional results on forecasting in abstract modalities

B.1 Oracle-independent mixed modality 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. This distribution contains 10 bbox, 3 transcription, 2 SigLIP, and 1 surface-normal rollout, with the composition selected on a validation set disjoint from the test set. The resulting mixed-modality distribution achieves a better energy score than any single-modality forecast distribution. Note that this modality-mixture forecast is only enabled by a model such as A2A-Video. While this modality-mixture forecast does not rely on an optimal forecast modality oracle, the gain is weaker than using the oracle, so we believe that the main takeaway from the forecast experiments is still the per-sample nature of the best forecast modality, and this modality mixture forecast experiment

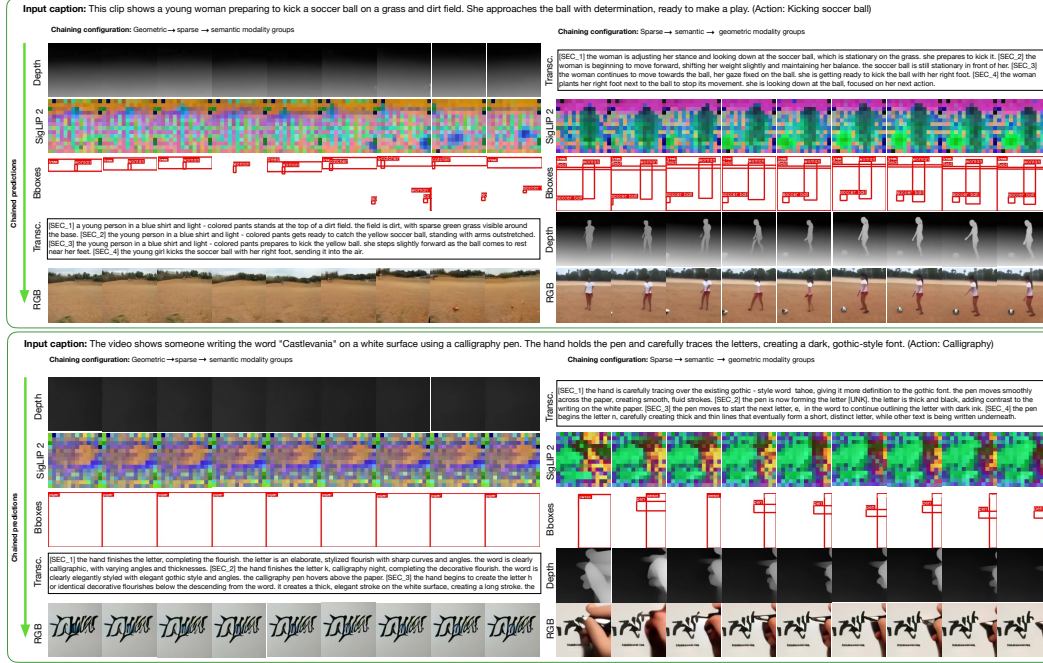


Figure 13: **Effect of reversing the chain order.** For the same input caption, we compare generations using the reversed chain (left: Geometric $\rightarrow$ Semantic $\rightarrow$ Sparse) and the default coarse-to-fine chain (right: Sparse $\rightarrow$ Semantic $\rightarrow$ Geometric). When the geometric 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.

serves as secondary strategy where combining forecasts from multiple modality spaces can partially recover the benefits of selecting the appropriate forecasting modality for each sample.

Table 9: **Oracle-independent mixed-modality forecasting.** Energy score (lower is better) for single-modality forecasts compared to the mixed-modality forecast, which combines rollouts across modalities without requiring oracle access at test time.

Forecast modality	Final frame (16)	Mean frames 6–16
Mixture	8.56 [8.36, 8.76]	7.73 [7.56, 7.91]
Bbox	8.67 [8.44, 8.89]	7.79 [7.61, 7.97]
SigLIP	8.94 [8.76, 9.12]	8.00 [7.84, 8.17]
Optical flow	8.91 [8.72, 9.10]	8.04 [7.87, 8.20]
DINOv2	9.04 [8.86, 9.22]	8.10 [7.93, 8.26]
Transcription	9.08 [8.90, 9.26]	8.20 [8.04, 8.37]
Normal	9.23 [9.06, 9.40]	8.27 [8.10, 8.43]
RGB	9.28 [9.09, 9.46]	8.30 [8.13, 8.47]
Depth	9.42 [9.25, 9.59]	8.45 [8.29, 8.61]

B.2 Energy scores

Gneiting & Raftery [19] 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 future prediction 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.

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

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

where d is a task-appropriate distance between rollouts. For the energy score to define a proper scoring rule, the distance metric d must be of negative type [19]. 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.

B.3 Qualitative results

Figure 14 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 A2A-Video.

C Additional results on multimodal conditioning for video editing and control

In Sec. 4.4, we showed different use cases of A2A-Video 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. 15 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.

D Pseudolabeling

No existing video dataset provides aligned ground truth across multiple modalities at the scale on which our pre-training 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 [7], we obtain additional modalities through pseudolabeling using task-specific models. Tab. 10 provides an overview of all the modalities supported by A2A-Video. 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 [46] and DINOv2

Table 10: **Overview of Modalities.** We group different modalities into four categories based on the type of information they capture. In total, A2A-Video supports 12 multimodal abstractions. For class category, we learn a token embedding layer.

Modality	Pseudo Labeler	Tokenizer
RGB	N/A	ST FSQ
<i>Geometric modalities</i>		
Depth	VDA-v2 [10]	FSQ
Surface Normal	NormalCrafter [4]	FSQ
3D Human Pose	TokenHMR [15]	VQ-VAE
<i>Motion modality</i>		
Optical Flow	DPFlow [37]	FSQ
<i>Semantic modalities</i>		
SigLIP 2	SigLIP 2 [46]	FSQ
DINOv2	DINOv2 [38]	FSQ
V-JEPA 2	V-JEPA 2 [2]	FSQ
Detection	Grounding DINO [32]	WordPiece
<i>1D/High-Level Modalities</i>		
Caption	Gemini-2.0 Flash [45]	WordPiece
Class category	N/A	Embed.
Transcriptions	Gemini-2.0 Flash [45]	WordPiece

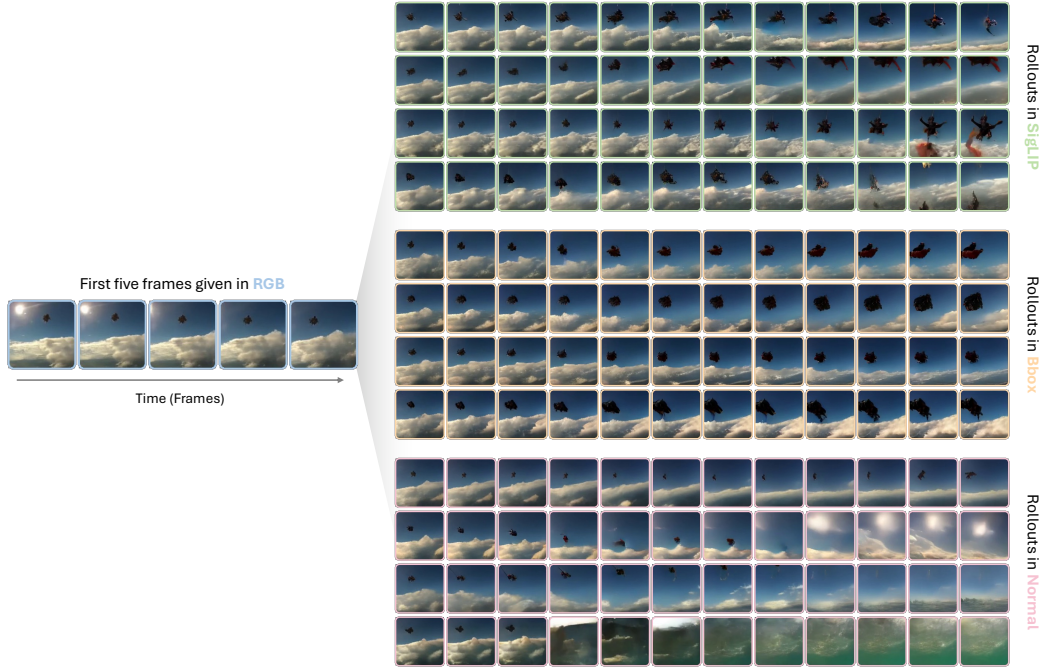


Figure 14: **Rollouts of Future Frames (Forecasting) in Different Modalities.** Future frames are rolled out in **SigLIP 2**, **Bbox**, and **Normal** modality spaces, from the same five initial RGB frames, using 4 different seeds each. Forecasts in **SigLIP 2** tend to preserve the semantics of the 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. 7 that the best forecasting modality depends on the sample and the task at hand.

[38], and 256 for VJEPa 2 [2]), 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 [20] 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 A2A-Video.

For human poses, we first generate 3D human poses using TokenHMR [15]. 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 [25] 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.

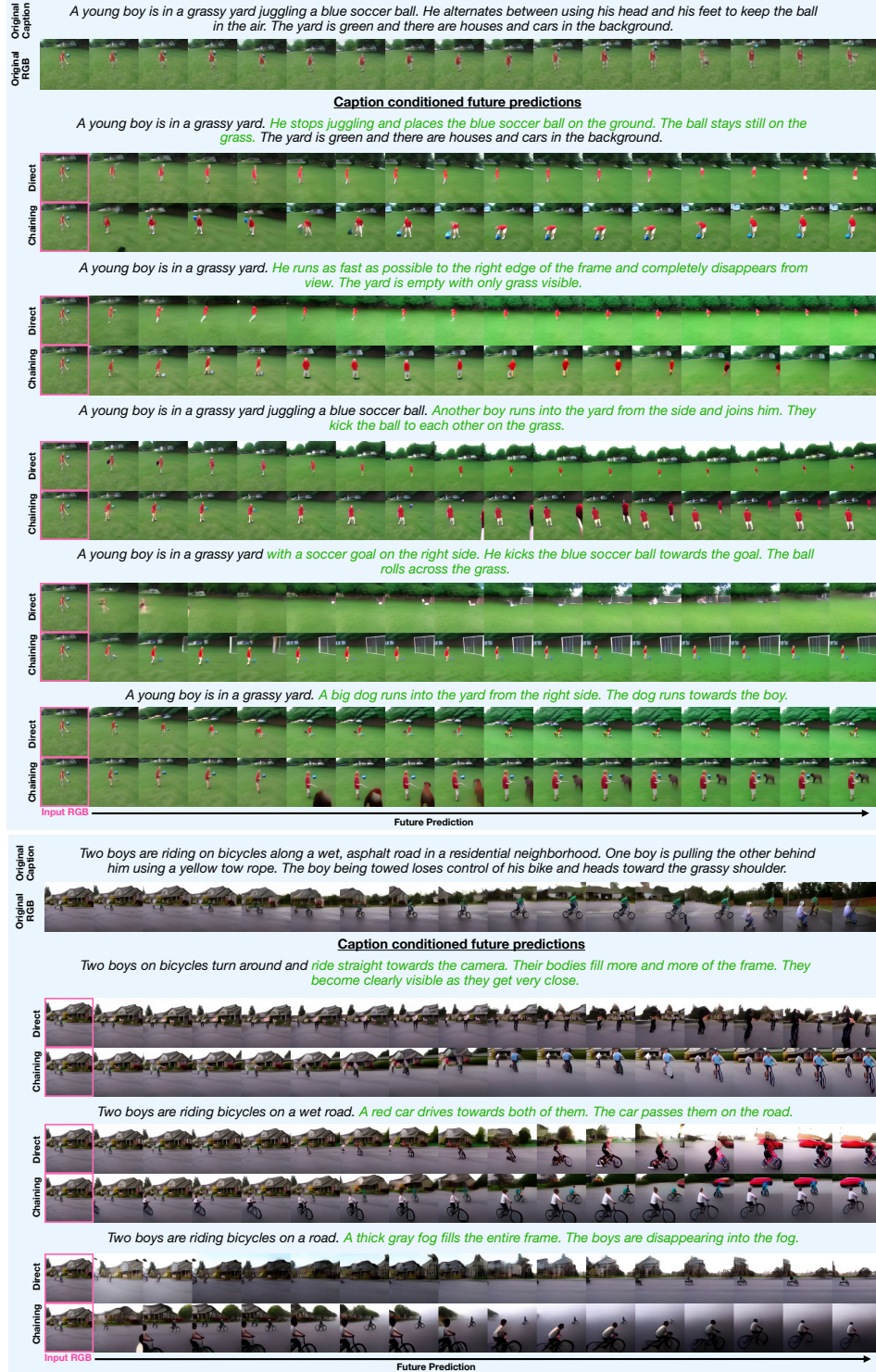


Figure 15: **Caption-conditioned with direct and chaining future predictions.** Given a single RGB frame and a caption text as inputs, we use A2A-Video to predict different future RGB trajectories by changing the caption. Notice how the chaining improves the future prediction.

E Tokenization

A2A-Video is pretrained on the tokenized K600-MM dataset. As discussed in Sec. 3.2, we train modality-specific discrete tokenizers. In Fig. 16, 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 more abstract 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.

E.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 [11] and UniTAB [54], as well as image-based any-to-any models [36, 17].

Human poses. For human poses, we use a learned VQ-VAE tokenizer from 4M-21 [17], which tokenizes each pose into 8 tokens, as detailed in Tab. 10. 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].

F Experimental setup details

F.1 Evaluation details.

For RGB generation tasks, we use Fréchet Video Distance (FVD) [47] 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 [24], we use a UMT-L [30] model finetuned on Kinetics-600 classification to measure class-video alignment (using the first 16 of 17 frames), and ViCLIP-InternVid-10M-FLT [51] 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.

F.2 Pre-training details

Implementation details. A2A-Video is trained at three model scales: A2A-Video-B with 198M parameters, A2A-Video-L with 705M parameters, and A2A-Video-XL with 2.8B parameters. For A2A-Video-B, we use a batch size of 11,200 (70 per GPU × 40 GPUs × 4 accumulation steps) and train for 1,500B tokens, taking approximately 1 day on H200 GPUs. For A2A-Video-L, we

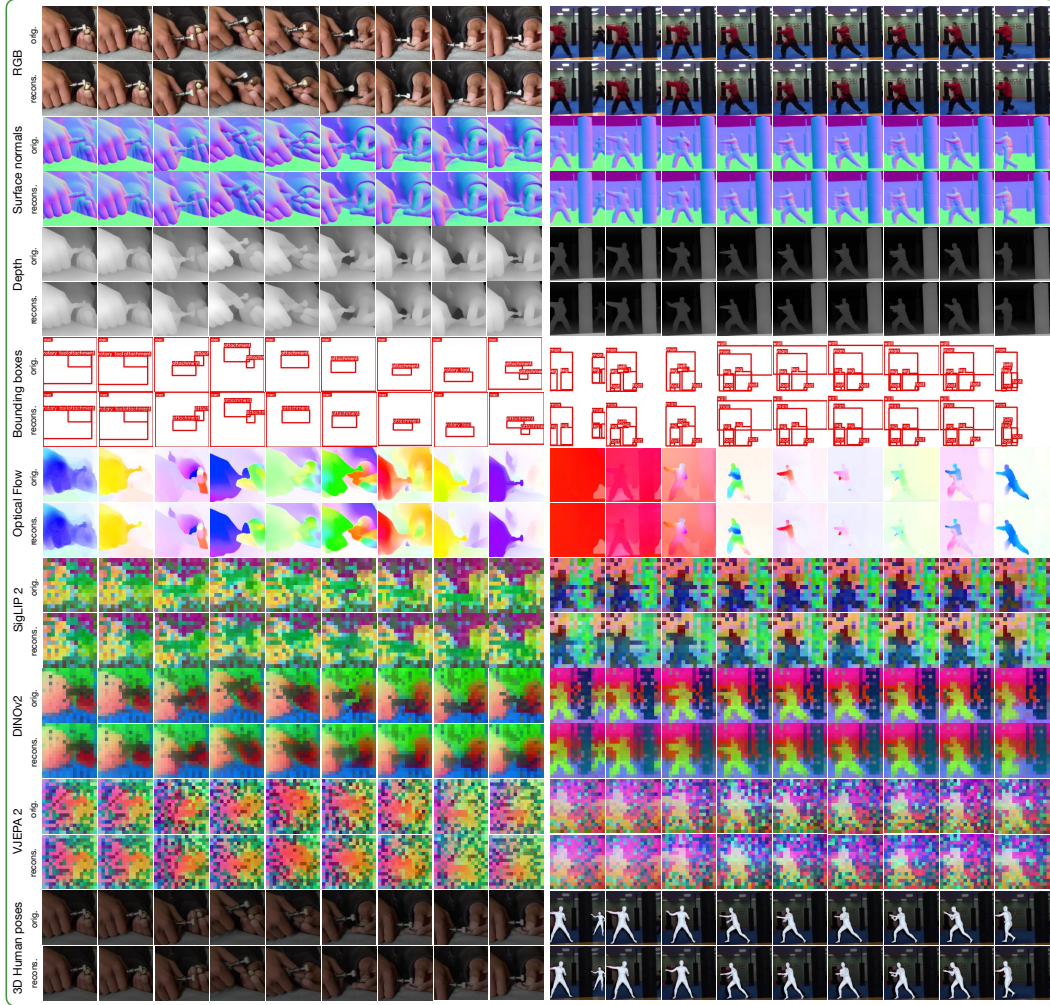


Figure 16: **Tokenizer reconstruction quality across modalities.** For each of the tokenized modalities, we show the original pseudolabeled input (orig.) alongside its reconstruction after tokenization (recons.), for two representative K600-MM clips: one with gradual motion (a person using a sewing machine, left) and one with abrupt motion changes (a person performing a kick, right). Reconstructions of abstract modalities such as feature maps, depth, and optical flow are highly consistent with the raw modalities, while RGB reconstructions contain small artifacts when fast or abrupt motion occurs.

use a batch size of 11,520 ($36 \text{ per GPU} \times 80 \text{ GPUs} \times 4 \text{ accumulation steps}$) and train for 1,500B tokens over approximately 2 days. For A2A-Video-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 A2A-Video-B and Fully Sharded Data Parallel (FSDP) for A2A-Video-L and A2A-Video-XL. We pretrain all variants of A2A-Video from scratch. Pretraining hyperparameters are shown in Tab. 11.

G Limitations

While A2A-Video demonstrates the benefits of abstract space modeling for video generation and prediction, 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,

Configuration	Base	Large	XL
Training length (n tokens)	1500B	1500B	1500B
Warmup length (n tokens)		10B	
Optimizer		AdamW [33]	
Opt. momentum		$\beta_1, \beta_2 = 0.9, 0.95$	
Base learning rate [21]		$2e-5$	
Batch size	16.3K	14.3K	6.4K
Weight decay		0.05	
Learning rate schedule		Cosine decay	
Feedforward activation		SwiGLU [42]	
Input token budget	512	512	512
Target token budget	512	512	512
Input and target α		Mixture [17]	
Masking strategy		Mixture [17]	
Image resolution		128^2	
Augmentation		Center Crop	
Repeated sampling [18]		4	
Data type		bfloat16 [6]	

Table 11: **Pre-training settings for scratch initialization.**

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 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 D, E), but the quality of abstract spaces 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 A2A-Video 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 future prediction.** Evaluating video generation and future prediction 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 A.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 future prediction metrics assume a single ground-truth future, whereas multiple plausible futures may exist for a given observation. We believe developing better evaluation protocols for both video generation and future prediction is an important direction for future work.

H Broader Impact

As demonstrated in our paper, a unified A2A-Video model 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. A2A-Video 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) abstract 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.