BRIDGING PERCEPTUAL AND ANALYTIC DYNAMICS VIA FUNCTION ALIGNMENT

Yuxuan Wu[♭] **Gus Xia**[♭♯]
[♭] Mohamed bin Zayed University of Artificial Intelligence
[♯] New York University Shanghai

ABSTRACT

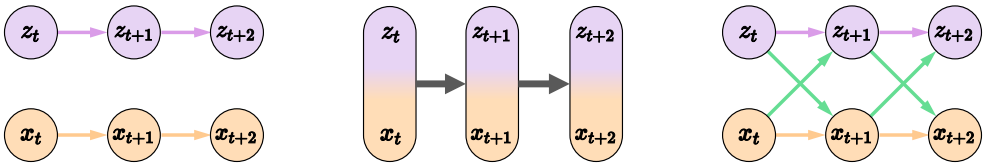
Human modeling of complex processes often involves multiple representations that capture different aspects of the same underlying reality. While recent approaches mostly unify such representations into a single predictive model, this unification could obscure the distinct functional roles associated with each representation. Inspired by the *function alignment* framework proposed recently, we study an alternative paradigm in which heterogeneous predictive dynamics are preserved and coupled through bidirectional alignment at the level of functions. We consider a setting with two representations of the same process paired in time: a high-dimensional ***perceptual*** sequence and a compact ***analytic*** state sequence, each governed by its own autoregressive dynamics. Rather than collapsing them into a unified model, we align their predictive functions using lightweight adapter modules that allow each dynamics to incorporate signals from the other during rollout. We conduct experiments on two physical prediction tasks exhibiting different functional roles of the two dynamic processes, and demonstrate that function alignment significantly improves long-horizon stability during joint rollout in both perceptual and analytic domains [1]. Together, our results provide a concrete instantiation of function alignment between perceptual and analytic dynamics, along with empirical evidence that preserving heterogeneous predictive dynamics can be critical for stable sequential prediction.

1 INTRODUCTION

Humans reason about the world through multiple levels of abstraction. The same underlying reality can be perceived, described, and predicted using different representations, ranging from low-level sensory patterns to high-level analytic structures (Fodor, 1975). These representations typically exhibit distinct predictive dynamics and serve different functional roles. For example, perceptual representations tend to capture rich and high-dimensional information to model short-range context dependencies in observations, whereas more abstract or analytic representations often trade perceptual detail for compact structure, focusing on longer-range temporal dependencies and more general regularities (Kahneman, 2011; LeCun, 2022; Bennis & Lahlou, 2025). These heterogeneous and often complementary functions of intelligence, sometimes referred to as *mode-1* and *mode-2*, coexist and interact in human cognition—reflecting the interplay between **sensation** and **reasoning** that enables flexible and robust understanding and generation across diverse scenarios.

In contrast, most contemporary artificial intelligence systems adopt a unified modeling paradigm for predictive dynamics. Transformer-based architectures have become a dominant framework in multimodal generative modeling across domains such as video, audio and text. To integrate these distinct sources of information, many approaches rely on fusion or alignment mechanisms at the *representational* level, where inputs from different modalities or abstraction levels are embedded into compatible forms of representation and processed by a single autoregressive backbone (Liu et al., 2023; Wang et al., 2023; Lu et al., 2024; Zhou et al., 2025a). Despite the empirical success on conditional generation tasks, such unified way of modeling introduces an unavoidable structural trade-off: either the dynamics of one modality are discarded and treated as static conditioning, or the dynamics of multiple modalities are absorbed into a single predictive process, *collapsing* their

[1] Project demo page: `https://irislucent.github.io/PA-Function-Alignment-Demo/`



(a) Standalone dynamics (b) Unified dynamics (c) Function alignment

Figure 1: Organization of predictive dynamics across two representations ($\boldsymbol{x}$ and $\boldsymbol{z}$) of the same underlying reality. (a) Standalone dynamics evolve independently. (b) Flat unified modeling collapses $\boldsymbol{x}$ and $\boldsymbol{z}$ into a single autoregressive state transition. (c) Function alignment preserves individual dynamics while enabling bidirectional cross-function adaptation (green arrows).

distinct inductive biases. In both cases, the resulting models do not fully exploit the complementary functional roles of different representations, limiting sample efficiency in long-horizon modeling and hindering extension to fully unconditional co-generation settings.

Inspired by the function alignment framework (Xia, 2025), we propose that aligning heterogeneous representations at the function level—while preserving their individual dynamics—can improve long-horizon prediction of both. We consider the setting where two temporally paired representations of the same underlying process follow distinct autoregressive dynamics. Specifically, we study the interaction between a ***perceptual*** dynamic process implemented by an autoregressive neural network operating directly on high-dimensional perceptual observations, and an ***analytic*** dynamic process defined over a compact state space that captures underlying physical regularities—two predictive functions mirroring the duality of *mode-1* and *mode-2* in human cognition.

As shown in Figure 1c, by introducing bidirectional adaptation while keeping each autoregressive function intact, the two dynamics are allowed to co-evolve over time, exchanging corrective cues while retaining their respective functional strengths. We evaluate this hypothesis on two physical prediction tasks: (1) a wind-affected bouncing-ball system, where the analytic process only models the ideal kinematics without wind effects, and (2) a double-pendulum system, where the analytic process relies on an approximate transition model. These settings represent two distinct functional trade-offs in the analytic dynamics: (1) missing latent information, and (2) suffering from model inaccuracy. In both cases, we find that function alignment consistently improves predictive performance on both perceptual and analytic representations over baselines that unify multiple dynamics within a single model, with particularly pronounced gains in long-horizon rollouts where error accumulation dominates. To sum up, our contributions are threefold:

- We present a concrete bidirectional instantiation of function alignment for predictive dynamics, bridging analytic and perceptual representations while preserving their individual autoregressive functions.
- We empirically demonstrate that function alignment outperforms unified-dynamics baselines and ablated variants, with consistent improvements in long-horizon prediction where error accumulation is critical.
- We show that function alignment is effective across distinct trade-offs in the analytic dynamic process, including settings where the analytic process either lacks latent information or structurally relies on approximate models.

2 Related Work

A large body of work in multimodal machine learning studies how to align heterogeneous modalities through shared or coordinated representations. Earlier approaches often employ multi-stream architectures, where different modalities are processed by separate encoding branches and interact through cross-modal attention mechanisms, enabling fine-grained exchange of information between modalities like vision and natural language (Tan & Bansal, 2019; Tsai et al., 2019; Lu et al., 2019; Zheng et al., 2022; Copet et al., 2023). Another line of work adopts unified or single-branch modeling strategies, where signals from different modalities are mapped into a common form of representation through discretization or projection, and then modeled by a single backbone (Wang et al.,

2023; Lu et al., 2024; Lei et al., 2025; Zhou et al., 2025a;b). This paradigm has been largely motivated by the success of large-scale contrastive pretraining (Radford et al., 2021; Jia et al., 2021; Wu et al., 2023), which learns modality-invariant embedding spaces that facilitate joint reasoning and generation across modalities (Kim et al., 2021; Liu et al., 2023; Li et al., 2023; Zhang et al., 2023). Despite their effectiveness, these approaches primarily focus on aligning the representations across modalities, while largely underutilizing the respective dynamics inherent to each modality, which have been shown to carry rich semantic structure (Eslami & de Melo, 2025; Abbasi et al., 2025; Takishita et al., 2025).

Beyond aligning heterogeneous inputs through neural representations, a different form of alignment has long appeared in sequential prediction systems that combine multiple models operating over the same underlying process. In such settings, distinct components are often responsible for modeling complementary aspects of the data—for example, local evidence from observations and higher-level regularities over symbol sequences—and their predictions are combined at inference time. Classical speech recognition systems exemplify this pattern by integrating acoustic models with language models through probabilistic decoding (Hinton et al., 2012; Hannun et al., 2014; Povey et al., 2016), while related designs appear in statistical machine translation and state estimation, where learned observation models are coupled with structured sequence priors via post-hoc inference (Bahdanau, 2014; Gulcehre et al., 2015). These approaches demonstrate the practical value of combining heterogeneous predictive dynamics, but the interaction between components is typically fixed and unidirectional, occurring only at the decoding or inference stage rather than through learned, time-coupled adaptation between predictive functions.

The recently proposed function alignment framework (Xia, 2025) offers a perspective that bridges representation-level alignment and modular system-wise composition by allowing multiple autoregressive models to interact during the prediction process. Several recent works explore related forms of cross-model interaction. For example, Bansal et al. (2024) inject latent information from one pretrained model into another via cross-attention to share model-exclusive knowledge, while Jiang et al. (2025) aligns two identical models to unify the processing of conditioning and target signals in conditional generation. These approaches demonstrate the benefits of cross-model information flow, but are limited to unidirectional adaptation and shared model architectures or modalities. Zayats et al. (2024) investigates mutual interaction between pretrained components through cross-attention layers, yet remains focused on conditional generation within the transformer architecture. Our work lifts function alignment to a more general setting, enabling bidirectional adaptation across different modalities, between heterogeneous autoregressive dynamics, and supporting fully unconditional joint generation.

3 METHODOLOGY

We consider two temporally paired sequences $\boldsymbol{x}_{1:T}$ and $\boldsymbol{z}_{1:T}$, where $\boldsymbol{x}$ denotes the raw perceptual signals and $\boldsymbol{z}$ denotes a compact representation corresponding to the same underlying observations. (x_t, z_t) are aligned in time at each time step t. We assume limited to training data consisting of paired trajectories $\{(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})\}$.

Each representation is equipped with its own autoregressive predictive dynamics. The perceptual sequence $\boldsymbol{x}$ is modeled by a high-capacity autoregressive model:

$$p_\theta(x_t \mid x_{<t}) \equiv f_x(x_{<t}, \theta), \tag{1}$$

which operates directly on perceptual observations. In this work, we instantiate the perceptual model f_x as a GPT-style transformer decoder, which serves as a representative example for illustrating the proposed alignment mechanisms (Vaswani et al., 2017; Radford et al., 2018). The analytic sequence $\boldsymbol{z}$ follows a separate analytic dynamics:

$$\hat{z}_t \equiv f_z(z_{<t}), \tag{2}$$

defined over a compact state space. Throughout this work, both perceptual and analytic dynamics are treated as distinct autoregressive processes with independent state transitions, unless explicitly coupled through alignment mechanisms.

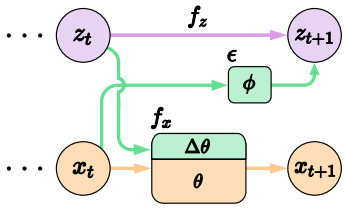


Figure 2: Bidirectional function alignment of perceptual dynamics f_x and analytic dynamics f_z. Green components denote the additional adaptation modules introduced for alignment.

3.1 Function Alignment Between Perceptual and Analytic Dynamics

Function alignment couples the perceptual and analytic predictive dynamics by enabling information exchange during autoregressive rollout. We introduce two directional alignment mechanisms corresponding to the two predictive processes.

Analytic-to-perceptual alignment ($z \to x$). To incorporate analytic state information into perceptual prediction, we augment the conditioning context of the perceptual dynamics. Specifically, the perceptual model is adapted as

$$p_{\theta+\Delta\theta}(x_t \mid x_{<t}, z_{<t}) \;\equiv\; f_x(x_{<t}, z_{<t}, \theta + \Delta\theta), \tag{3}$$

allowing perceptual predictions to depend on past analytic states in addition to perceptual history. In practice, this is implemented by first projecting $z_{<t}$ to the same representation space as $x_{<t}$, and concatenating them to the perceptual token embeddings and enforcing causal self-attention, such that x_t may attend to past analytic states. Note that while the perceptual model parameters are optimized to accommodate the additional analytic context, the original parameters θ that support standalone perceptual prediction are preserved.

Perceptual-to-analytic alignment ($x \to z$). To allow perceptual information to influence the analytic dynamics, we introduce an additional adapter network that projects perceptual cues into the analytic state space. The modified prediction on the state space is defined as

$$\tilde{z}_t \equiv f_z(z_{<t}) + \epsilon(x_{<t}, \phi), \tag{4}$$

where $\epsilon(\cdot)$ is implemented as a multi-layer perceptron (MLP) operating on a short history of $\boldsymbol{x}$.

As shown in Figure 2, the alignment mechanisms enable bidirectional information flow between the perceptual and analytic dynamics. Note that although we instantiate the alignment mechanisms with certain architectural choices (e.g., transformer decoder for f_x and MLP for ϵ), the framework is broadly applicable to any pair of autoregressive functions and the alignment modules can be flexibly designed accordingly.

3.2 Training Objectives and Optimization

The above function alignment introduces two sets of trainable parameters: the parameter update $\Delta\theta$ applied to the perceptual model f_x, and the parameters ϕ of the perceptual-to-analytic adapter $\epsilon(\cdot;\phi)$. These parameters are optimized using separate training objectives corresponding to the two alignment directions.

To train the analytic-to-perceptual alignment, we fine-tune the perceptual model with additional conditioning on $\boldsymbol{z}$. Given paired trajectories, we minimize the negative log-likelihood of perceptual observations:

$$\mathcal{L}_{z\to x}(\Delta\theta) \;=\; \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z})} \sum_{t=1}^{T} -\log p_{\theta+\Delta\theta}(x_t \mid x_{<t}, z_{<t}). \tag{5}$$

Similarly, the perceptual-to-analytic adapter is trained with a regression loss in the analytic space:

$$\mathcal{L}_{x\to z}(\phi) \;=\; \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z})} \sum_{t=1}^{T} \|z_t - \tilde{z}_t\|^2 . \tag{6}$$

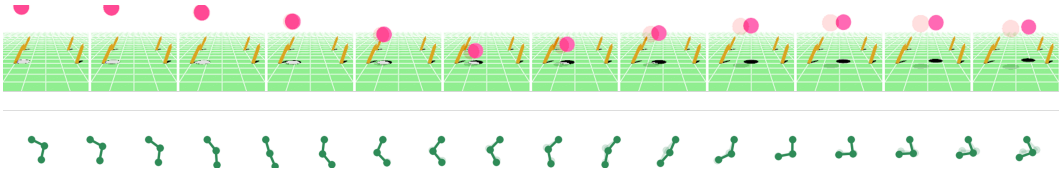


Figure 3: Example frames (from left to right) for the two experimental environments. Up: Wind-Affected Bouncing Ball (WBB), where the ball in dimmed pink renders the kinematic rollout without wind effects. Down: Double Pendulum (DP), where the pendulum in dimmed green shows the Euler-integrated simulation. Note that the dimmed renderings are only for visualization.

Note that throughout training, we apply teacher forcing on the alignments of both directions—the ground truth x_t and z_t instead of the model predictions $\hat{x}_t$ and $\hat{z}_t$ are used as supervision in Equations 5 and 6. This is appropriate because the alignment modules are trained to adapt to an already well-initialized dynamics model on the opposite side. Teacher forcing allows the alignment to be learned from stable optimization targets under the assumption that the opposite dynamics can provide sufficiently reliable context during rollout. Since the two alignment objectives involve disjoint parameter sets, we apply early stopping to each based on their respective validation losses.

To efficiently optimize analytic-to-perceptual alignment without disrupting the pretrained dynamics θ, we parameterize the update $\Delta\theta$ using low-rank adaptation (LoRA) Hu et al. (2022). Concretely, $\Delta\theta$ is realized as a low-rank decomposition added to selected weight matrices of θ, while the original parameters θ are kept frozen throughout training.

4 Experiments

In this section, we instantiate the function alignment framework with specific model architectures and evaluate its empirical behavior on two physical prediction environments, under both conditional and unconditional rollout settings on perceptual ($\boldsymbol{x}$) and analytic ($\boldsymbol{z}$) trajectories.

4.1 Experimental Settings and Implementation

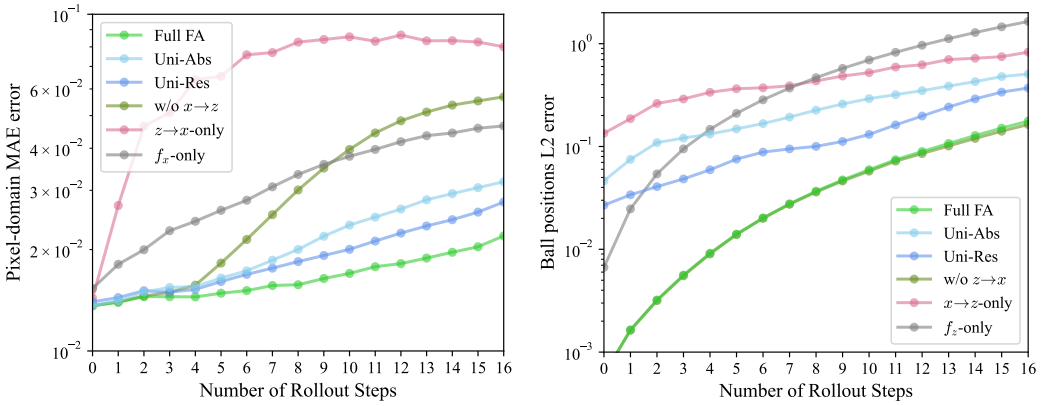
We consider two simulated physical systems rendered as videos, each paired with an analytic state sequence describing the same underlying process.

Wind-Affected Bouncing Ball (WBB). A ball moves in a 3D space under gravity, with approximately elastic collisions that exhibit mild energy dissipation, and a wind field with small Gaussian perturbations over time. The wind direction and magnitude are visually indicated by waving wheat-like textures but are not observed in the analytic state. The analytic representation $\boldsymbol{z}_t$ consists of the ball's position and velocity vectors, and the analytic dynamics f_z follow ideal kinematic equations without wind terms.

Double Pendulum (DP). A chaotic double-pendulum system is rendered as video frames. The analytic representation $\boldsymbol{z}_t$ consists of the two joint angles and angular velocities. f_z is implemented using Euler integration of the physical equations of motion, introducing accumulated numerical error over long horizons.

Example frames of these datasets are shown in Figure 3, where we also visualize the f_z rollouts using dimmed colors. Other details of the datasets can be found in Appendix B.

For both environments, we implement the perceptual dynamics f_x as a GPT-style transformer decoder with causal self-attention. We first train a ResNet-based VQ-VAE (He et al., 2016; Van Den Oord et al., 2017) to learn discrete token representations of video frames, and then train the transformer to autoregressively model the token sequences. We adopt the rotary positional encoding scheme in RoFormer (Su et al., 2024) for temporal position embedding and a trainable position embedding to distinguish patches in every frame. Accordingly, the analytic-to-perceptual alignment is implemented by concatenating the projected analytic tokens to the temporal-spatial visual patches (separated with an additional token), and fine-tuning the RoFormer layers with LoRA to attend to the



(a) Perceptual-domain prediction error.

(b) Analytic-domain prediction error.

Figure 4: Unconditional co-generation performance on the WBB environment given a prompt of 3 steps. Prediction error in both the perceptual and analytic domains is shown as a function of rollout horizon. Errors are plotted on a logarithmic scale.

newly added analytic context causally. The perceptual-to-analytic alignment is implemented as an MLP that takes the visual tokens of past three frames and outputs a corrective signal in the analytic state space. Details of the model architectures are provided in Appendix C.

4.2 Baselines and Ablations

All compared models in this section can be viewed as *structural variants* obtained by progressively removing, collapsing, or restricting components of full bidirectional function alignment.

Unified Dynamics Baseline. We consider a unified autoregressive transformer as a primary baseline, which concatenates perceptual and analytic tokens into a single sequence and models their co-evolution by fine-tuning the pretrained f_x. During training, both x_t and z_t can attend to the full history of both modalities, and $\boldsymbol{z}$ is predicted with an additional output head. The unified model has two variants: (i) *Uni-Abs*, which models directly the absolute analytic state z_t, and (ii) *Uni-Res*, which models the difference $z_t - f_z(z_{<t})$ to leverage the base analytic dynamics. Both variants have comparable numbers of trainable parameters to the full function alignment model.

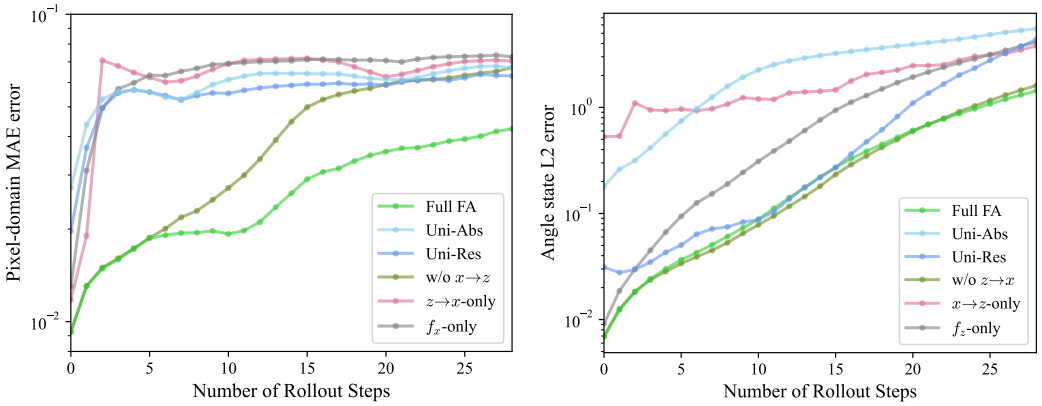
Ablation Variants. We evaluate a set of ablated variants that progressively degrade full bidirectional function alignment. Specifically, we consider: (i) removing a single adaptation pathway while retaining the other (*w/o* $x \to z$ or *w/o* $z \to x$); (ii) removing the base predictive dynamics and relying solely on cross-function adaptation ($x \to z$*-only* or $z \to x$*-only*); and (iii) standalone dynamics continuously trained without any alignment (f_x*-only* or f_z*-only*).

4.3 Quantitative Results

We evaluate all models under **unconditional co-generation**, where perceptual and analytic trajectories are jointly rolled out without access to any ground-truth signals beyond the initial paired prompts. This setting exposes long-horizon error accumulation and requires mutual correction between heterogeneous predictive dynamics.

Figure 4 reports unconditional co-generation performance on the WBB environment given a 3-step prompt. We measure the average prediction error in both representation spaces. In the perceptual domain, error is measured as pixel-domain mean absolute error (MAE) between predicted and ground-truth video frames. In the analytic domain, error is measured as the ℓ_2 distance between predicted and ground-truth ball positions.

Figure 4a shows perceptual prediction error as a function of rollout horizon. The full function alignment model (Full FA) consistently achieves the lowest pixel-domain error and exhibits the



(a) Perceptual-domain prediction error.

(b) Analytic-domain prediction error.

Figure 5: Unconditional co-generation performance on the DP environment given the initial state.

slowest error growth over time. In contrast, unified-dynamics baselines (Uni-Abs and Uni-Res) suffer from steady error accumulation as rollout length increases, despite having access to the entire history of both modalities during training. As for ablated variants, removing a single adaptation pathway (w/o $x \to z$) leads to delayed but inevitable divergence, while continuously training the perceptual dynamics (f_x-only) shows steady error growth.

Similarly, Figure 4b reports analytic state prediction error under joint rollout. Full FA consistently exhibits superior performance over other models. Standalone analytic dynamics (f_z-only) diverge rapidly due to unmodeled wind effects, and unified-dynamics baselines suffer from considerable error from the very beginning. Note that although removing one adaptation pathway (w/o $z \to x$) also shows competence here, it can solely rely on f_x-only in the perceptual domain rollout.

Taken together, we observe a **win-win** effect of two dynamics enabled by function alignment—accurate analytic cues improve perceptual prediction in the short term, which in turn provide more reliable perceptual feedback for future analytic updates, and vice versa. Breaking this feedback loop in either direction leads to asymmetric and ultimately unstable behavior—although w/o $x \to z$ appears competitive early in perceptual prediction (Figure 4a), analytic error accumulates and eventually degrades the perceptual dynamics. In contrast, such a win–win interaction is not observed in unified-dynamics baselines, which do not leverage the respective strengths of the dynamics of each modality, despite incorporating f_z explicitly in Uni-Res.

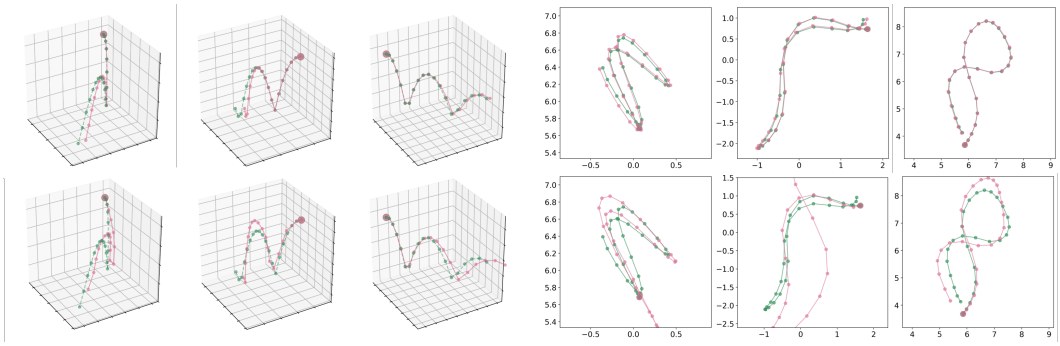
A similar pattern of results is observed in the DP environment, as shown in Figure 5. Full FA alignment consistently outperforms baselines and ablated variants in both perceptual and analytic domains. Notably, in this environment, the single-direction ablation (w/o $x \to z$) performs better than Uni-Abs and Uni-Res in the perceptual domain, suggesting that a sufficiently accurate initial analytic estimation is critical for modeling a chaotic system.

Results under conditional generation are provided in Appendix E.1. Conditional generation corresponds to in-distribution, teacher-forced inference on a single modality and only probes single-direction alignment. FA remains competitive and exhibits particularly strong performance on analytic prediction.

4.4 Qualitative Analysis

Figure 6 provides a qualitative comparison of long-horizon analytic rollouts between full function alignment (FA) and the unified-dynamics baseline (Uni-Res) under unconditional co-generation. Although both models are initialized from identical states, their subsequent trajectory evolution differs substantially over time.

In the WBB environment (Figure 6a), full function alignment closely tracks the ground-truth motion in three-dimensional position space over extended rollouts. The predicted trajectories preserve both



(a) WBB trajectories in the 3D position space.

(b) DP trajectories in two pendulums' angle space.

Figure 6: Representative long-horizon rollouts in the analytic state space under unconditional co-generation. Up: rollouts produced by Full FA (pink) compared against ground truth (green). Bottom: Uni-Res (pink) compared against ground truth (green). All trajectories are generated from identical initial conditions, marked by a larger point. In DP trajectories, the horizontal axis corresponds to the angle of the first pendulum, and the vertical axis corresponds to the angle of the second pendulum.

the overall structure and local variations of the motion. In contrast, while Uni-Res also exhibits trajectory turning consistent with the presence of wind, its predictions deviate increasingly from the ground truth, indicating accumulated error that cannot be corrected once perceptual and analytic dynamics are collapsed into a single transition.

A similar but more pronounced effect is observed in the DP environment (Figure 6b), where the underlying dynamics are chaotic and highly sensitive to initial conditions. Under full function alignment, predicted trajectories remain qualitatively consistent with the ground truth in the joint angle space over long-horizon rollout. Uni-Res, however, diverges rapidly and suffers from severe error accumulation, leading to qualitatively incorrect trajectory evolution away from the ground-truth angles. This behavior is further reflected in the joint error space of the two pendulum angles (Appendix Figure 10), where full function alignment remains the slowest in error growth compared to unified dynamics and FA variants.

5 Discussion

In this section, we discuss and interpret the empirical findings from a structural perspective, focusing on why bidirectional function alignment is effective and when unified-dynamics modeling may be insufficient.

5.1 Error Profiles of Latent Representations

To better understand why bidirectional function alignment could work, we analyze the sources of prediction error that may arise when modeling an observed process with latent representations.

- **Estimation error.** This error arises from training the predictive function from finite data. Due to limited training samples or optimization stochasticity, the learned parameters $\hat{\theta}$ may deviate from the optimal parameters θ^*.
- **Representational error.** This error occurs when encoding the observed sequence to a lossy latent representation sequence, filtering out information that is relevant for predicting the observed process.
- **Model bias.** The model class used to fit the predictive dynamics may not be expressive enough to capture the true underlying process, leading to persistent approximation error even with infinite data.

We claim that for two distinct representations $\boldsymbol{x}$ and $\boldsymbol{z}$ of the same underlying process, these three sources of error typically manifest in complementary ways. Different representations impose dif-

ferent structural constraints on predictive modeling, leading to distinct mixtures of estimation error, representational error, and model bias. For example, a high-dimensional perceptual representation $\boldsymbol{x}$ may have low representational error but suffer from high estimation error due to the complexity of modeling rich sensory patterns. Conversely, a compact analytic representation $\boldsymbol{z}$ may have low estimation error due to its training simplicity but incur higher representational error if it omits relevant details, or suffer from model bias if the model expressiveness is structurally bounded.

This complementarity suggests that predictive functions defined on different representations can, in principle, assist each other if allowed to interact at the level of dynamics. In our experiments, the perceptual dynamics in both environments mainly exhibit estimation error due to the complexity of modeling video sequences, while the analytic dynamics suffer from either representational error (WBB) or model bias (DP). We have demonstrated through the two cases that function alignment can leverage the complementary error profiles to enable mutual correction. While in this work we instantiated the alignment using perceptual and analytic dynamics as a concrete and illustrative example, the function alignment framework itself is broadly applicable to any pair of predictive functions with complementary error profiles.

5.2 The Structural Drawbacks of Unified Dynamics Modeling

Unified modeling of dynamics can be limited by the mismatch of inductive biases preferred by different representations. For example, in physical systems, analytic states are typically continuous-valued and governed by smooth dynamics, whereas transformer-based perceptual models rely on discrete tokenization and are not naturally suited for modeling continuous state transitions. Also, in chaotic settings, approximate analytic dynamics can provide accurate short-term predictions and strong initial conditions, an advantage that is often diluted when perceptual and analytic processes are collapsed into a single transition. More broadly, different representations are frequently modeled using heterogeneous architectures that encode substantial prior knowledge, whereas a unified model may not be able to merge the prior knowledge of both effectively. Function alignment avoids these issues by preserving the architectures of both representations.

5.3 Limitations

This work focuses on settings where two representations of the same underlying process are temporally aligned and paired at each time step. We do not address scenarios in which such pairing must be inferred, relaxed, or learned from unaligned sequences. Extending function alignment beyond strict time alignment remains an open challenge. Moreover, while both predictive dynamics considered in this work are capable of autonomous rollout, their interaction is purely predictive and passive. The aligned dynamics exchange corrective signals during rollout but do not actively influence the underlying process through actions or interventions. As a result, the current framework captures alignment in perception and prediction, but does not model how aligned representations might dynamically adjust behavior through control or decision-making.

6 Conclusion

In this work, we studied how heterogeneous representations of the same underlying process can be jointly modeled through function alignment, without collapsing their distinct predictive dynamics. We proposed a concrete and fully specified instantiation in which perceptual and analytic dynamics are preserved as separate autoregressive processes and coupled through bidirectional adaptation. Through experiments on two physical prediction problems, we demonstrated that function alignment consistently improves long-horizon stability under unconditional co-generation, and that such mutual benefits arise across settings where the analytic dynamics play different functional roles. More broadly, this work provides empirical evidence that preserving representation-specific predictive dynamics is beneficial for stable sequential prediction, and that respecting the functional form of each dynamics can be as important as improving representational capacity.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9308–9317, 2025.

Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. Llm augmented llms: Expanding capabilities through composition. In *The Twelfth International Conference on Learning Representations*, 2024.

Mehdi Bennis and Salem Lahlou. Semantic communication meets system 2 ml: How abstraction, compositionality and emergent languages shape intelligence. *arXiv preprint arXiv:2505.20964*, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720, 2023.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*, 2015.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Junyan Jiang, Daniel Chin, Liwei Lin, Xuanjie Liu, and Gus Xia. Versatile symbolic music-for-music modeling via function alignment. *arXiv preprint arXiv:2506.15548*, 2025.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *arXiv preprint arXiv:2504.10462*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 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.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pp. 2751–2755, 2016.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sho Takishita, Jay Gala, Abdelrahman Mohamed, Kentaro Inui, and Yova Kementchedjhieva. Llms can compensate for deficiencies in visual representations. *arXiv preprint arXiv:2506.05439*, 2025.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, pp. 6558, 2019.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Gus G Xia. Function alignment: A new theory of mind and intelligence, part i: Foundations. *arXiv preprint arXiv:2503.21106*, 2025.

Vicky Zayats, Peter Chen, Melissa Ferrari, and Dirk Padfield. Zipper: A multi-tower decoder architecture for fusing modalities. In *NeurIPS Efficient Natural Language and Speech Processing Workshop*, pp. 123–135. PMLR, 2024.

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.

Jiahao Zheng, Sen Zhang, Zilu Wang, Xiaoping Wang, and Zhigang Zeng. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE transactions on multimedia*, 25:2213–2225, 2022.

Bo Zhou, Liulei Li, Yujia Wang, Huafeng Liu, Yazhou Yao, and Wenguan Wang. Unialign: Scaling multimodal alignment within one unified model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29644–29655, 2025a.

Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *The Thirteenth International Conference on Learning Representations*, 2025b.

A Appendix

B Dataset Details

Both datasets used in this work consist of image sequences rendered at a spatial resolution of 128×128. For each environment, we generate large-scale simulated trajectories and split them into disjoint subsets for representation learning, alignment training, and evaluation.

Specifically, for each dataset we generate $100{,}000$ sequences for pretraining the perceptual models (VQ-VAE and RoFormer), and an additional $100{,}000$ sequences for training the function alignment modules. Evaluation is performed on held-out data consisting of $10{,}000$ validation sequences and $10{,}000$ test sequences. All splits are non-overlapping and sampled independently.

B.1 Wind-Affected Bouncing Ball (WBB)

The analytic dynamics for the Wind-Affected Bouncing Ball (WBB) environment are implemented using explicit Euler integration. The analytic state at each time step consists of the ball position $\mathbf{p}_t \in \mathbb{R}^3$ and velocity $\mathbf{v}_t \in \mathbb{R}^3$.

The simulation uses a fixed time step of $\Delta t = 0.12$ seconds. The ball has radius $r = 0.5$ and moves under gravity with acceleration $\mathbf{g} = (0, 0, -9.81)$. Ground contact occurs at $z = 0$, and collisions with the ground are modeled using a restitution coefficient of 0.85.

Wind is modeled as a stochastic external force evolving according to a Brownian motion in the horizontal plane. At each time step, the wind vector $\mathbf{w}_t \in \mathbb{R}^3$ is updated as

$$\mathbf{w}_{t+1}^{(x,y)} = \mathbf{w}_t^{(x,y)} + \boldsymbol{\epsilon}_t \sqrt{\Delta t}, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \sigma^2 I),$$

where $\sigma = 0.05$, and the vertical component of wind is fixed to zero. Wind effects are rendered visually but are not included in the analytic state.

The ball acceleration is computed as

$$\mathbf{a}_t = \frac{\mathbf{f}_{\text{gravity}} + \mathbf{f}_{\text{wind}}}{m},$$

where $\mathbf{f}_{\text{gravity}} = m\mathbf{g}$ and $\mathbf{f}_{\text{wind}} = \mathbf{w}_t$. Velocity and position are then updated using Euler integration:

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \mathbf{a}_t \Delta t, \qquad \mathbf{p}_{t+1} = \mathbf{p}_t + \mathbf{v}_{t+1} \Delta t.$$

B.2 DOUBLE PENDULUM (DP)

The analytic dynamics for the Double Pendulum (DP) environment are defined by the standard equations of motion for a planar double pendulum system. The analytic state at each time step is a four-dimensional vector consisting of the angles and angular velocities of the two pendulums.

Physical parameters are fixed across all trajectories. Both pendulums have unit length ($L_1 = L_2 = 1.0$) and unit mass ($M_1 = M_2 = 1.0$), and gravity is set to $g = 9.81$. Trajectories are simulated over a time horizon of $t_{\text{stop}} = 3$ seconds.

Ground-truth trajectories are generated using a high-accuracy numerical solver based on adaptive Runge–Kutta integration (RK45). Specifically, the system of ordinary differential equations is solved using `solve_ivp` with dense evaluation at fixed time steps.

The analytic dynamics used for prediction are implemented using explicit Euler integration. Formally, let $\mathbf{z}_t$ denote the analytic state consisting of the two pendulum angles and their angular velocities. Given the continuous-time dynamics $\dot{\mathbf{z}} = f(\mathbf{z}, t)$, explicit Euler integration updates the state as

$$\mathbf{z}_{t+1} = \mathbf{z}_t + f(\mathbf{z}_t, t)\,\Delta t. \tag{7}$$

Given an initial state, the system is evolved forward using a fixed step size of $\Delta t = 0.1$. To reduce numerical instability, Euler integration is performed with s substeps per time step. The integration step is divided into smaller increments of size $\Delta t/s$, where s denotes the number of substeps, and the resulting trajectory is downsampled to the original temporal resolution. We set s to 5 in our experiments.

Due to the chaotic nature of the double pendulum, small numerical errors introduced by Euler integration accumulate rapidly over time, leading to divergence from the ground-truth RK45 trajectories under long-horizon rollout.

C MODEL ARCHITECTURES

C.1 PERCEPTUAL MODEL: VQ-VAE AND ROFORMER

For both environments, perceptual observations are modeled using a two-stage architecture consisting of a VQ-VAE for discrete representation learning followed by an autoregressive transformer for temporal modeling.

VQ-VAE. Video frames are encoded using a ResNet-18 backbone as the encoder of the VQ-VAE. The encoder maps each frame to a grid of discrete latent codes drawn from a learned codebook. For the DP environment, the VQ-VAE uses a codebook size of 256 with latent dimensionality $d_{\text{code}} = 32$ and embedding dimension $d_{\text{emb}} = 512$. The decoder mirrors the encoder structure with 6 layers and 2 residual blocks per layer, each with 512 hidden channels. Unless otherwise specified, the same VQ-VAE architecture is used for WBB.

Perceptual Dynamics. The perceptual dynamics f_x are modeled using a GPT-style transformer decoder with causal self-attention and rotary temporal positional embeddings. For the WBB environment, the transformer uses a hidden dimension of $d_{\text{model}} = 768$, 12 layers, and 8 attention heads, with dropout set to 0.1 and a maximum sequence length of 512 tokens. For the DP environment, the transformer uses a hidden dimension of 512 with 6 layers. In both cases, the transformer operates on the discrete token sequences produced by the VQ-VAE and predicts future tokens autoregressively.

C.2 ALIGNMENT MODULES

Perceptual-to-analytic alignment ($x \to z$) is implemented using a lightweight multi-layer perceptron. Specifically, a three-layer MLP takes perceptual tokens from the recent history as input and outputs a corrective signal in the analytic state space.

Analytic-to-perceptual alignment ($z \to x$) is implemented by projecting analytic states into token embeddings and concatenating them with temporal-spatial image tokens, as shown in Figure 7.

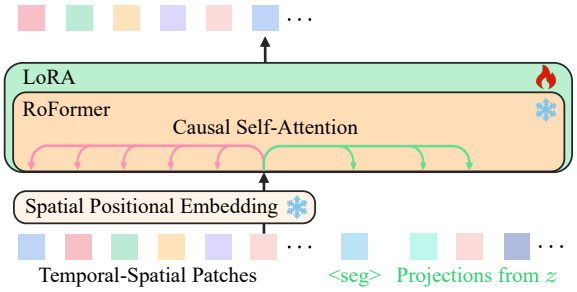


Figure 7: Adaptive training of f_x for $z \to x$ alignment in our experiments. The RoFormer f_x is fine-tuned with LoRA to attend to additional context tokens causally, as colored in green.

The perceptual transformer is adapted to attend to these additional tokens causally using low-rank adaptation (LoRA).

LoRA Configuration. For full function alignment and variants, LoRA adapters use rank $r = 16$ with scaling factor $\alpha = 64$. For unified-dynamics baselines, because of the absence of perceptual-to-analytic adaptation, larger LoRA adapters are used with rank $r = 24$ and scaling factor $\alpha = 96$, ensuring comparable adaptation capacity across models. In all cases, the original transformer backbone parameters are kept frozen during alignment training.

D Training Details

All models are trained on a single NVIDIA RTX 5000 Ada GPU using mixed-precision training with `bf16` precision. Optimization is performed using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay set to 0.01. The base learning rate is 1×10^{-4}, except when performing continuous training on f_x on the alignment dataset where we start from 1×10^{-5}. A cosine annealing learning rate scheduler is used, with a minimum learning rate factor of 0.01. The scheduler operates over $20{,}000$ steps, and the learning rate decay is further modulated by a decay factor of 0.99 applied every $1{,}000$ steps, with a minimum decay factor of 0.1.

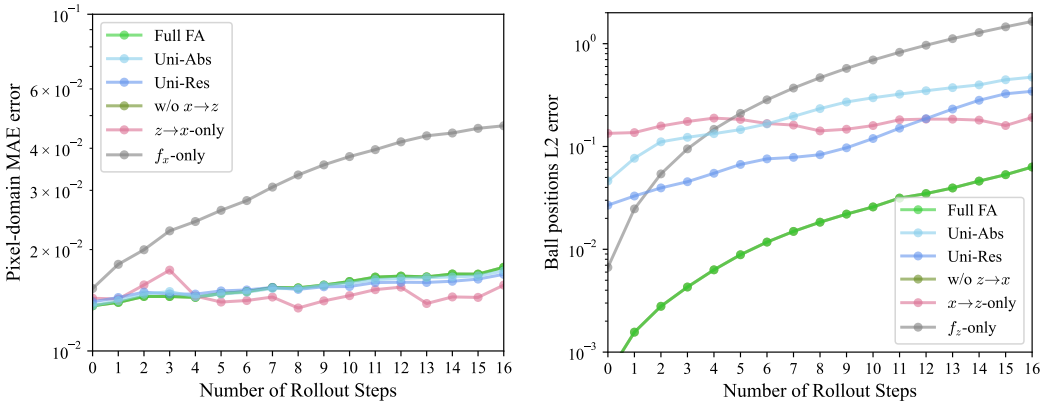
For unified-dynamics baselines and function alignment models, only the designated adapter parameters and LoRA modules are updated during alignment training, while the pretrained perceptual backbone remains frozen. Early stopping is applied based on validation loss for each alignment module independently.

E Additional Experimental Results

E.1 Conditional Generation Results

We additionally report results under conditional generation as a sanity check in Figure 8 and Figure 9. In this setting, one modality is generated autoregressively while the other modality is clamped to ground truth at each time step. As a result, conditional generation corresponds to in-distribution inference with teacher forcing and does not expose long-horizon error accumulation.

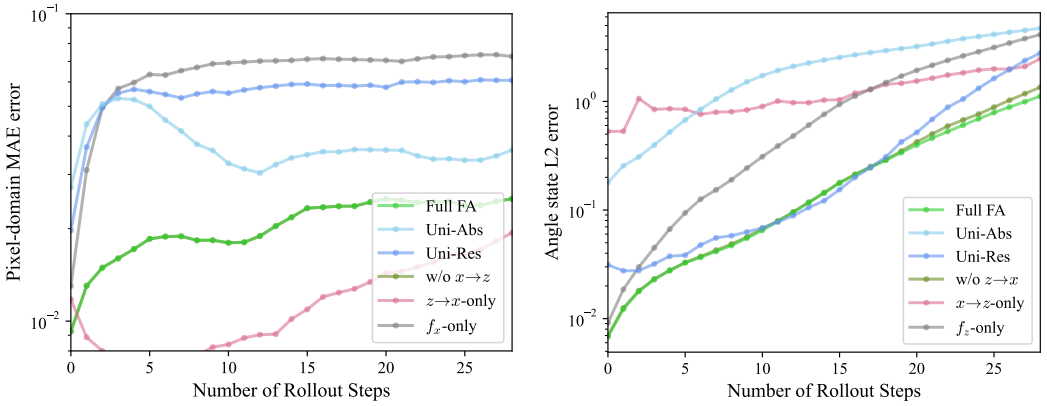
Under this regime, single-direction alignment variants (e.g., $x \to z$-only or $z \to x$-only) can perform competitively and, in some cases, outperform full function alignment. This behavior is expected, as the availability of ground-truth context reduces the need for bidirectional error correction. However, these advantages do not transfer to unconditional co-generation, where errors from both modalities accumulate and must be mutually corrected. Consistent with our main results, full function alignment is most effective in the unconditional setting, which constitutes the primary evaluation in this work.



(a) Perceptual-domain prediction error conditioned on ground-truth analytic history.

(b) Analytic-domain prediction error conditioned on ground-truth perceptual history.

Figure 8: Conditional co-generation performance on the WBB environment given a prompt of 3 steps. Prediction error in both the perceptual and analytic domains is shown as a function of rollout horizon. Errors are plotted on a logarithmic scale.



(a) Perceptual-domain prediction error conditioned on ground-truth analytic history.

(b) Analytic-domain prediction error conditioned on ground-truth perceptual history.

Figure 9: Conditional co-generation performance on the DP environment given a prompt of 3 steps. Prediction error in both the perceptual and analytic domains is shown as a function of rollout horizon. Errors are plotted on a logarithmic scale.

E.2 Additional Qualitative Results

Figure 10 visualizes the joint evolution of analytic prediction errors for the two pendulum angles under unconditional rollout in the DP environment. Each curve traces how the prediction errors of the first and second pendulums grow together over time, starting from the origin. Full FA remains **closest to the origin** and exhibits the slowest joint error growth, indicating suppressed error coupling across state dimensions. In contrast, unified dynamics (Uni-Res) and standalone analytic dynamics (f_z-only) drift rapidly toward the upper-right region, reflecting strong amplification of coupled errors in the chaotic regime. The single-direction variant (w/o $z \to x$) shows intermediate behavior, suggesting that bidirectional alignment is necessary to stabilize joint error growth.

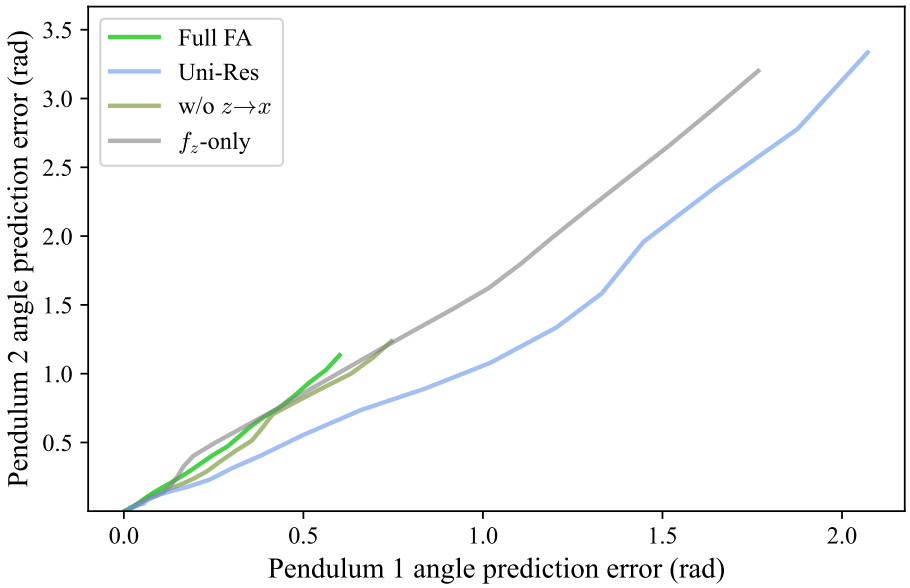


Figure 10: Joint evolution of prediction errors for the two pendulum angles under unconditional rollout. Full function alignment suppresses coupled error growth, while unified and single-direction dynamics diverge rapidly in joint error space.