

# EgoTraj-Bench: Towards Robust Trajectory Prediction under Ego-view Noisy Observations

Jiayi Liu<sup>1</sup>, Jiaming Zhou<sup>1</sup>, Ke Ye<sup>1</sup>, Kun-Yu Lin<sup>2</sup>, Allan Wang<sup>3</sup>, and Junwei Liang<sup>1,4\*</sup>

**Abstract**—Reliable trajectory prediction from an ego-centric perspective is crucial for robotic navigation in human-centric environments. However, existing methods typically assume noiseless observation histories, failing to account for the perceptual artifacts inherent in first-person vision, such as occlusions, ID switches, and tracking drift. This discrepancy between training assumptions and deployment reality severely limits model robustness. To bridge this gap, we introduce EgoTraj-Bench, built upon the TBD dataset, which is the first real-world benchmark that aligns noisy, first-person visual histories with clean, bird’s-eye-view future trajectories, enabling robust learning under realistic perceptual constraints. Building on this benchmark, we propose BiFlow, a dual-stream flow matching model that concurrently denoises historical observations and forecasts future motion. To better model agent intent, BiFlow incorporates our EgoAnchor mechanism, which conditions the prediction decoder on distilled historical features via feature modulation. Extensive experiments show that BiFlow achieves state-of-the-art performance, reducing minADE and minFDE by 10–15% on average and demonstrating superior robustness. We anticipate that our benchmark and model will provide a critical foundation for robust real-world ego-centric trajectory prediction. The benchmark library is available at: <https://github.com/zoeyliu1999/EgoTraj-Bench>.

## I. INTRODUCTION

Pedestrian trajectory prediction [1], [2], aiming to estimate the multimodal future paths of agents in dynamic environments, serves as a foundation for safe, socially compliant motion planning in autonomous systems such as mobile robots, intelligent prosthetics, and service vehicles [3]–[7]. Although extensively studied, most existing methods are developed and evaluated under idealized bird’s-eye view (BEV) settings with globally consistent observations and flawless agent tracking [2]. However, these conditions rarely hold in real-world deployment. Autonomous agents, such as mobile robots, typically perceive the environment through front-facing cameras, where observations are inherently incomplete and noisy as illustrated in Fig. 1: pedestrians may be occluded, enter or exit the field of view (FOV), or suffer from physical sensing errors such as perspective distortion. These imperfections substantially violate the idealized historical assumptions in BEV settings. Therefore, trajectory prediction under ego-centric (first-person view, FPV) noisy observations

is essential for enabling robust deployment in real-world scenarios.

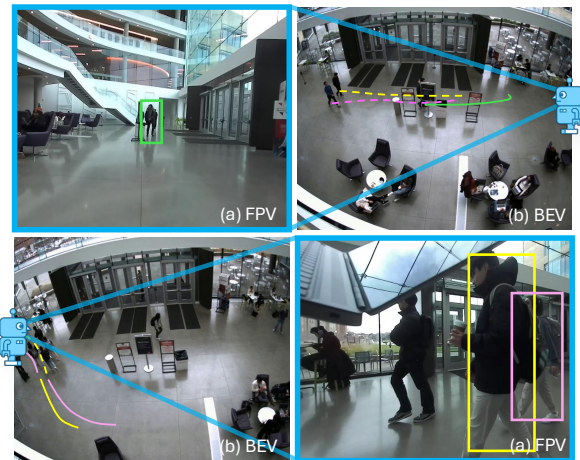


Fig. 1: Illustration of key challenges under ego-view observations. **Top row: Occlusion.** In the first-person view (a), only one pedestrian (green box) is visible; the corresponding BEV (b) shows two additional agents (pink and yellow) behind the green pedestrian. Dashed lines indicate trajectories visible in BEV but not in FPV. **Bottom row: ID Switch and Perspective Distortion.** Two pedestrians (yellow and pink) cross paths, causing an ID swap in the FPV tracking output (a). Additionally, individuals near the image corners suffer significant perspective distortion, making accurate localization challenging.

Some prior work [6], [8], [9] predicts trajectories from ego-centric input, typically predicting future positions in image space, e.g., bounding boxes or keypoints. Despite operating from an ego-centric view, these methods lack spatial reasoning in real-world space, assume idealized tracking results in image space, and thus leave the modeling of fine-grained interactions unresolved in real-world trajectory prediction. In contrast, methods [10], [11] that predict trajectories in global metric spaces (e.g., world coordinates) enable precise spatial reasoning about proximity, collision risk, and social norms, and we therefore focus on the latter paradigm.

In addition, some studies [12], [13] simulate ego-centric conditions by rendering BEV data into synthetic views using simulators. While this approximates the visual input of moving entities, the rule-based agent motion and simplified rendering in the simulator fail to capture the intricate and subtle motion patterns and visual nuances present in authentic scenes. Moreover, the utilized BEV data [14], [15] are collected in open and uncluttered environments such as streets

<sup>1</sup>The Hong Kong University of Science and Technology (Guangzhou). Jiayi.LIU@connect.hkust-gz.edu.cn, junweiliang@hkust-gz.edu.cn \* Corresponding author.

<sup>2</sup>The University of Hong Kong.

<sup>3</sup>Miraikan – The National Museum of Emerging Science and Innovation.

<sup>4</sup>The Hong Kong University of Science and Technology.

This work was supported by the National Natural Science Foundation of China (No. 62306257).

with few static obstacles, resulting in overly clean inputs that do not reflect the perception challenges of dense, interactive environments. These limitations highlight the critical need for a real-world benchmark for robust trajectory prediction with ego-centric noisy observations.

To this end, we introduce EgoTraj-Bench, the first real-world benchmark for trajectory prediction under ego-centric noise. Built upon the TBD dataset [16], EgoTraj-Bench first derives historical trajectories with noise from real ego-view videos, capturing deployment-realistic imperfections such as occlusions, mis-tracked IDs, FOV truncations, and perspective distortions. Furthermore, the observed ego-centric trajectory with noise is projected into world coordinates and paired with the corresponding clean, human-verified future trajectory from the BEV view, ensuring metric-consistent supervision while preserving the realism of ego-centric input conditions. This practice can transfer the disturbance from the ego-view noise to the widely used BEV-based trajectory prediction framework, thereby providing a fairer and trustworthy platform for systematically evaluating existing BEV-based trajectory prediction methods. The benchmarking results show that state-of-the-art BEV-based models suffer significant performance degradation when their input of historical observations is disturbed by the ego-view noise, underscoring the need for new frameworks for robust trajectory prediction under real-world ego-view perturbations.

To address this problem, we propose BiFlow, a novel noise-resistant dual-stream flow matching model as an example solution for our benchmark. BiFlow jointly recovers the observed noisy historical trajectories and predicts future trajectories. By jointly learning latent features across the two tasks, the model implicitly leverages denoised historical semantics to guide future trajectory predictions, improving robustness while maintaining parameter efficiency. In addition, we introduce EgoAnchor, a mechanism to distill compact, ego-centric tokens from agent- and scene-level histories. These intent-aware representations, extracted via attention mechanism during history reconstruction, are injected into the decoder via feature-wise affine modulation, providing a robust intent prior to stabilize prediction under partial or corrupted input.

The main contributions of our work are: 1) We introduce EgoTraj-Bench, the first real-world benchmark for trajectory prediction under deployment-realistic conditions, enabling rigorous evaluation of models under authentic ego-centric noisy perturbations; 2) We propose a novel dual-stream flow matching framework with a distillation mechanism, which jointly recovers noisy historical observations and predicts future trajectories, aiming to leverage clean historical semantics to facilitate and stabilize future forecasting; 3) Our experiments demonstrate the significant impact of ego-view noise on existing models and the robustness of our proposed approach, which outperforms baselines by over 10% in minADE and 13% in minFDE averaged over datasets, highlighting the importance of noise-aware modeling and providing valuable insights for future research in ego-view realistic trajectory prediction.

## II. RELATED WORK

### A. Pedestrian Trajectory Prediction

Various models have been proposed for pedestrian trajectory prediction. Social-LSTM [1] pioneers agent interaction modeling via pooled LSTM states. Social GAN [2] and AC-VRNN [17] introduce generative frameworks to capture multimodal futures. More recently, Transformer-based architectures like TUTR [18] leverage self-attention for long-range spatiotemporal reasoning. Denoising models, particularly diffusion models and flow-based generative models [19], [20], further improve the modeling of complex multimodal trajectory distributions via iterative refinement. MID [21] first applies diffusion to predict future trajectory, while Leapfrog [22] introduces a deterministic latent initializer to speed up sampling. However, both sample futures independently per agent, leading to spatially redundant trajectories. MoFlow [23] uses flow matching (FM) to jointly model futures conditioned on past motion, promoting learning of diverse future trajectories. These advances motivate adopting FM to generate future trajectories from partial and noisy ego-centric inputs, leveraging its ability to condition on corrupted histories and support diverse predictions.

Beyond architecture, several methods incorporate auxiliary inputs, e.g., goal estimates [11], [24] or scene graphs [10], to enrich contextual cues. However, these typically assume clean, globally observed trajectories, an assumption violated in real-world ego-centric settings. While some studies attempt to enhance robustness via random history masking [25]–[27], such artificial missingness fails to reflect the complex and structured nature of real ego-centric perception. This gap motivates the need for trajectory prediction approaches under realistic ego-centric perturbations.

### B. Trajectory Prediction under Ego-centric Noise

Recent efforts address ego-centric prediction via FPV or BEV paradigms. FPV-based methods, such as [8] and [9], predict pedestrian locations (e.g., bounding boxes or keypoints) in image space. While effective for visual tracking or localization, these rely on idealized image-space observations and unscaled pixel coordinates, limiting the ability to support metric-space reasoning, such as real-world proximity or collision risk.

BEV-based approaches project perturbed ego-centric trajectories to world coordinates, enabling prediction within a shared metric framework (e.g., in meters), where spatial relationships can be precisely computed. Fvtraj [12] and T2FPV [13] generate synthetic FPV videos by rendering BEV trajectories in simulators. While enabling controlled study of perception noise, their rule-based agent behaviors and simplified visual rendering fail to capture subtle motion patterns and realistic visual nuances, consequently limiting the model’s realism and generalizability. Moreover, underlying BEV datasets (i.e., ETH-UCY [14], [15]) were collected in open, sparse environments with minimal obstacles, yielding rendered FPV sequences failing to reflect the dense, dynamic scenes typical of real-world deployment. While

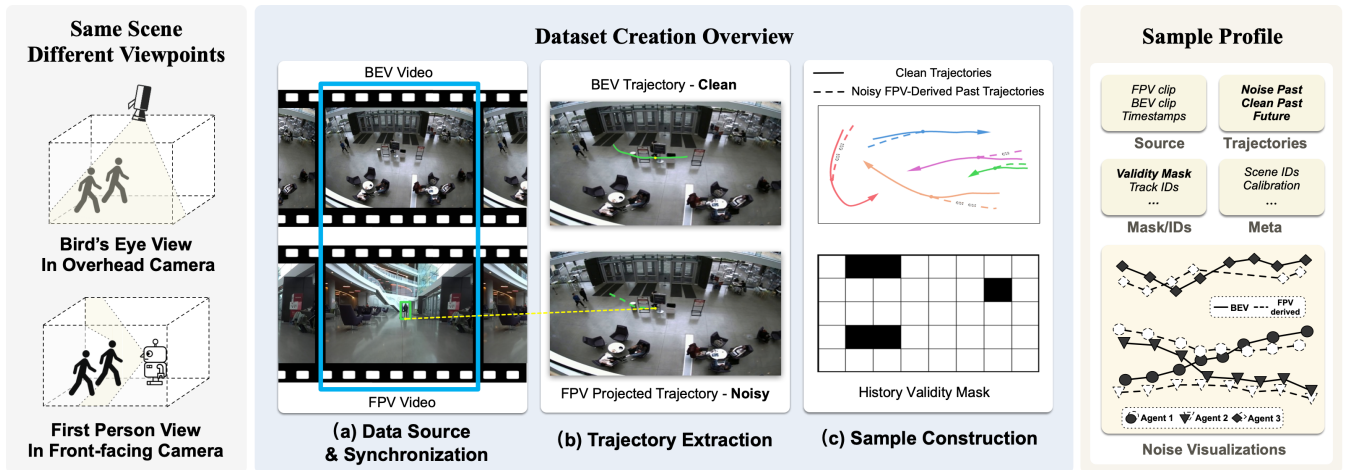


Fig. 2: EgoTraj-Bench Overview: **Left** contrasts the same scene in BEV vs FPV, where limited visibility and viewpoint changes introduce noisy histories. **Mid** shows sample construction: synchronized BEV–FPV clips provide noisy FPV-derived past trajectories as inputs and clean metric BEV trajectories as supervision, with a history validity mask. **Right** illustrates what the benchmark records and the key noise patterns that drive robustness gaps beyond idealized BEV settings.

several real-world ego-centric datasets exist [28], [29], most focus on past or present state estimation without providing future ground truth in world coordinates, or involve agents (e.g., vehicles) with motion patterns fundamentally different from pedestrian-scale robots.

Methodologically, T2FPV’s CoFE module denoises historical perturbations using clean history as supervision [13], but only corrects missing positions when constructing the input of prediction model. This hybrid representation, combining uncorrected observed positions with corrected ones, risks spatial-temporal discontinuities at segment boundaries, potentially causing artificial motion jumps during prediction. These limitations motivates a unified framework that avoids such patchwork reconstruction, together with a real-world benchmark capturing both noisy perception inputs and metric-accurate ground truth.

### III. EGOTRAJ-BENCH: EGO-VIEW TRAJECTORY PREDICTION BENCHMARK

To bridge the gap between idealized BEV-based trajectory prediction and real-world deployment under ego-view perception noise, we introduce EgoTraj-Bench, a novel real-world benchmark that transfers realistic ego-view induced noise into BEV coordinate space, enabling fair evaluation of existing BEV-based models and fine-grained spatial reasoning under deployment-realistic conditions.

#### A. Real-world Dataset Creation

We construct the core of EgoTraj-Bench by deriving noisy trajectories from real-world first-person video as shown in Fig. 2. Rather than simulating perception artifacts as in [13], we capture authentic perception artifacts, e.g., occlusions, identity switches, and ego-motion drift, arising from dynamic human-robot interactions and physical sensor limitations in unstructured environments.

The generated pixel-space trajectories from authentic videos are projected into the global BEV coordinate system

using calibrated camera intrinsics and time-synchronized robot ego-motion (the information recorded during data collection), ensuring metric consistency with BEV-based prediction models. Each ego-view-derived noisy history trajectory in BEV space is then temporally aligned with its corresponding clean past and future trajectory extracted from overhead cameras. This paired structure, i.e., noisy history as input, clean past and future as supervision, enables rigorous and fair evaluation of trajectory prediction robustness under realistic ego-view noise, while preserving the spatial reasoning supporting navigation in crowded scenes.

In the following sections, we detail the full pipeline across three stages: Data Source and Synchronization, Trajectory Extraction, and Sample Construction.

**- Data Source and Synchronization.** To establish a foundation for injecting and evaluating real-world ego-view noise, we build upon the publicly available TBD dataset [16], which uniquely provides synchronized BEV videos from overhead cameras and ego-view videos from mobile robots. While THör [31] also offers dual-view data, it was collected in a controlled laboratory setting, resulting in less natural human behavior and reduced environmental diversity. We select segments where the robot is actively moving and collecting pedestrian data, and for each ego-view clip, extract its temporally aligned BEV counterpart. This synchronization allows every trajectory extracted from ego-view video, inherently noisy and perspective-distorted, to be geometrically projected into the shared world coordinate system, where they can be directly spatially aligned with clean ground-truth trajectories for fair and consistent evaluation.

**- Trajectory Extraction.** To ensure perception noise reflects real deployment conditions, we extract pedestrian trajectories from raw FPV video using YOLOv8 [32] for detection, selected for its robustness in crowded scenes, and BotSort [33] for tracking, which fuses motion and appearance cues to mitigate ID switches. Visibility is quantified per frame via segmentation masks from YOLOv8-seg. All hyperparameters

TABLE I: Dataset Statistics Comparison.

Dataset	Fold	Duration (min)	Label Freq (Hz)	# Traj	Ego-noise Involved			Agent Scale	FPV Noisy Rate <sup>†</sup>	History MSE <sup>*</sup> (m)
					Perceptual	Real-world	Physical			
WildTrack [30]	-	3	2	313	✗		✗	pedestrian	-	-
THoR [31]	-	60	100	600	✓		✗	pedestrian	-	-
nuScenes [28]	-	330	12	-	✓		✓	✗ vehicle	-	-
T2FPV-ETH [13]	ETH	9	15	181	✓		✗	pedestrian	0.44	5.66
	Hotel	13	15	1,053	✓		✗	pedestrian	0.51	4.55
	Zara1	6	2.5	5,939	✓		✗	pedestrian	0.28	1.23
	Zara2	7	2.5	17,608	✓		✗	pedestrian	0.32	2.50
	Univ	3.5	2.5	24,334	✓		✗	pedestrian	0.45	2.47
EgoTraj-TBD	-	210	30	36,947	✓		✓	pedestrian	0.37	0.66

<sup>†,\*</sup> **FPV noisy rate** indicates the average probability of being marked as invalid. **History MSE** indicates the distance between estimated and ground-truth history. Further details, such as density statistics, are available in TBD [16].

are tuned for different scenes. The final 2D bounding box bottom centers are back-projected to the ground plane using calibrated intrinsics and synchronized ego-motion from the TBD dataset, yielding BEV-space trajectories that retain realistic noise such as occlusion, ID instability, and localization error. For the supervision of clean past and future trajectories, we use the BEV annotations provided in the TBD dataset, benefiting from occlusion-resilient multi-view coverage and semi-automated human verification.

**- Sample Construction.** To enable supervised learning under ego-view noise, we align each noisy ego-view derived history with its corresponding clean past and future trajectory from BEV annotations. Due to unaligned track IDs from independent FPV and BEV pipelines, input and ground-truth tracks are associated using Hungarian matching as in [13], [16], [34]. Instead of relying solely on mean squared error (MSE) in absolute position, we compute weighted MSEs incorporating location, velocity, and acceleration to enhance matching robustness under noise. Following common practice [14], [15], [35], we adopt an 8-second sliding window for each sample: 8 frames (3.2 s) for observation and subsequent 12 frames (4.8 s) for prediction. Only samples with at least three valid observation frames are retained, where validity is determined by visibility (more than 100 pixels in the segmentation mask) and motion plausibility (estimated frame speed less than 2 m/s). Missing observations within valid samples are marked and can be filled via techniques such as linear interpolation to ensure fixed-length inputs, while the robot’s trajectory is included to model agent-robot interaction. The final dataset contains 36,947 aligned pairs, each linking noisy BEV-aligned histories to clean BEV past and future.

### B. Statistics and Analysis

As summarized in Table I, EgoTraj-Bench provides 210 minutes of real-world recordings at 30 Hz, offering extended interaction diversity and fine temporal resolution for ego-centric modeling. While the total number of detected trajectories is comparable to synthetic benchmarks T2FPV-ETH, the generation strategy differs fundamentally: T2FPV-ETH treats every agent as a virtual ego-observer, artificially multiplying samples per scene, whereas EgoTraj-Bench preserves

natural perceptual bias by including only one true mobile observer per scene.

EgoTraj-TBD covers two key challenge types: (1) inherent perceptual artifacts (e.g., occlusion, FOV truncation, ID switching); and (2) physical sensor artifacts (e.g., lens distortion, projection error, ego-motion drift), important to real-world systems. EgoTraj-Bench also focuses on pedestrian-scale agents in human-centric environments, whose surroundings and motion patterns differ fundamentally from vehicle-centric benchmarks such as nuScenes [28]. As reported in Table I, the dataset features a moderate noise rate and lower historical alignment error, which could be explained by higher-fidelity ground truth and a more noise-aware processing pipeline as mentioned in Sec. III-A. During learning, the dataset is chronologically split into train, validation, and test sets (70%-10%-20%) to ensure temporal coherence and avoid data leakage.

### C. Evaluation Framework and Findings

**- Datasets.** We utilize EgoTraj-TBD, a real-world pedestrian trajectory dataset with ego-centric perturbations, recorded indoors with complex layouts and dynamic obstacles. In addition, we include the simulated T2FPV dataset for reference, which was collected outdoors across five folds at varying times and locations.

**- Metrics.** Following the most commonly used ones in trajectory prediction, we adopt Average Displacement Error (ADE) and Final Displacement Error (FDE) as primary metrics, measuring average and final Euclidean distance between estimated and ground-truth trajectories. Since we evaluate multimodal methods that generate K candidate trajectories per agent, we report minADE@K and minFDE@K, which compute ADE and FDE of the best-matching trajectory among K outputs, rewarding models for diverse yet accurate predictions.

**- Models.** To assess robustness under noisy FPV observations, we evaluate three classes of state-of-the-art trajectory forecasting models: 1) Recurrent models that iteratively model temporal dynamics through stochastic latent variables, including VRNN [36], AC-VRNN [17], and SGNet [11]; 2) Transformer-based models that leverage self-attention mechanisms to capture long-range dependencies and social

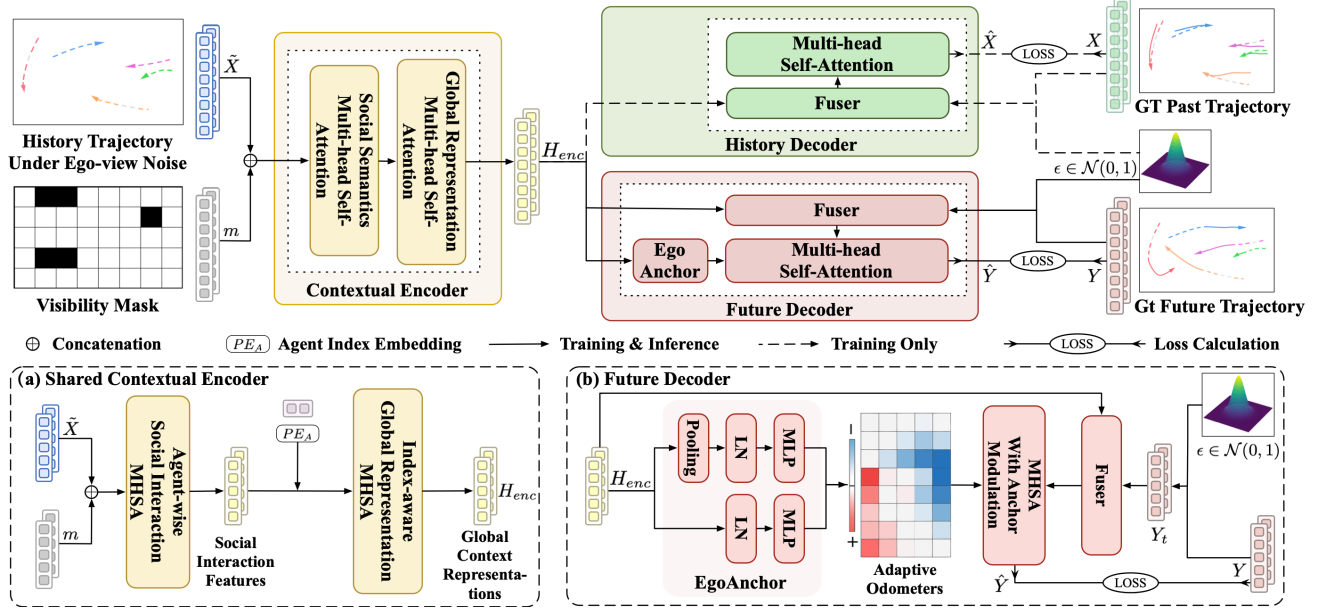


Fig. 3: Overview of our BiFlow. The input consists of a noisy historical trajectory  $\tilde{X}$  and its corresponding validity mask  $m$ . During training, the model is supervised with clean ground-truth past  $X$  and future  $Y$  trajectories to jointly learn reconstruction and prediction. At inference, only the noisy history and mask are used as input to predict the future trajectory  $\hat{Y}$ .

interactions, represented by TUTR [18]; and 3) Flow-based generative models that learn data distributions via iterative noise refinement, specifically MoFlow [23]. This diverse model suite enables a comprehensive analysis of real-world ego-view perceptual challenges. Additionally, to enable a direct comparison with the T2FPV framework, we include its Correction of FPV Errors (CoFE) module [13], a refinement component trained end-to-end with the prediction model as a baseline for noise correction.

- **Empirical Findings.** As shown in Table II, our benchmark reveals a significant gap: all BEV-trained state-of-the-art models exhibit substantial degradation under ego-view perception noise. Particularly, on the widely adopted ETH-UCY datasets, the minADE@20 rises to 0.67m, compared to approximately 0.20m under clean historical trajectories [23]. This highlights a critical limitation that existing methods are highly sensitive to perception artifacts prevalent in ego-view data. These findings underscore the need for noise-aware modeling in trajectory prediction.

## IV. PROPOSED METHOD

### A. Problem Formulation

We consider the task of multi-agent trajectory prediction under ego-centric observation noise. Let  $A$  denote the total number of agents in the scene. The input consists of: (1) the observed history  $T_p$  steps of all agents  $\tilde{X} \in \mathbb{R}^{A \times 2T_p}$ , structured as  $\tilde{X} = [\tilde{X}_{\text{other}}; \tilde{X}_{\text{ego}}]$ , where  $\tilde{X}_{\text{other}} \in \mathbb{R}^{(A-1) \times 2T_p}$  contains noisy positions of non-ego agents observed from FPV video, and  $\tilde{X}_{\text{ego}} \in \mathbb{R}^{2T_p}$  contains clean, fully observable positions of the ego-agent from robot odometry; (2) a binary validity mask  $m \in \{0, 1\}^{A \times T_p}$ , structured as  $m = [m_{\text{other}}; \mathbf{1}_{T_p}]$ , where  $m_{\text{other}} \in \{0, 1\}^{(A-1) \times T_p}$  indicates validity of non-ego agents, and  $\mathbf{1}_{T_p}$  denotes full observability of the ego-agent across all  $T_p$  steps.

The goal is to predict the future trajectories  $Y \in \mathbb{R}^{A \times 2T_f}$  of all agents over the next  $T_f$  time steps.

### B. Overview

Motivated by our benchmark findings that ego-view noise severely degrades the trajectory prediction performance of existing methods, and by the strong performance and efficient conditional generation of recent flow-matching approaches, we propose **BiFlow**, a dual-stream framework that jointly reconstructs clean agent history trajectories and predicts their future trajectories.

As illustrated in Fig. 3, BiFlow’s core idea is to transfer denoised motion patterns from history reconstruction to stabilize future prediction under partial or corrupted observations. This is realized through three key components: (1) a noise-aware contextual encoder that models Social Interactions under occlusion; (2) an EgoAnchor mechanism that distills intent priors from history hidden features without future supervision; and (3) a dual decoder architecture that shares latent representations of the encoder’s output but modulates future prediction via historical confidence.

Our proposed BiFlow adopts a dual-stream flow matching framework to jointly learn two mappings from the same input  $\tilde{X}$ : (1) reconstructing the clean history trajectory  $X_1 \in \mathbb{R}^{A \times 2T_p}$ , and (2) predicting the clean future trajectory  $Y_1 \in \mathbb{R}^{A \times 2T_f}$ . For each task, we sample a noise vector (i.e.,  $X_0 \sim N(0, I)$  for history,  $Y_0 \sim N(0, I)$  for future) and construct interpolated states:

$$X_t = (1-t)X_0 + tX_1, \quad t \in [0, 1] \quad (1)$$

$$Y_t = (1-t')Y_0 + t'Y_1, \quad t' \in [0, 1] \quad (2)$$

where  $t$  and  $t'$  are sampled independently to provide diverse supervision in training.

The model is trained to denoise  $X_1$  and  $Y_1$  conditioned on the observation  $\tilde{X}$ , time  $t$ , and interpolated states  $X_t$  or  $Y_t$ , using a MoFlow-style [23] multi-candidate objective to ensure trajectory coherence and diversity. The total loss combines the flow matching losses from two trajectory prediction tasks,

$$L_{\text{total}} = \lambda_1 L_{\text{recon}} + \lambda_2 L_{\text{pred}}, \quad (3)$$

where  $L_{\text{recon}}$  and  $L_{\text{pred}}$  will be elaborated in Sec. IV-E.

### C. Contextual Encoder: Modeling Noisy Social Dynamics

To capture the noisy dynamics and social interactions of agents' past trajectories, we design an agent-aware contextual encoder using the Transformer architecture.

Given the noisy historical trajectory data  $\tilde{X}$  and validity mask  $m$ , we first apply multi-head self-attention (MHSA) to model inter-agent social dependencies. We then add a learnable agent-identity encoding  $\text{PE}_A$  and apply a second MHSA layer to aggregate a coherent scene-level representation:

$$H_0 = \text{MLP}(\tilde{X} \oplus m), \quad (4)$$

$$H_{\text{soc}} = \text{MHSA}(Q = H_0, K = H_0, V = H_0), \quad (5)$$

$$H'_{\text{soc}} = H_{\text{soc}} + \text{PE}_A, \quad (6)$$

$$H_{\text{enc}} = \text{MHSA}(Q = H'_{\text{soc}}, K = H'_{\text{soc}}, V = H'_{\text{soc}}), \quad (7)$$

The output  $H_{\text{enc}} \in \mathbb{R}^{A \times D}$  serves as the shared latent representation for both reconstruction and prediction streams, where  $D$  is the feature dimension.

### D. EgoAnchor: Intent Prior Distillation

Drawing on intent-driven models [11], [24] that infer long-term goals to guide behavior prediction, we introduce **EgoAnchor**, a lightweight mechanism that distills intent priors from  $H_{\text{enc}}$ . These priors are constructed to encode historical context to stabilize predictions under partial or corrupted observations.

$$\text{Ach}_{\text{agent}} = \text{MLP}(\text{LayerNorm}(H_{\text{enc}})), \quad (8)$$

$$\text{Ach}_{\text{scene}} = \text{MLP}(\text{LayerNorm}(\text{Mean}(\text{Ach}_{\text{agent}}))), \quad (9)$$

where  $\text{Ach}_{\text{agent}} \in \mathbb{R}^{A \times D}$  is used to capture agent-level motion tendencies, and  $\text{Ach}_{\text{scene}} \in \mathbb{R}^D$  summarizes global context. We use layer normalization to handle crowd size, validity, and camera noise variations in  $H_{\text{enc}}$ , making the output anchor comparable across agents and scenes.

To reduce computation, EgoAnchor operates in a self-supervised manner by eschewing additional intent labels. The extracted anchors are integrated into the future decoder via feature-wise affine modulation [37], modulating feature distribution based on agent/scene-level intent and reliability.

### E. Dual Decoder with Multi-Candidate Prediction

Motivated by leveraging clean motion patterns from history reconstruction to future prediction, we employ two independent decoders that share the encoder output  $H_{\text{enc}}$ , as shown in Figure 3.

In **future prediction** stream, we predict clean future trajectory  $Y_1$  from  $H_{\text{enc}}$ , time step  $t'$ , and noisy intermediate

state  $Y_{t'}$ . These inputs are first fused into a hidden state, which is then processed through  $K$ -to- $K$  MHSA blocks to model interactions among  $K$  candidate trajectories for diverse yet coherent predictions. The resulting hidden state  $Z$  is modulated using intent priors from the EgoAnchor module:

$$(\beta_{\text{Ach}}, \gamma_{\text{Ach}}) = \text{MLP}(\text{Ach}_{\text{agent}} + \text{Ach}_{\text{scene}}), \quad (10)$$

$$Z_{\text{Ach}} = (1 + \gamma_{\text{Ach}}) \odot Z + \beta_{\text{Ach}}, \quad (11)$$

$$H_{\text{dec}} = \text{MHSA}(Q = Z_{\text{Ach}}, K = Z_{\text{Ach}}, V = Z_{\text{Ach}}) \quad (12)$$

where  $\beta_{\text{Ach}}$  and  $\gamma_{\text{Ach}}$  are affine parameters adaptively modulating feature space based on historical confidence, analogous to adaptive odometers (e.g., amplifying high-confidence features). To refine inter-agent dependencies, a subsequent MHSA is employed. Finally, an MLP head maps  $H_{\text{dec}}$  to candidate predictions  $\hat{Y}^{1:K}$  and logits  $\hat{c}_y$ . The training objective for the future prediction task is:

$$L_{\text{pred}} = \mathbb{E}_{t', Y_1, Y_{t'}} \left[ \|\hat{Y}^{1:K} - Y_1\|_2^2 + \text{CE}(c_y^{1:K}, j_y^*) \right] \quad (13)$$

$$j_y^* = \arg \min_j \|\hat{Y}^j - Y_1\|_2^2 \quad (14)$$

where  $j_y^*$  identifies the best-matching candidate, and  $\text{CE}(\cdot, \cdot)$  is the cross-entropy loss over mode selection.

The **history reconstruction** stream mirrors identical structure but sets  $\beta_{\text{Ach}}$  and  $\gamma_{\text{Ach}}$  to 0, enabling pure reconstruction learning. During inference, the history reconstruction decoder is no longer activated. Instead, we use only the noisy input trajectory  $\tilde{X}$ , which is processed through the shared contextual encoder and the prediction decoder and conditioned on intent priors from the EgoAnchor module, to forecast future trajectories.

## V. EXPERIMENTS

### A. Implementation Details

We evaluate BiFlow and existing baseline methods on our proposed EgoTraj-Bench, as detailed in Sec. III-C. In our approach, all trajectories are normalized to stabilize training. In the history reconstruction branch, we use absolute coordinates as targets to facilitate denoising; in the future prediction branch, we adopt displacement-based (relative) targets to improve temporal coherence. For feature modulation in the future stream, we integrate EgoAnchor through a 4-layer MHSA block, which enables progressive fusion of structured prior information into the decoder. We sample from the model using 10 denoising steps based on a logit-normal time scheduler. The model is trained for 150 epochs with a batch size of 64 and a latent dimension of 128. We use the AdamW optimizer with an initial learning rate of 0.001, a cosine annealing learning rate schedule with warmup, and a weight decay of 0.01.

### B. Quantitative Results

We present performance comparisons across all models on both datasets in Table II. Across architectures, ego-view noise substantially degrades performance relative to clean-history results. Recurrent models (VRNN/AC-VRNN/SGNet) degrade more severely due to error accumulation under corrupted histories, while the attention-based

TABLE II: Quantitative results across EgoTraj-TBD and T2FPV-ETH using minADE@20 / minFDE@20 (in meters). Best and second-best performances are highlighted in **bold** and underlined, respectively.

Model →	VRNN [36]		AC-VRNN [17]		SGNet [11]		TUTR* [18]		MoFlow [23]		BiFlow
	–	+ CoFE	–	+ CoFE	–	+ CoFE	–	+ CoFE	–	+ CoFE	
ETH <sup>†</sup>	1.35/2.00	1.52/2.35	1.39/2.04	1.47/2.18	1.43/1.97	0.98/1.32	1.25/1.54	1.02/1.41	0.85/1.22	0.88/1.26	<b>0.66/0.85</b>
HOTEL	1.30/1.73	1.06/1.53	1.31/1.75	1.16/1.72	0.72/1.00	0.59/0.76	1.04/1.47	0.85/1.26	0.63/0.89	0.62/0.87	<b>0.49/0.59</b>
ZARA1	1.14/1.68	1.65/2.10	1.04/1.40	1.03/1.48	0.58/0.79	0.55/0.76	0.77/1.00	0.62/0.93	0.46/0.63	0.52/0.67	<b>0.42/0.58</b>
ZARA2	1.54/2.10	1.06/1.63	1.47/1.93	1.31/1.69	0.78/0.91	0.73/0.86	0.75/0.91	0.72/0.82	<b>0.50/0.60</b>	0.53/0.63	<u>0.50/0.62</u>
UNIV	2.24/2.89	1.27/1.62	2.26/3.00	1.54/1.88	1.48/1.73	1.23/1.48	1.10/1.30	1.02/1.24	<u>0.92/1.10</u>	0.92/1.08	<b>0.91/1.08</b>
AVG	1.51/2.08	1.31/1.84	1.49/2.03	1.30/1.79	1.00/1.28	0.82/1.04	0.98/1.12	0.85/1.13	0.67/0.88	0.69/0.90	<b>0.60/0.74</b>
<b>Ego-TBD</b>	0.76/1.26	0.68/1.19	0.82/1.38	0.63/1.09	0.34/0.52	0.37/0.58	0.58/0.72	0.52/0.67	<u>0.21/0.29</u>	0.26/0.36	<b>0.19/0.27</b>

<sup>†</sup> ETH/HOTEL/ZARA1/ZARA2/UNIV are the five folds from T2FPV-ETH.

\* The TUTR architecture is adapted to support multimodal output to ensure a fair comparison.

TABLE III: Ablation study on EgoTraj-TBD.

Model	Components			ADE/FDE@K		
	SI	EA	SE	k=1	k=5	k=10
MoFlow	-	-	-	0.84/1.45	0.46/0.76	0.31/0.48
BiFlow	✓	✗	✗	0.76/1.37	0.42/0.72	0.29/0.47
	✓	✓	✗	0.73/1.32	0.40/0.69	0.28/0.45
	✓	✓	✓	<b>0.70/1.21</b>	<b>0.38/0.63</b>	<b>0.26/0.41</b>

predictor TUTR is generally more competitive but still struggles under realistic noise. Among prior methods, the flow-matching model MoFlow is the strongest competitor, suggesting that conditional generative modeling is promising for forecasting from noisy histories, yet it still degrades under corrupted observations.

Under ego-view noise, our method is both more robust and more efficient. On T2FPV-ETH, it achieves state-of-the-art (SOTA) performance with minADE@20 of 0.60 and minFDE@20 of 0.74, outperforming the prior SOTA by over 11% and 15%, respectively. On EgoTraj-TBD, it shows consistent improvements over existing methods. Notably, as shown in Table III, our model achieves significant improvements when generating fewer future trajectory candidates (i.e., smaller  $K$ ). Specifically, with the same reduced number of samples, our approach improves minADE and minFDE by around 16% compared to the prior SOTA. The strong performance indicates a predicted distribution more closely aligned with true trajectories, underscoring enhanced robustness and predictive efficiency under noisy conditions.

Additionally, while integrating CoFE brings improvements for some models, the gains are limited, and can even degrade flow-based methods. This suggests that correcting only missing positions is insufficient, as ego-centric observations contain structured noise beyond occlusion, including tracking errors and perspective distortion. Effective denoising therefore requires holistic trajectory modeling rather than patchwise correction, which is validated by our SOTA results.

Note that for all models, including BiFlow, the ADE/FDE values on EgoTraj-TBD are lower than those evaluated on T2FPV-ETH. A key reason is our higher-fidelity ground truth and more noise-aware processing pipeline, which results in a significantly lower historical MSE (0.66m) in EgoTraj-

TBD, as shown in Table I. These more meaningful historical representations enable more reliable training and contribute to overall better performance across methods.

### C. Ablation Study

Table III presents an ablation study of BiFlow on EgoTraj-TBD: Social Interaction (SI) within the contextual encoder, EgoAnchor (EA) distillation, and Shared Encoder (SE) used by both streams. We compare against MoFlow as a strong baseline, which can be viewed as a representative single-flow predictor that directly forecasts futures from past motion. Although it is not a strictly architecture-matched dual-versus-single control, BiFlow’s consistent improvements over MoFlow suggest that jointly modeling history denoising and future forecasting is beneficial under ego-view noise.

We then ablate components by enabling them progressively. When only SI is added, BiFlow achieves notable improvements over MoFlow, reducing minADE/minFDE by over 9% at  $K=1$ . Incorporating EA further enhances performance, yielding a 13% improvement with the same  $K$ . With all components (SI, EA, and SE) integrated, the full model achieves the best performance, improving minADE and minFDE by 16%, 17%, and 16% at  $K=1$ ,  $K=5$ , and  $K=10$ , respectively. These results demonstrate that integrating social interaction cues and ego-centric intent patterns, particularly through EgoAnchor, significantly improves prediction accuracy and robustness under the deployment-realistic condition.

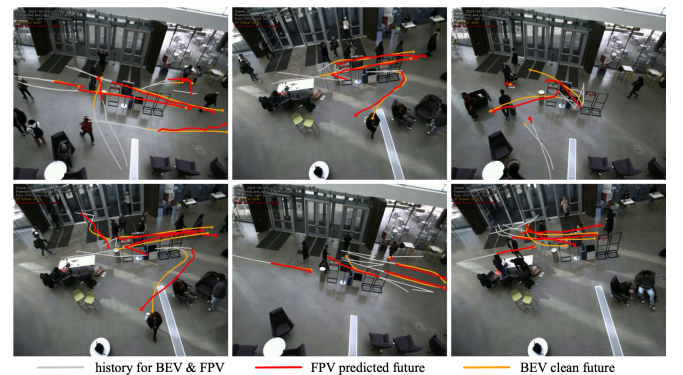


Fig. 4: Visualization under ego-view noise. **Gray**: past trajectories (BEV and noisy FPV-derived histories). **Orange**: clean BEV futures (GT). **Red**: predicted futures.

## D. Qualitative Results

Fig. 4 provides qualitative examples on EgoTraj-TBD. Although the BEV history and the FPV-derived history (both shown in gray) can differ substantially due to ego-view perception noise, BiFlow predicts future trajectories (red) that closely follow the clean BEV ground truth futures (orange). The examples highlight that BiFlow maintains accurate and physically plausible motion under deployment-realistic history corruption, indicating improved robustness to noisy and incomplete observations.

## VI. CONCLUSION

We introduce EgoTraj-Bench, a new benchmark that pairs noisy first-person-view observations with human-verified metric-space ground truth, capturing authentic deployment-level perturbations. The benchmark highlights a critical gap in trajectory prediction: the disconnect between idealized BEV-based evaluation and real-world ego-centric perception noise. We further propose BiFlow, a dual-stream framework with the EgoAnchor mechanism for intent-aware prediction. Experiments on our EgoTraj-Bench verify the effectiveness of our designs and demonstrate clear advantages in noisy and resource-constrained settings.

While our approach shows competitive performance in the current setup, our benchmark relies on projecting FPV-derived tracks into metric space, which may introduce residual alignment error, and we do not yet quantify the impact of individual noise types. Generalization may also vary with camera geometry and sensing characteristics. Future work will improve calibration-robust projection, noise-stratified evaluation, and adaptation to diverse platforms and environments. We hope that EgoTraj-Bench and BiFlow will support the community with models more robust to the ego-centric observation challenges and progress toward reliable real-world deployment.

## REFERENCES

- [1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *CVPR*, 2016.
- [2] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.
- [3] L. Sun, Z. Yan, S. M. Mellado, M. Hanheide, and T. Duckett, "3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data," in *ICRA*, 2018.
- [4] H. Tonoki, A. Yorozu, and M. Takahashi, "Model-based pedestrian trajectory prediction using environmental sensor for mobile robots navigation," *IJACSA*, 2017.
- [5] K. Li, M. Shan, K. Narula, S. Worrall, and E. Nebot, "Socially aware crowd navigation with multimodal pedestrian trajectory prediction for autonomous vehicles," in *ITSC*, 2020.
- [6] K. Chen, H. Zhu, D. Tang, and K. Zheng, "Future pedestrian location prediction in first-person videos for autonomous vehicles and social robots," *Image and Vision Computing*, 2023.
- [7] S. Poddar, C. Mavrogiannis, and S. S. Srinivasa, "From crowd motion prediction to robot navigation in crowds," in *IROS*, 2023.
- [8] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *CVPR*, 2018, pp. 7593–7602.
- [9] A. Rasouli, "A novel benchmarking paradigm and a scale-and motion-aware model for egocentric pedestrian trajectory prediction," in *ICRA*, 2024.

- [10] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectory++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *ECCV*, 2020.
- [11] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," *RAL*, 2022.
- [12] H. Bi, R. Zhang, T. Mao, Z. Deng, and Z. Wang, "How can i see my future? fvtraj: Using first-person view for pedestrian trajectory prediction," in *ECCV*, 2020.
- [13] B. Stoler, M. Jana, S. Hwang, and J. Oh, "T2fpv: Dataset and method for correcting first-person view errors in pedestrian trajectory prediction," in *IROS*, 2023.
- [14] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, 2007.
- [15] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *ECCV*, 2010.
- [16] A. Wang, D. Sato, Y. Corzo, S. Simkin, A. Biswas, and A. Steinfeld, "Tbd pedestrian data collection: Towards rich, portable, and large-scale natural pedestrian data," in *ICRA*, 2024.
- [17] A. Bertugli, S. Calderara, P. Coscia, L. Ballan, and R. Cucchiara, "Acvrnn: Attentive conditional-vrnn for multi-future trajectory prediction," *Computer Vision and Image Understanding*, 2021.
- [18] L. Shi, L. Wang, S. Zhou, and G. Hua, "Trajectory unified transformer for pedestrian trajectory prediction," in *ICCV*, 2023.
- [19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.
- [20] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.
- [21] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu, "Stochastic trajectory prediction via motion indeterminacy diffusion," in *CVPR*, 2022.
- [22] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog diffusion model for stochastic trajectory prediction," in *CVPR*, 2023.
- [23] Y. Fu, Q. Yan, L. Wang, K. Li, and R. Liao, "Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation," in *CVPR*, 2025.
- [24] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *ICCV*, 2021.
- [25] Y. Liu, R. Yu, S. Zheng, E. Zhan, and Y. Yue, "Naomi: Non-autoregressive multiresolution sequence imputation," *NeurIPS*, 2019.
- [26] M. Qi, J. Qin, Y. Wu, and Y. Yang, "Imitative non-autoregressive modeling for trajectory forecasting and imputation," in *CVPR*, 2020.
- [27] P. S. Chib, A. Nath, P. Kabra, I. Gupta, and P. Singh, "Ms-tp: imputation aware pedestrian trajectory prediction," in *ICML*, 2024.
- [28] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.
- [29] R. Martin-Martin, M. Patel, H. Rezaatfighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, "Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments," *TPAMI*, 2021.
- [30] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret, "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection," in *CVPR*, 2018.
- [31] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal, "Thör: Human-robot navigation data collection and accurate motion trajectories dataset," *RAL*, 2020.
- [32] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.
- [33] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.
- [34] X. Weng, B. Ivanovic, K. Kitani, and M. Pavone, "Whose track is it anyway? improving robustness to tracking errors with affinity-based trajectory prediction," in *CVPR*, 2022.
- [35] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory prediction in crowded scenes," in *ECCV*, 2017.
- [36] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *NeurIPS*, 2015.
- [37] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.