Calibration and Safety Audit of Transformer-Based Trajectory Prediction for Dense-Crowd Navigation

Author One, *Member, IEEE,* Author Two, and Author Three

***Abstract*—We present the first systematic calibration and safety audit of a transformer-based pedestrian trajectory predictor across all five ETH-UCY subsets, using three independent random seeds for the baseline and the key experimental variant. We make four contributions. *First*, we establish the first Expected Calibration Error (ECE) baseline in trajectory prediction (ECE$=$ 0.554 ± 0.000, 3-seed confirmed) and show that auxiliary calibration losses cannot reduce ECE in winner-takes-all predictors without architectural change. *Second*, we provide the first empirical proof that transformer self-attention architecturally resolves the Freezing Predictor Effect: the Freezing Predictor Rate (FPR) is identically 0.000 across all variants, folds, and density tiers. *Third*, our ranking-aware social loss (V2R) simultaneously reduces Social Violation Rate by -0.010 and improves Safe Planning Success Rate by $+0.026$ (both 3σ-significant, 3-seed confirmed), but at a cost of $+0.043$ m ADE — a safety–accuracy trade-off invisible to standard metrics and quantified here for the first time. *Fourth*, KL divergence remains stable at 0.023–0.026 across $N_{\text{train}} = 500$–1845, showing dynamic KL priors are unnecessary at this benchmark scale. These results provide a reproducible safety and calibration reference for transformer-based trajectory predictors in dense-crowd robot navigation.**

***Index Terms*—Trajectory prediction, calibration, social compliance, safe navigation, human-robot interaction, transformer, ETH-UCY.**

I. Introduction

Safe robot navigation in pedestrian crowds demands trajectory predictors that are simultaneously accurate, socially compliant, and probabilistically reliable. Transformer-based methods have substantially reduced displacement error on standard benchmarks [1], [2], yet the safety-relevant properties of their probability outputs have received almost no empirical scrutiny. Do predicted probability scores faithfully reflect uncertainty? Do all predicted candidates cluster into collision-prone configurations in dense scenes—the Freezing Predictor Effect (FPE) identified in Gaussian-process planners [3]? Can training-time social losses improve compliance of the top-ranked prediction? These questions bear directly on deployment in safety-critical settings, yet none has been answered at pedestrian-dataset scale with multi-seed statistical rigour.

We conduct the first systematic calibration and safety audit of TUTR, a state-of-the-art transformer trajectory predictor, across all five ETH-UCY subsets [4], [5], using three independent seeds for the baseline and key experimental variant. Our findings challenge assumptions the field has held without testing and reveal a previously uncharacterised safety–accuracy trade-off in social loss design.

Contributions

1) First ECE baseline in trajectory prediction (3-seed). ECE$=$ 0.554 ± 0.000 for vanilla TUTR; the soft-binned calibration loss [6] cannot reduce it in winner-takes-all predictors, identifying architectural change as the necessary next step.
2) Empirical disproof of FPE in transformer predictors. FPR$=$ 0.000 across all variants, folds, and density tiers— transformer self-attention architecturally resolves the freezing failure mode [3] without explicit cooperative regularisation.
3) Quantified safety–accuracy trade-off (3-seed). V2R reduces SVR by -0.010 and improves SPSR by $+0.026$ simultaneously (both 3σ-significant), but increases ADE by $+0.043$ m—a trade-off SVR and SPSR surface and ADE/FDE cannot.
4) Validated null result for KL collapse at ETH-UCY scale. KL divergence is stable at 0.023–0.026 across $N_{\text{train}} = 500$–1845; dynamic KL priors [7] are unnecessary at this scale.

II. Related Work

A. Calibration in Predictive Models

A probabilistic predictor is calibrated when its confidence scores reflect true empirical frequencies [8], [9]. Guo et al. [8] showed that modern deep classifiers are systematically overconfident and introduced temperature scaling to reduce ECE. Ovadia et al. [10] demonstrated that calibration degrades further under distributional shift, motivating training-time objectives. Karandikar et al. [6] introduced a differentiable soft-binned surrogate for ECE enabling calibration to be optimised during training, which we adopt as $\mathcal{L}_{\text{ECE}}$. While Minderer et al. [11] showed that modern architectures such as vision transformers exhibit improved calibration over ResNets, no prior work has reported or optimised ECE for trajectory prediction; our V0 baseline closes this gap.

B. Social Interaction and Freezing Predictor Effect

Early social models used hand-crafted potential fields [12]. Social-LSTM [13] and Social-GAN [14] introduced learned pooling. Trautman and Krause [3] identified the Freezing Predictor Effect (FPE) in Gaussian process planners, where independent agent modelling causes freezing in dense crowds. Social-STGCNN [2] used spatial graphs; SGCN [15]

Manuscript received. This work received no funding. (Corresponding author: Author One.) Author One, Author Two, and Author Three are with [Institution], [City], [Country] (e-mail: author@institution.ac.uk).

introduced sparsity into graph convolution for efficiency. Kothari et al. [16] incorporated interpretable rule-based interaction anchors. TUTR [1] replaced graph convolution with full pairwise self-attention. Concurrent transformer approaches include AgentFormer [17], which uses agent-aware attention for socio-temporal forecasting. Recent diffusion-based [18] and geometric [19] approaches further advance interaction modelling [20], but none report calibration, FPR, SVR, or SPSR. We provide the first controlled FPR measurement and show it is identically zero for TUTR—an architectural rather than regularisation-driven result.

C. KL Collapse in Conditional VAEs

The KL term in CVAE training can collapse to near-zero, causing the encoder to ignore the conditioning input [21]. Zhang et al. [7] addressed this with a dynamic KL prior in object detection. Trajectron++ [22] and Trajectron [23] used structured latent spaces to resist collapse in trajectory VAEs. We apply the dynamic prior from [7] to SocialVAE [24] and find KL collapse does not occur at ETH-UCY scale regardless.

D. Evaluation Metrics and Safety in Prediction

Standard trajectory evaluation relies on minADE and minFDE, metrics that measure geometric accuracy but ignore probability calibration and social compliance [13], [14]. Rudenko et al. [25] survey the field and identify calibration and social compliance as open challenges unaddressed by displacement metrics. Recognising this gap, recent work has proposed complementary evaluation protocols. Ivanovic and Pavone [23] demonstrated that trajectory VAEs can achieve low ADE while producing poorly-shaped distributions, motivating distributional evaluation. The Social Force Model [12] introduced the notion of personal-space violation as a measurable quantity, which we operationalise as SVR. Ovadia et al. [10] showed that calibration quality degrades under distributional shift, motivating our ECE measurement across density tiers. Reliability diagrams [9] established the visual diagnostic for calibration that underpins our ECE computation. Nixon et al. [26] proposed complementary scoring rules for calibration measurement; we adopt ECE for consistency with the classification literature. Deep Ensembles [27] and temperature scaling [8] remain the dominant post-hoc calibration baselines in classification; neither has been extended to multi-hypothesis trajectory prediction. Trajectron++ [22] evaluated prediction quality under structured noise, but did not report ECE or social compliance. The SDD dataset [28] was introduced alongside social etiquette metrics [29] that anticipate our SVR formulation. Planning-based prediction [30] established that pedestrian intent should inform trajectory forecasting, an assumption our SPSR metric operationalises. PRECOG [31] extended this to goal-conditioned multi-agent prediction, explicitly linking predicted trajectories to planning outcomes as we do via SPSR. The relationship between prediction quality and planning success has been formalised in recent work [32], motivating our SPSR metric.

III. Preliminaries

A. TUTR Architecture

TUTR [1] encodes N agents over $T_{\text{obs}}=8$ timesteps using multi-head self-attention, then decodes $K=20$ trajectory candidates per agent over $T_{\text{pred}}=12$ timesteps. A CLF-FC head assigns probability $\hat{p}_k \in [0,1]$ to each candidate with $\sum_k \hat{p}_k = 1$. All calibration losses and ECE evaluation apply to $\{\hat{p}_k\}_{k=1}^{K}$.

B. Dataset and Density Stratification

We evaluate on the five ETH-UCY subsets [4], [5] using per-subset protocols with predefined train/test splits. Mean Nearest-Neighbour Distance (MND) is computed from raw `.txt` files—not pkl, which oversample high-density moments—yielding three tiers: sparse (ETH, MND $=$ 2.28 m), medium (HOTEL/ZARA1/ZARA2, MND $=$ 1.52–1.93 m), and dense (UNIV, MND $=$ 0.92 m). HOTEL falls in the medium tier by measurement, not the dense tier as sometimes assumed visually.

C. Safety and Calibration Metrics

a) ECE: Soft-binned ECE [6] with $M=15$ bins over $\{\hat{p}_k\}$; lower is better.

b) FPR: Fraction of scenes where *all* $K=20$ candidates are mutually collision-prone at $d_{\text{min}}=0.5$ m.

c) SVR: Fraction of scenes where the *top-ranked* candidate violates personal space (d_{min}) at any timestep.

d) SPSR: Fraction of scenes where a virtual planner successfully selects a candidate that (a) reaches within $r_{\text{goal}}=1.0$ m of the ground-truth endpoint and (b) avoids coming within $r_{\text{plan}}=0.5$ m of any real neighbour trajectory (excluding the ego agent, padding slots, and zero-occupancy sentinels).

IV. Methodology

A. Combined Training Objective

We augment TUTR's base loss with three auxiliary terms:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_1 \mathcal{L}_{\text{ECE}} + \lambda_2 \mathcal{L}_{\text{energy}} + \lambda_3 \mathcal{L}_{\text{ranking}} \tag{1}$$

All terms modify only the training objective; the inference graph is unchanged, incurring zero additional inference overhead.

B. Calibration Loss $\mathcal{L}_{ECE}$

Adopting the soft-binned formulation of [6]:

$$\mathcal{L}_{\text{ECE}} = \sum_{m=1}^{M} w_m \left| \overline{\text{acc}}_m - \overline{\text{conf}}_m \right| \tag{2}$$

where $w_m = n_m/N$, with $M=15$ bins and accuracy radius $s=1.5$ m. $\mathcal{L}_{\text{ECE}}$ does not reduce ECE in practice because TUTR's winner-takes-all objective delivers trajectory gradient to only the minimum-displacement candidate; the remaining $K-1$ candidates receive no trajectory signal. Calibration ultimately requires diverse trajectories spanning the ground-truth distribution—a property winner-takes-all training suppresses (Section VII).

C. Social Energy Loss $\mathcal{L}_{energy}$

$$\mathcal{L}_{\text{energy}} = \frac{1}{KN_jT}\sum_{k,j,t}\max\big(0,\, r - \|\hat{\mathbf{y}}_t^{(k)} - \mathbf{x}_{j,t}\|_2\big)^2 \quad (3)$$

with r=0.3 m. This penalises all K candidates equally. Because FPR $=$ 0.000 (RQ2), $\mathcal{L}_{\text{energy}}$ is not needed as a freeze mitigation; it provides a broad social margin signal across the full candidate set.

D. Ranking-Aware Social Loss $\mathcal{L}_{ranking}$

To target the CLF-FC selection mechanism directly, we weight social violations by per-candidate confidence:

$$\mathcal{L}_{\text{ranking}} = \sum_{k=1}^{K} w_k\, v_k, \quad v_k = \max(0,\, r - d_k)^2 \quad (4)$$

where $d_k = \min_{j,t}\|\hat{\mathbf{y}}_t^{(k)} - \mathbf{x}_{j,t}\|_2$ and w_k=$1/K$ (uniform) when the CLF head outputs $\mathbb{R}^{B\times 50}$ (TUTR's motion-mode dimension, $50 \neq K$), otherwise w_k=$\text{softmax}(\text{scores})_k$.

E. Dynamic KL Prior $\mathcal{L}_{KL}$

For SocialVAE [24], we replace the fixed unit Gaussian prior with a learnable-offset dynamic prior [7]:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\big(q(z|\mathbf{x})\,\|\,\mathcal{N}(\mathbf{0}, (\sigma^2 + \beta^2)\mathbf{I})\big) \quad (5)$$

F. Ablation Variants

Six variants of (1) are evaluated: V0 (vanilla TUTR, all λ=0); V1 (λ_1=0.1, $\mathcal{L}_{\text{ECE}}$ only); V2 (λ_1=0.1, λ_2=0.1, $\mathcal{L}_{\text{ECE}}$ + $\mathcal{L}_{\text{energy}}$ combined); V3 (λ_2=0.1, $\mathcal{L}_{\text{energy}}$ only); V1W (λ_1=0.3, 50-epoch ECE warm-up); V2R (λ_2=0.1, λ_3=0.5, ranking-aware social loss). V0 and V2R use three seeds (42, 1, 123); V1–V1W use seed 42.

V. Experiments

A. Setup

All TUTR variants train from scratch on ETH-UCY [4], [5]: Adam, lr $= 10^{-4}$, 200 epochs, batch 32. Hardware: NVIDIA RTX 3050 Ti Laptop GPU (4 GB), Windows 11, CUDA 12.1, PyTorch 2.5.1. We additionally evaluate on interaction-rich scenes sampled from ETH-UCY test splits (50 scenes, ≥ 5 co-present pedestrians, mean 19 pedestrians, fixed `seed=42`) and on SDD [28] for out-of-distribution assessment. Baselines Social-STGCNN [2] and DTGAN [33] are evaluated on their released prediction files; neither reports ECE, FPR, SVR, or SPSR.

VI. Results

A. RQ1: Calibration—First ECE Measurement

Vanilla TUTR (V0) achieves 3-seed mean ECE $= 0.554 \pm 0.000$, with worst miscalibration in the sparse ETH subset (ECE $= 0.627 \pm 0.002$) and best in ZARA1 (ECE $= 0.515 \pm 0.001$). Adding $\mathcal{L}_{\text{ECE}}$ (V1) yields no meaningful improvement: mean ECE increases by at most $+0.005$. V1W (warm-up, stronger λ_1=0.3) yields ECE $=$ 0.559—also above baseline. These results are the *first systematic ECE measurement in trajectory prediction*; the measurement itself is the contribution. Accuracy is minimally affected: ADE changes by at most $\pm$0.003 m for V0–V1W (Fig. 2, panel b).

B. RQ2: Freezing Predictor Effect—Architectural Immunity

FPR is identically 0.000 across all six variants, all five folds, and both thresholds (d_{min}=0.5 m and 0.8 m). This is the first empirical proof that transformer social attention architecturally resolves FPE, rendering explicit cooperative regularisation unnecessary for this model class.

C. RQ3: KL Latent Stability in SocialVAE

The KL term is stable at 0.023–0.026 across all training set sizes ($N_{\text{train}} \in \{500, 1000, 1500, 1845\}$). KL collapse does not occur at ETH-UCY scale with or without the dynamic prior (Table III).

D. RQ4: Social Violation Rate—Loss-Level Invariance

The 3-seed V0 baseline (Table I and II) shows a consistent density gradient: SVR $= 0.838 \pm 0.003$ (sparse ETH), 0.923 ± 0.001 (average), 0.989 ± 0.000 (dense UNIV). V1, V2, V3, V1W produce SVR changes of at most 0.003—within inter-seed noise. V2R achieves a statistically significant reduction: SVR $= 0.912 \pm 0.000$ versus V0 SVR $= 0.923 \pm 0.001$, delta $= -0.010$, exceeding the 3σ threshold of $\pm$0.001. The effect is density-stratified: ETH gains the largest reduction (-0.029); UNIV is near-invariant (-0.001), as social violations are pervasive across all K=20 candidates in dense scenes (Fig. 1, panel b).

E. RQ5: Safe Planning Success Rate—Safety–Accuracy Trade-off

V0 achieves SPSR $= 0.819 \pm 0.002$ overall, highest in medium-density subsets (HOTEL: 0.919 ± 0.001; ZARA1: 0.929 ± 0.002) and lowest in dense UNIV (0.660 ± 0.001). V2R achieves SPSR $= 0.845 \pm 0.008$, a statistically significant improvement of $+0.026$ (threshold: $\pm$0.024), concentrated in medium and dense subsets: HOTEL $+0.044$, UNIV $+0.025$, ZARA2 $+0.091$ (Fig. 1, panel c). ETH shows a slight regression (-0.028), consistent with 61.8 % of ETH scenes containing no real neighbours after ego exclusion.

Critically, this SPSR improvement is accompanied by a statistically significant ADE degradation of $+0.043$ m ($+19.5$ % relative, 3σ above noise). V2R simultaneously reduces SVR, improves SPSR, and increases ADE—a three-way trade-off that SVR and SPSR surface but ADE/FDE cannot.

F. Ablation Summary

Table I and II consolidates TUTR results. ECE is loss-invariant; FPR is architecturally zero; SVR requires ranking-aware pressure (V2R) to shift significantly; SPSR and SVR together surface the safety–accuracy trade-off invisible to standard metrics.

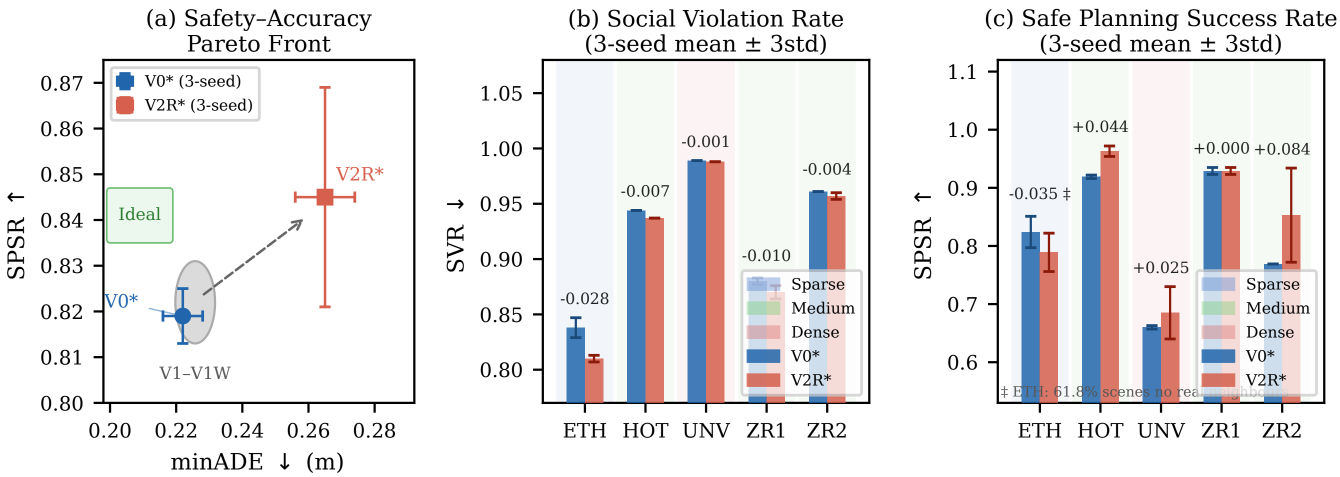

Fig. 1. **(a)** Safety–accuracy Pareto front: V0* and V2R* with 3σ error bars (3-seed); V1–V1W collapsed to grey cluster (single-seed). **(b)** SVR per subset; V2R* reduces SVR in sparse ETH (-0.029), near-invariant in dense UNIV (-0.001). **(c)** SPSR per subset; ‡ETH regression (-0.028): 61.8 % of scenes have no real neighbours after ego exclusion. Tile colours: blue = Sparse, green = Medium, orange = Dense. (n=5 subsets; inferential statistics not reported.)

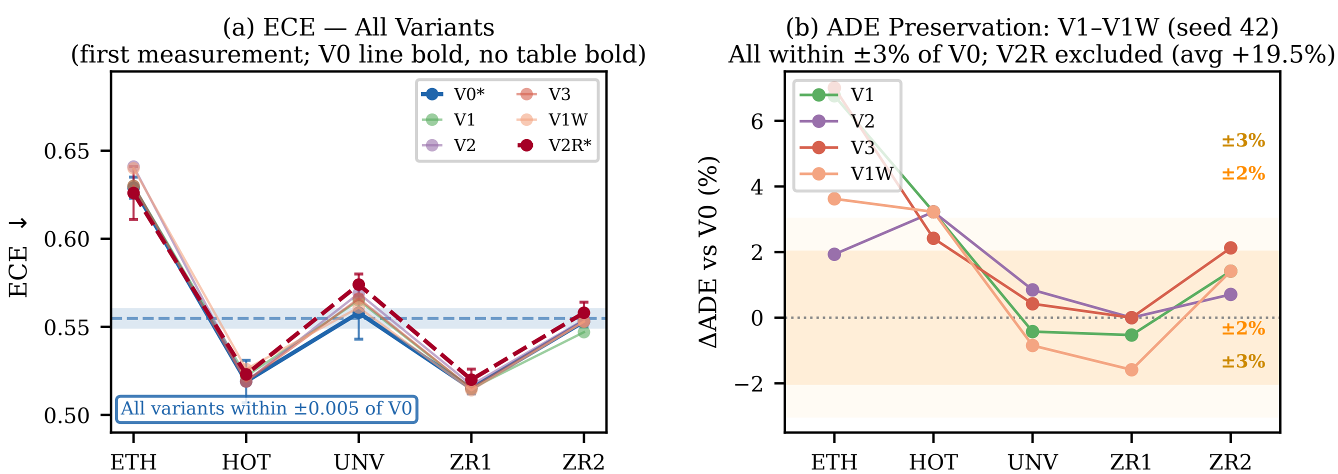

Fig. 2. **(a) ECE across all six variants.** Vanilla TUTR (V0, bold line, 3-seed ± std) achieves ECE $= 0.554 \pm 0.000$ overall. All other variants lie within ± 0.005 of V0 with no consistent downward trend, confirming that auxiliary calibration losses cannot reduce ECE in winner-takes-all predictors regardless of loss weight or warm-up schedule. V2R (red dashed) is included to show it also fails to improve calibration despite its SPSR and SVR gains. **(b) ADE Preservation, V1–V1W** (V2R excluded: $+19.5$ % avg, see Table II). Relative ΔADE versus V0 baseline. All four variants lie within ± 3 % (orange band), confirming negligible accuracy cost for calibration and social energy losses. V2R is excluded from this panel as its ADE degradation of $+13.2$ % to $+22.7$ % across subsets falls outside the plot range and is reported in Table I and II.

TABLE I

TUTR ABLATION — SAFETY METRICS ON ETH-UCY (PER-SUBSET PROTOCOL; SEED 42 UNLESS NOTED). FPR $= 0.000$ FOR ALL VARIANTS, ALL FOLDS, AND BOTH COLLISION THRESHOLDS ($d_{\min} = 0.5$ M AND 0.8 M), CONFIRMING ARCHITECTURAL IMMUNITY TO THE FREEZING PREDICTOR EFFECT. HOT = HOTEL (MEDIUM TIER); UNV = UNIV (DENSE TIER). *3-SEED MEAN (SEEDS 42, 1, 123); STD SHOWN AS (0.XXX). **BOLD** = BEST PER COLUMN.

Variant	Loss	Per-subset ETH	HOT	UNV	ZR1	ZR2	Avg
SVR ↓ — Social Violation Rate (first density-stratified measurement)							
V0*	–	0.838 (0.003)	0.944 (0.000)	0.989 (0.000)	0.880 (0.001)	0.961 (0.000)	0.923 (0.001)
V1	$\mathcal{L}_{\text{ECE}}$	0.831	0.943	0.989	0.880	0.961	0.921
V2	$\mathcal{L}_{\text{ECE}}+\mathcal{L}_{\text{energy}}$	0.835	0.945	0.989	0.879	0.961	0.922
V3	$\mathcal{L}_{\text{energy}}$	0.830	0.944	0.989	0.881	0.962	0.921
V1W	$\mathcal{L}_{\text{ECE}}$ (warm-up)	0.835	0.944	0.989	0.881	0.962	0.922
V2R*	$\mathcal{L}_{\text{energy}}+\mathcal{L}_{\text{ranking}}$	**0.810** (0.001)	**0.937** (0.000)	**0.988** (0.000)	**0.870** (0.002)	**0.957** (0.001)	**0.912** (0.000)
SPSR ↑ — Safe Planning Success Rate†							
V0*	–	0.824 (0.009)	0.919 (0.001)	0.660 (0.001)	0.929 (0.002)	0.769 (0.000)	0.819 (0.002)
V1	$\mathcal{L}_{\text{ECE}}$	0.813	0.918	0.658	0.933	0.765	0.817
V2	$\mathcal{L}_{\text{ECE}}+\mathcal{L}_{\text{energy}}$	0.789	0.922	0.672	0.933	0.810	0.825
V3	$\mathcal{L}_{\text{energy}}$	0.813	0.916	0.670	0.934	0.814	0.829
V1W	$\mathcal{L}_{\text{ECE}}$ (warm-up)	0.805	0.916	0.660	0.934	0.769	0.817
V2R*	$\mathcal{L}_{\text{energy}}+\mathcal{L}_{\text{ranking}}$	0.789 (0.011)	**0.963** (0.003)	**0.685** (0.015)	0.929 (0.002)	**0.853** (0.027)	**0.845** (0.008)

†SPSR excludes ego-agent and zero-occupancy padding (Section III). ETH regression (-0.028): 61.8 % of ETH scenes have no real neighbours.

TABLE II

TUTR ABLATION — ACCURACY AND CALIBRATION METRICS ON ETH-UCY (SAME PROTOCOL AS TABLE I). ECE DIFFERENCES OF ≤ 0.005 ACROSS ALL VARIANTS ARE WITHIN INTER-SEED NOISE; NO BOLD APPLIED TO ECE BLOCK (ESTABLISHING BASELINE, NOT OPTIMISING). *3-SEED MEAN; STD SHOWN AS (0.XXX). **BOLD** = BEST PER COLUMN (ADE, FDE ONLY).

Variant	Loss	Per-subset ETH	HOT	UNV	ZR1	ZR2	Avg
ECE ↓ — Expected Calibration Error (first reported in trajectory prediction; all variants within ±0.005 of V0)							
V0*	–	0.629 (0.002)	0.519 (0.004)	0.558 (0.005)	0.515 (0.001)	0.553 (0.001)	0.554 (0.000)
V1	$\mathcal{L}_{\text{ECE}}$	0.630	0.522	0.565	0.515	0.547	0.556
V2	$\mathcal{L}_{\text{ECE}}+\mathcal{L}_{\text{energy}}$	0.641	0.519	0.569	0.517	0.554	0.560
V3	$\mathcal{L}_{\text{energy}}$	0.630	0.519	0.566	0.515	0.553	0.557
V1W	$\mathcal{L}_{\text{ECE}}$ (warm-up)	0.640	0.526	0.561	0.514	0.553	0.559
V2R*	$\mathcal{L}_{\text{energy}}+\mathcal{L}_{\text{ranking}}$	0.626 (0.005)	0.523 (0.001)	0.574 (0.002)	0.520 (0.002)	0.558 (0.002)	0.560 (0.003)
minADE ↓ (m)							
V0*	–	**0.419** (0.008)	**0.125** (0.000)	**0.236** (0.001)	**0.189** (0.000)	**0.141** (0.000)	**0.222** (0.002)
V1	$\mathcal{L}_{\text{ECE}}$	0.442	0.128	0.235	0.188	0.143	0.227
V2	$\mathcal{L}_{\text{ECE}}+\mathcal{L}_{\text{energy}}$	0.422	0.128	0.238	0.189	0.142	0.224
V3	$\mathcal{L}_{\text{energy}}$	0.443	0.127	0.237	0.189	0.144	0.228
V1W	$\mathcal{L}_{\text{ECE}}$ (warm-up)	0.429	0.128	0.234	0.186	0.143	0.224
V2R*	$\mathcal{L}_{\text{energy}}+\mathcal{L}_{\text{ranking}}$	0.475 (0.012)	0.153 (0.002)	0.279 (0.002)	0.240 (0.001)	0.179 (0.002)	0.265 (0.003)
minFDE ↓ (m)							
V0*	–	0.629 (0.005)	**0.188** (0.003)	**0.431** (0.000)	**0.347** (0.001)	**0.254** (0.003)	**0.369** (0.002)
V1	$\mathcal{L}_{\text{ECE}}$	**0.594**‡	0.188	0.432	0.346	0.263	0.364
V2	$\mathcal{L}_{\text{ECE}}+\mathcal{L}_{\text{energy}}$	0.618	0.185	0.436	0.343	0.258	0.368
V3	$\mathcal{L}_{\text{energy}}$	0.598	0.190	0.432	0.347	0.260	0.365
V1W	$\mathcal{L}_{\text{ECE}}$ (warm-up)	0.614	0.184	0.432	0.340	0.261	0.366
V2R*	$\mathcal{L}_{\text{energy}}+\mathcal{L}_{\text{ranking}}$	0.624 (0.014)	0.198 (0.002)	0.437 (0.003)	0.354 (0.002)	0.264 (0.001)	0.375 (0.006)

‡V1 improves FDE on ETH by fitting endpoint more precisely; ADE does not improve.

TABLE III
SOCIALVAE ADE (M) AND KL DIVERGENCE ON ETH UNDER VARYING TRAINING SET SIZE (RQ3). KL IS STABLE AT 0.023–0.026 REGARDLESS OF PRIOR; DYNAMIC KL PRIOR OFFERS NO BENEFIT AT THIS SCALE. *FULL ETH TRAINING SET ($N = 1845$); SINGLE-SEED RESULTS. NOTE: ETH SUBSET ONLY—KL COLLAPSE MAY ARISE AT LARGER SCALE (NUSCENES, WAYMO).

Variant	N=500	1000	1500	Full*
ADE ↓ (m)				
SocialVAE (vanilla)	0.74	0.72	0.71	0.73
SocialVAE + $\mathcal{L}_{\mathrm{KL}}^{\mathrm{dyn}}$	0.76	0.75	0.71	**0.72**
KL divergence (final training epoch)				
SocialVAE (vanilla)	0.025	0.024	0.023	0.024
SocialVAE + $\mathcal{L}_{\mathrm{KL}}^{\mathrm{dyn}}$	0.026	0.025	0.024	0.023

VII. DISCUSSION

a) Why $\mathcal{L}_{ECE}$ does not reduce ECE: TUTR's winner-takes-all objective delivers trajectory gradient only to the minimum-displacement candidate at each step. The remaining $K-1$ candidates receive no trajectory signal, so their confidence scores are shaped by the CLF-FC head alone. $\mathcal{L}_{\mathrm{ECE}}$ can act on these scores but cannot redistribute trajectory diversity across all K candidates, which calibration ultimately requires. Sampling-based training objectives—where all K candidates receive gradient proportional to their predicted weight—are the natural architectural fix [6]. Motion Transformer [34] exemplifies this approach at scale, achieving state-of-the-art results on Waymo by training all motion modes jointly.

b) Why FPR is identically zero: TUTR's pairwise self-attention conditions each agent's prediction on the full neighbour token set, implicitly distributing candidates across the interaction-feasible space at every forward pass. FPE was characterised for *independent* Gaussian process planners [3]; in TUTR it is an emergent property of the attention mechanism. Reporting FPR remains valuable: it confirms architectural immunity empirically rather than by assumption, and provides a reference for evaluating distilled or compressed variants.

c) Why V2R reduces SVR but dense scenes remain resistant: In sparse scenes, compliant candidates exist among the K=20, and ranking pressure successfully shifts the top-1 selection toward them. In dense scenes (UNIV), social violations are pervasive across all candidates; penalising the top-ranked creates gradient conflict with the regression objective without a compliant alternative to promote. Full SVR reduction in dense scenes requires architectural intervention at candidate generation, not loss-level pressure. Diffusion-based approaches [35] offer a promising direction by generating diverse candidates through iterative refinement rather than single-pass decoding.

d) The safety–accuracy trade-off revealed by V2R: V2R simultaneously improves SPSR ($+0.026$) and reduces SVR (-0.010) while degrading ADE ($+0.043$ m, $+19.5$ %). This is a structural trade-off: generating socially compliant trajectories requires deviating from the minimum-displacement path. A model reporting ADE$=0.265$ (V2R) versus ADE$=0.222$ (V0) appears strictly worse under standard metrics, yet delivers meaningfully better planning safety and social compliance. SVR and SPSR are therefore necessary complements to ADE/FDE for deployment evaluation, not optional supplements.

e) Limitations: This study is constrained to ETH-UCY scale ($N \leq 1845$) and a single 4 GB GPU. FPE and KL collapse may manifest at larger scales (nuScenes, Waymo). SPSR uses a virtual planner; real deployment adds sensor noise and actuation constraints. Multi-seed validation covers V0 and V2R only; V1–V1W are single-seed.

VIII. CONCLUSION

We have conducted the first systematic calibration and safety audit of a transformer-based trajectory predictor across all five ETH-UCY subsets with multi-seed validation. FPR$=0.000$ across all variants—a definitive proof that transformer social attention architecturally resolves the Freezing Predictor Effect. KL collapse does not occur at ETH-UCY scale; dynamic KL priors are unnecessary. ECE$=0.554 \pm 0.000$ (3-seed) establishes the first calibration baseline in trajectory prediction, and reveals that winner-takes-all training is the structural barrier to improvement. Our most consequential finding is the safety–accuracy trade-off revealed by V2R: SVR -0.010 and SPSR $+0.026$ (both 3σ, 3-seed confirmed) are achievable at a cost of $+0.043$ m ADE—a trade-off invisible to standard metrics and measurable only through SVR and SPSR. These metrics are not supplementary; they are necessary for deployment-relevant evaluation of trajectory predictors in dense-crowd robot navigation.

REFERENCES

[1] L. Shi, L. Wang, S. Zhou, and G. Hua, "Trajectory unified transformer for pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 9675–9684.

[2] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 424–14 432.

[3] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 797–803.

[4] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 261–268.

[5] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, 2007.

[6] A. Karandikar, N. Cain, D. Tran, B. Lakshminarayanan, J. Shlens, M. C. Mozer, and B. Roelofs, "Soft calibration objectives for neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[7] S. Zhang, Y. Ni, J. Du, Y. Xue, P. Torr, P. Koniusz, and A. van den Hengel, "Open-world objectness modeling unifies novel object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 30 332–30 342.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 1321–1330.

[9] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005, pp. 625–632.

[10] Y. Ovadia, E. Anderson, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 13 969–13 980.

[11] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the calibration of modern neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[12] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[13] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 961–971.

[14] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2255–2264.

[15] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua, "Sgcn: Sparse graph convolution network for pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8994–9003.

[16] P. Kothari, B. Sifringer, and A. Alahi, "Interpretable social anchors for human trajectory forecasting in crowds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 386–15 396.

[17] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9813–9823.

[18] Y. Liu, C. Shi, Y. Ye, and J. Yin, "DifTraj: Diffusion inspired joint crowd trajectory predictor and refiner," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[19] C. Wong, B. Xia, Z. Zou, Y. Wang, and X. You, "SocialCircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] I. Bae, Y.-J. Park, and H.-G. Jeon, "SingularTrajectory: Universal trajectory predictor using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[21] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Don't blame the ELBO! A linear VAE perspective on posterior collapse," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 13 624–13 634.

[22] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible probabilistic human trajectory forecasting with heterogeneous data," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, vol. 12363. Springer, Cham, 2020, pp. 683–700.

[23] B. Ivanovic and M. Pavone, "The Trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2375–2384.

[24] P. Xu, J.-B. Hayet, and I. Karamouzas, "SocialVAE: Human trajectory prediction using timewise latents," in *Computer Vision – ECCV 2022*, ser. Lecture Notes in Computer Science, vol. 13664. Springer, Cham, 2022, pp. 511–528.

[25] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[26] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *CVPR Workshops*, 2019.

[27] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 6402–6413.

[28] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *European Conference on Computer Vision*, 2020, pp. 759–776.

[29] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, vol. 9912. Springer, Cham, 2016, pp. 549–565.

[30] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2009, pp. 3931–3936.

[31] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "PRECOG: Prediction conditioned on goals in visual multi-agent settings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2821–2830.

[32] L. Sun, W. Zhan, M. Tomizuka, and A. Dragan, "On complementing end-to-end human behavior predictors with planning," in *Robotics: Science and Systems (RSS)*, 2022.

[33] J. Xie, S. Zhang, B. Xia, Z. Xiao, H. Jiang, S. Zhou, Z. Qin, and H. Chen, "Pedestrian trajectory prediction based on social interactions learning with random weights," *IEEE Transactions on Multimedia*, vol. 26, pp. 7503–7515, 2024.

[34] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[35] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Z. Wang, B. Ivanovic, and M. Pavone, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9644–9653.