PlayGen-MoG: Framework for Diverse Multi-Agent Play Generation via Mixture-of-Gaussians Trajectory Prediction

Supplementary Material

Comparison with Generative Baselines

We compare PlayGen-MoG against two standard generative approaches trained on the same American football tracking data: a conditional VAE (CVAE) with an MLP decoder, and a Leapfrog Diffusion Model (LED) with formation conditioning. For the CVAE, we applied β-warmup, target KL annealing, and a weak (MLP-only) decoder to mitigate posterior collapse. For LED, we followed the two-stage training procedure (denoiser then initializer) with $K{=}6$ trajectory proposals and 5 denoising steps.

Table 1. Comparison with generative baselines on American football tracking data (9,934 plays). ADE/FDE in yards. APD measures diversity across $K{=}10$ samples per formation. †LED reports best-of-6 proposals (standard LED evaluation); single-proposal ADE is substantially higher.

Method	ADE↓	FDE↓	APD↑	Params
CVAE (MLP decoder)	2.88	6.56	0.10	11M
LED†	**1.34**	**3.14**	—	3.4M
PlayGen-MoG (Ours)	1.68	3.98	**1.13**	1.3M

The CVAE achieves moderate ADE but suffers from severe posterior collapse: the learned KL divergence converges to near-zero (4.3×10^{-6}), and diverse latent samples decode to near-identical trajectories (APD $=$ 0.10 yards). Despite anti-collapse strategies, the decoder learns to ignore the latent code entirely.

LED achieves the lowest best-of-K ADE by selecting the closest of 6 proposals to the ground truth, but its individual proposals exhibit high variance with trajectories spanning unrealistic distances. As shown in Figure 1, LED samples lack the spatial coherence of real plays, with player trajectories crossing the entire field in an uncoordinated fashion. LED also does not provide a mechanism for enumerating distinct play concepts: diversity is an artifact of diffusion noise rather than structured variation.

PlayGen-MoG strikes a middle ground between these two failure modes. Its ADE of 1.68 yards is competitive with LED's best-of-K selection, indicating that generated routes are individually realistic and spatially coherent. At the same time, the mixture structure produces meaningful diversity (APD $=$ 1.13 yards): each of the 8 components learns a distinct play concept—short passes, deep routes, screens, etc.—rather than collapsing to a single output or producing random noise. This combination of low error *and* structured diversity is what makes MoG practical for play design: a coach can enumerate all 8 concepts from a formation and get realistic, coordinated routes for each one, rather than sifting through identical outputs (CVAE) or discarding incoherent ones (LED). PlayGen-MoG also achieves this with the smallest model (1.3M parameters), making it efficient to deploy in interactive tools.

Generation Quality Across Time Horizons

Table 2 evaluates PlayGen-MoG at truncated prediction horizons to characterize how generation quality evolves over time. All metrics are computed on the validation set at temperature 0.8.

Table 2. Generation quality at different prediction horizons. T is the number of predicted frames at 10 fps. ADE/FDE in yards. APD measures diversity across $K{=}10$ samples.

T	Duration	ADE↓	FDE↓	APD↑
10	1.0 s	0.22	0.51	0.08
20	2.0 s	0.63	1.50	0.31
30	3.0 s	0.96	2.28	0.66
40	4.0 s	1.13	2.66	1.17
49	4.9 s	1.19	2.80	1.69

ADE degrades gracefully from 0.22 yards at 1 s to 1.19 yards at 4.9 s, confirming that the non-autoregressive absolute-displacement architecture avoids cumulative drift even at longer horizons. FDE follows a similar trend. Notably, generative diversity (APD) increases with the prediction horizon: at $T{=}10$ all play concepts share similar early motion (formation release), while by $T{=}49$ the mixture components have diverged into distinct route patterns. This matches the intuition that play concepts become distinguishable only after players have had time to separate from the formation.

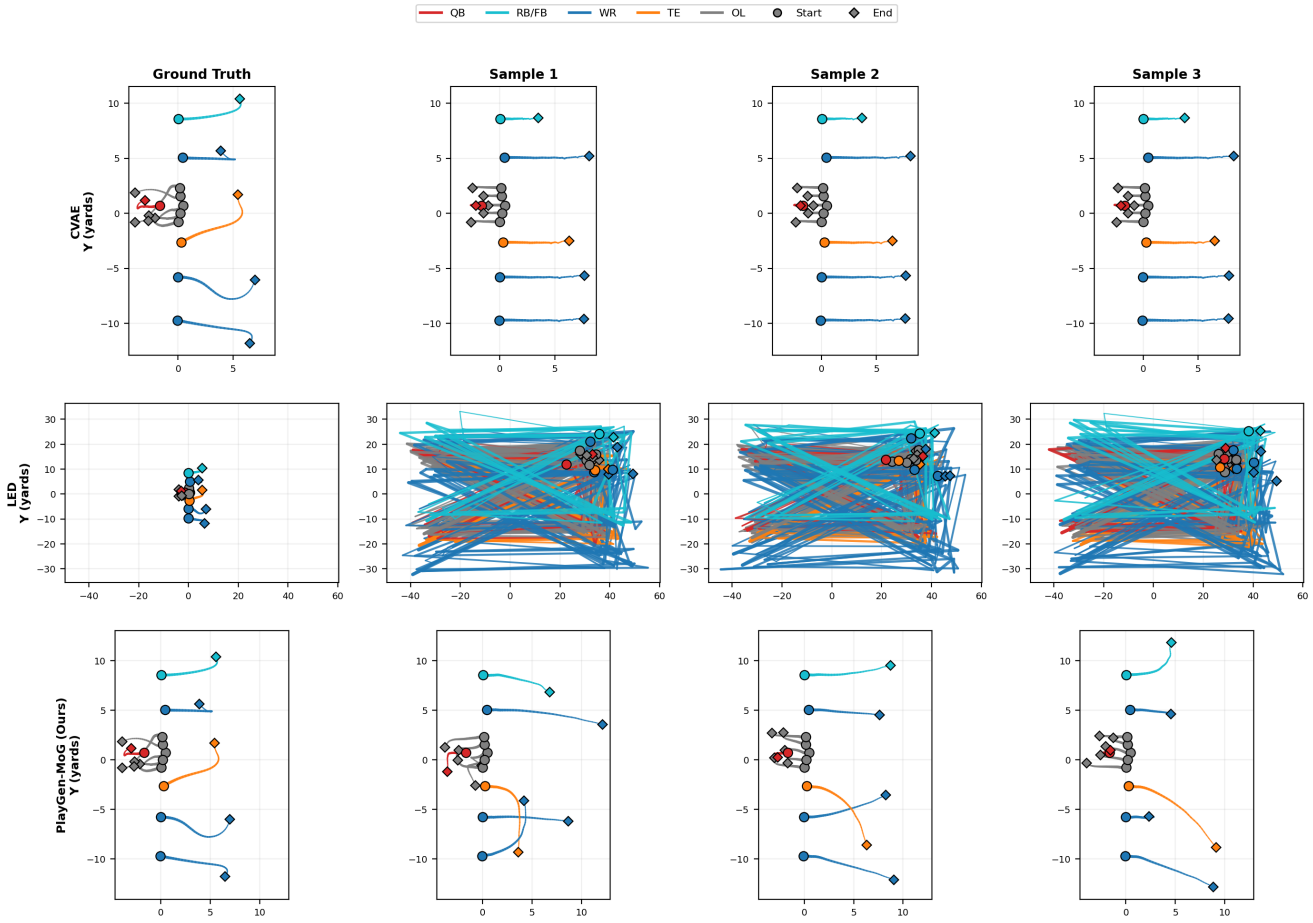

Figure 1. **Qualitative comparison of generative baselines.** Each row shows three independent samples from the same formation. **Top (CVAE):** Posterior collapse—all samples are nearly identical despite different latent draws. **Middle (LED):** Diffusion produces high-variance, spatially incoherent trajectories spanning the full field. **Bottom (PlayGen-MoG):** Each sample represents a distinct, realistic play concept with coordinated player motion.

Figure 2 visualizes a single generated play at each horizon. At T=10 (1 s), players have barely left the formation; by T=30 (3 s), distinct route shapes are emerging; and at T=49 (4.9 s), the full play concept is visible with receivers completing their routes downfield.

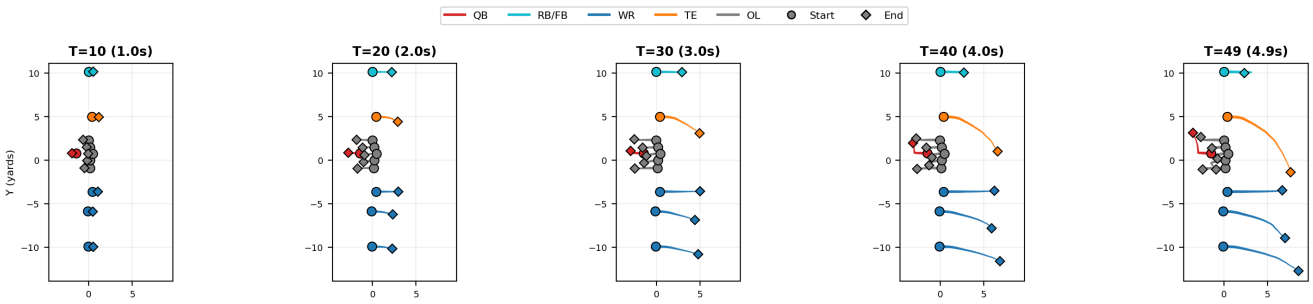

Figure 2. A single generated play shown at increasing prediction horizons. Circles mark starting positions; diamonds mark endpoints at each horizon. Trajectory width tapers to indicate direction of movement. Routes become progressively distinguishable as the horizon extends.