

1 Per-Class Performance Comparison on Human3.6M

We conduct extensive experiments on the Human3.6M dataset, which encompasses 13 diverse action categories including Directions, Discussion, Eating, Greeting, Phoning, Photo, Posing, Purchases, Sitting, SittingDown, Smoking, Waiting, Walking, WalkDog, and WalkTogether. Table 1 presents a detailed comparison of our method against several state-of-the-art approaches across all action categories. From Table 1, it can be observed that our method demonstrates strong performance in terms of both ADE and FDE across a wide range of action categories. In particular, our approach achieves significant improvements on two walking-related actions — Walking and WalkTogether — where precise motion prediction is crucial. This indicates the model’s enhanced ability to capture and forecast dynamic motion patterns, especially in scenarios involving coordinated or continuous movement.

2 Wavelet guided masked fusion module

The detailed structure of our proposed Wavelet-guided masked fusion module is shown in Fig 1. The module consists of a hierarchical arrangement of two distinct layer types. The lower section comprises N local fusion layers, each built with a sophisticated architecture incorporating CrossAttn, SelfAttn, and FFN components. Within these local fusion layers, the cross-attention mechanism facilitates information exchange between frequency domain features and masked vectors in the latent space, enabling effective multi-modal integration at a fine-grained level. The upper section contains N global context layers, constructed exclusively with self-attention mechanisms and feed-forward networks. These layers are specifically designed to consolidate global information, progressively refining representations to produce a condition vector z that encapsulates comprehensive contextual understanding. In our implementation, we set $N = 6$, creating a balanced architecture with sufficient capacity for both local feature fusion and global context integration.

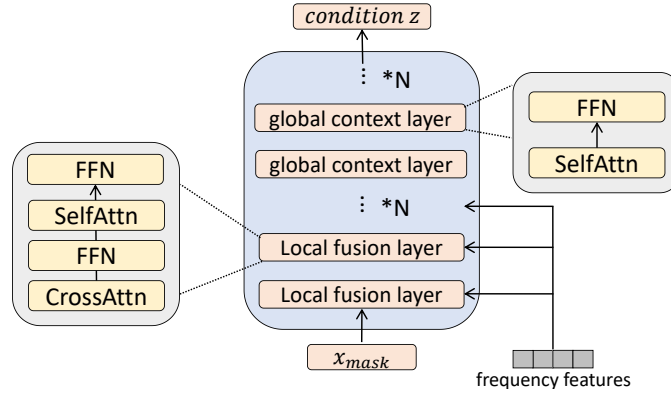


Figure 1: The detailed architecture of the Wavelet guided masked fusion module

3 More ablation study

In this part, we conduct additional ablation studies to further analyze the effectiveness of key components in our method. Specifically, we examine: (1) the impact of varying the number of layers in the masked fusion module, and (2) the design of the denoising steps within the diffusion process. For the first ablation study, we investigate how different configurations of local fusion layers and total layers affect model performance. As shown in Table 2, we systematically vary the number of local fusion layers (ranging from 4 to 12) while adjusting the total number of layers (ranging from 8 to 14). The results demonstrate that a proper configuration of these architectural components significantly impacts model performance. Specifically, the model achieves optimal performance on our task when using 6 local fusion layers within a total of 12 layers, yielding the best overall performance. For the second ablation study, we examine different noise schedule configurations on model performance and inference efficiency. Table 3 shows that reducing both noising steps during training and DDIM steps during inference to 10 maintains competitive performance while drastically

Classes		APD	ADE	FDE	MMADE	MMFDE	Classes		APD	ADE	FDE	MMADE	MMFDE
Directions	TPK	6.510	0.447	0.482	0.523	0.544	Sitting	TPK	6.417	0.400	0.547	0.461	0.548
	DLow	11.874	0.415	0.465	0.499	0.514		DLow	11.425	0.364	0.513	0.440	0.523
	GSPS	15.398	0.407	0.477	0.492	0.522		GSPS	14.966	0.323	0.454	0.411	0.484
	DivSamp	15.663	0.389	0.463	0.502	0.523		DivSamp	15.614	0.317	0.465	0.417	0.490
	BelFusion	7.090	0.378	0.422	0.484	0.494		BelFusion	6.495	<u>0.306</u>	<u>0.446</u>	0.400	0.461
	HumanMAC	6.357	0.391	0.456	0.475	0.475		HumanMAC	5.941	0.312	0.456	<u>0.404</u>	0.472
	CoMusion	7.527	0.372	0.417	0.454	0.435		CoMusion	6.237	0.307	0.448	0.406	0.468
	Ours	4.673	0.371	<u>0.421</u>	<u>0.467</u>	<u>0.461</u>		Ours	4.080	0.302	0.443	0.437	0.492
Discussion	TPK	6.966	0.511	0.581	0.570	0.600	SittingDown	TPK	7.393	0.496	0.678	0.531	0.666
	DLow	11.872	0.472	0.536	0.533	0.549		DLow	12.044	0.451	0.605	0.495	0.606
	GSPS	14.099	0.448	0.541	<u>0.526</u>	0.563		GSPS	13.725	0.406	0.561	0.461	0.565
	DivSamp	15.310	0.432	0.526	0.534	0.557		DivSamp	14.899	0.413	0.579	0.478	0.586
	BelFusion	9.172	0.420	0.507	0.512	0.530		BelFusion	9.026	0.413	0.585	0.468	0.587
	HumanMAC	7.496	0.434	0.533	0.547	0.571		HumanMAC	6.871	<u>0.381</u>	0.530	0.471	0.568
	CoMusion	8.747	0.409	<u>0.497</u>	0.527	0.523		CoMusion	7.253	0.378	<u>0.546</u>	0.472	0.578
	Ours	5.420	0.402	0.489	0.534	0.541		Ours	4.993	0.395	0.581	0.504	0.589
Eating	TPK	6.412	0.388	0.473	0.452	0.472	Smoking	TPK	6.522	0.422	0.529	0.509	0.560
	DLow	11.603	0.358	0.433	0.439	0.452		DLow	11.549	0.400	0.515	0.490	0.537
	GSPS	15.570	0.334	0.419	0.424	0.448		GSPS	14.822	0.466	0.485	0.472	0.530
	DivSamp	<u>15.681</u>	0.321	0.419	0.428	0.445		DivSamp	15.688	0.353	0.486	0.475	0.523
	BelFusion	5.954	0.310	0.381	0.418	0.420		BelFusion	6.780	0.341	0.467	0.467	0.512
	HumanMAC	4.817	0.305	0.374	<u>0.411</u>	<u>0.409</u>		HumanMAC	5.415	0.339	0.475	<u>0.445</u>	0.501
	CoMusion	6.149	0.295	0.366	0.408	0.395		CoMusion	6.802	0.311	0.443	0.427	0.458
	Ours	3.461	<u>0.303</u>	<u>0.371</u>	0.420	0.413		Ours	4.222	<u>0.322</u>	<u>0.447</u>	0.451	<u>0.486</u>
Greeting	TPK	6.779	0.555	0.615	0.571	0.598	Waiting	TPK	6.631	0.480	0.584	0.526	0.568
	DLow	11.897	0.530	0.590	0.561	0.564		DLow	11.680	0.441	0.541	0.497	0.534
	GSPS	14.974	0.502	0.592	<u>0.532</u>	0.577		GSPS	15.000	0.400	0.514	<u>0.475</u>	0.529
	DivSamp	15.447	0.489	0.575	0.535	0.562		DivSamp	15.455	0.387	0.517	0.486	0.535
	BelFusion	8.482	0.482	0.544	0.524	0.540		BelFusion	7.747	0.390	0.507	0.471	<u>0.511</u>
	HumanMAC	7.939	0.499	0.571	0.573	0.592		HumanMAC	6.506	0.385	0.532	0.496	0.557
	CoMusion	8.946	<u>0.481</u>	0.556	0.558	0.552		CoMusion	7.690	0.358	0.484	0.487	0.476
	Ours	5.444	0.473	<u>0.551</u>	0.554	<u>0.551</u>		Ours	4.434	<u>0.368</u>	<u>0.506</u>	0.515	0.538
Phoning	TPK	6.410	0.377	0.475	0.468	0.507	WalkDog	TPK	7.384	0.560	0.694	0.592	0.665
	DLow	11.542	0.343	0.444	0.451	0.487		DLow	11.882	0.490	0.566	0.539	0.570
	GSPS	15.050	0.311	0.413	0.436	0.476		GSPS	13.746	0.459	0.564	0.530	0.587
	DivSamp	15.751	0.296	0.400	0.437	0.471		DivSamp	15.616	0.439	0.555	0.532	0.577
	BelFusion	6.649	0.283	0.375	0.426	0.445		BelFusion	9.335	0.432	0.530	<u>0.527</u>	<u>0.569</u>
	HumanMAC	5.069	0.287	0.383	0.405	0.431		HumanMAC	7.741	0.441	0.547	0.543	0.591
	CoMusion	6.427	<u>0.268</u>	0.363	0.390	0.399		CoMusion	9.154	0.426	0.540	0.520	0.554
	Ours	4.013	0.264	0.363	0.404	<u>0.421</u>		Ours	5.823	<u>0.431</u>	<u>0.536</u>	0.566	0.601
Photo	TPK	6.894	0.541	0.689	0.548	0.633	WalkTogether	TPK	6.718	0.443	0.548	0.535	0.573
	DLow	11.931	0.507	0.655	0.516	0.596		DLow	11.951	0.395	0.495	0.503	0.530
	GSPS	14.310	0.485	0.663	<u>0.502</u>	0.606		GSPS	15.030	0.316	0.440	0.473	0.516
	DivSamp	15.330	0.474	0.665	0.506	0.607		DivSamp	16.095	0.321	0.458	0.486	0.525
	BelFusion	8.446	<u>0.434</u>	0.601	0.462	0.546		BelFusion	6.378	0.296	0.393	0.484	0.495
	HumanMAC	7.505	0.438	<u>0.600</u>	0.511	0.619		HumanMAC	4.336	0.298	0.387	<u>0.447</u>	0.454
	CoMusion	8.923	0.422	0.606	0.503	0.611		CoMusion	6.512	<u>0.270</u>	<u>0.372</u>	0.435	0.431
	Ours	5.522	0.445	0.591	0.531	0.590		Ours	4.128	0.256	0.357	0.449	<u>0.451</u>
Posing	TPK	6.520	0.466	0.538	0.542	0.565	Walking	TPK	6.708	0.455	0.533	0.538	0.558
	DLow	11.875	0.442	0.521	0.510	<u>0.525</u>		DLow	11.904	0.428	0.518	0.516	0.539
	GSPS	15.149	0.415	0.527	0.498	0.543		GSPS	14.797	0.351	0.469	0.490	0.528
	DivSamp	15.429	0.395	0.499	0.510	0.541		DivSamp	15.964	0.373	0.535	0.508	0.547
	BelFusion	8.438	0.406	0.510	<u>0.498</u>	0.531		BelFusion	5.116	0.367	0.471	0.530	0.546
	HumanMAC	7.320	0.407	0.530	0.512	0.553		HumanMAC	4.306	0.321	0.447	0.472	0.485
	CoMusion	8.236	<u>0.393</u>	<u>0.501</u>	0.492	0.499		CoMusion	6.487	<u>0.308</u>	<u>0.443</u>	0.447	0.465
	Ours	5.143	0.389	0.497	0.532	0.537		Ours	3.522	0.263	0.406	0.452	0.474
Purchases	TPK	7.450	0.505	0.522	0.535	0.538							
	DLow	11.947	0.430	0.422	0.493	0.477							
	GSPS	13.969	0.414	0.429	0.497	0.497							
	DivSamp	14.967	0.388	0.404	0.502	0.478							
	BelFusion	10.272	0.410	<u>0.409</u>	<u>0.494</u>	0.472							
	HumanMAC	8.601	<u>0.403</u>	0.410	0.506	<u>0.439</u>							
	CoMusion	9.484	0.405	0.426	0.496	0.425							
	Ours	5.185	<u>0.403</u>	0.421	0.542	0.460							

Table 1: Comparison of different methods on various classes and metrics.

reducing computational costs compared to larger step configurations. This demonstrates our approach can maintain high prediction accuracy even with a significantly accelerated sampling process, making it more practical for real-time applications.

local_layer	total_layer	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
4	8	5.013	0.372	0.484	0.529	0.542
5	10	4.766	0.361	0.467	0.515	0.533
6	12	4.458	0.347	0.452	0.513	0.535
12	12	4.229	0.354	0.461	0.520	0.536
7	14	4.587	0.362	0.458	0.522	0.540

Table 2: Performance comparison with different configurations of local fusion layers and total layers.

39

Noising steps	DDIM steps	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
1000	100	5.172	0.350	0.454	0.518	0.533
100	10	4.574	0.363	0.468	0.520	0.536
10	10	4.458	0.347	0.452	0.513	0.535

Table 3: Experiment results of the ablation study on diffusion steps

4 More visualization results

To further illustrate the effectiveness of our model, we provide a variety of qualitative results in the supplementary materials. Specifically, the included ZIP file contains several animated GIF sequences showcasing motion predictions generated by our model across different action categories. These visualizations highlight the model’s ability to generate smooth, realistic, and temporally coherent motion, and offer an intuitive understanding of its performance beyond numerical metrics.