A Temporal and Content Co-Awareness Latent Diffusion for Controllable Hand Image Generation

Supplementary Material

In the supplemental material, we provide:

- the time-segmented feature injection study in Sec. 1,
- the details of architecture in Sec. 2,
- qualitative results on the synthetic datasets in Sec. 3,
- more results on hand pose evaluation in Sec. 4,
- the ablation studies about each component in Sec. 5,
- the ablation studies about the number of reference appearance images in Sec. 6,
- temporal dynamics of the context factor in Sec. 7,
- the study of appearance modulation weight α_{a} in Sec. 8,
- efficiency analysis in Sec. 9.

1. The Time-Segmented Feature Injection

As illustrated in Fig. 2, we conduct a time-segmented feature injection study of the denoising process to unravel the dynamic roles of pose and appearance controls. On the left, the quantitative trajectories are evaluated using PCK for geometric consistency and LPIPS for appearance similarity. These curves show how the generation quality drops when removing certain control signals during specific time windows, such as the 0-5 or 5-10 step intervals. Specifically, losing pose control in early stages leads to a sharp drop in structural accuracy , while missing appearance control later mainly degrades appearance fidelity. These findings reveal that the pose condition dominates the global structure in early denoising timesteps, while the appearance condition gradually refines local textures in later denoising timesteps.

Furthermore, the qualitative comparisons on the right across different levels of pose and texture complexity highlight two critical insights: (1) Early-stage Coupling: Removing pose control within the 0-10 step interval not only causes structural collapse but also induces severe color drift and texture blurring. This phenomenon proves that pose and appearance exhibit a strong coupling effect in the early denoising phases. (2) Content Sensitivity: For complex poses (e.g., Row 1 of Fig. 2), the model maintains a high dependency on pose guidance even in the later denoising stages; lacking pose control in the 20-30 or 30-40 step intervals still leads to joint misalignment. Our qualitative results across varying conditions reveal that control requirements are highly sensitive to content complexity. These findings essentially reflect the characteristic of the diffusion process—a progressive refinement driven by the denoising state and the complexity of conditions.

Table 1. Ablation studies about each component.

PIAE	TCCA		FID↓	LPIPS↓	SSIM↑	PSNR↑
	TA	CA				
✓	✓	✓	9.046	0.4078	0.5714	14.965
	✓	✓	13.668	0.5049	0.4635	12.651
✓		✓	9.804	0.4572	0.5693	14.910
✓	✓		9.652	0.4299	0.5673	14.774
			18.015	0.5401	0.4428	11.814

Figure 1. Quantitative results of ablation studies about each component.

2. Architecture Details

In this section, we introduce the details of the Temporal and Content Co-Awareness (TCCA) module. In this module, we jointly model three semantic factors: context, pose, and appearance, through three sets of learnable queries. These factors are represented as $\mathbf{Q}_p,\ \mathbf{Q}_a,\ \mathbf{Q}_c \in \mathbb{R}^{N\times D}$, where each query set contains N learnable vectors of dimension D. To incorporate temporal information, the timestep embedding $\mathbf{t}_{\text{emb}}$ is concatenated with $(\mathbf{Q}_p, \mathbf{Q}_c)$ for pose modeling and with $(\mathbf{Q}_a, \mathbf{Q}c)$ for appearance modeling. The fused tokens are processed by transformer encoders to enable cross-domain interaction. Specifically, the timestep embedding $\mathbf{t}$emb serves a role analogous to a learnable [CLS] token in standard Transformers. Through the self-attention mechanism, it **aggregates global semantic information from both temporal cues and context-aware factors.** As a result, the updated time token becomes aware not only of the current denoising state but also of the content complexity of

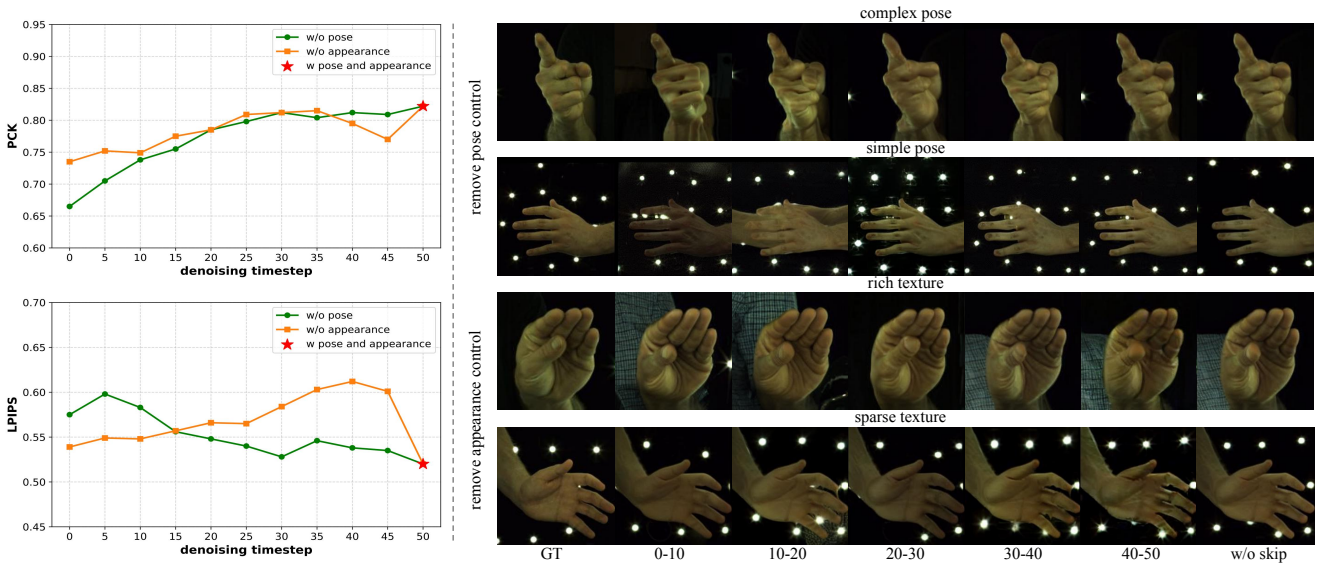

Figure 2. The time-segmented feature injection study investigates the distinct impact of pose and appearance controls at different denoising timesteps. Left: Quantitative analysis using LPIPS to assess appearance fidelity and PCK to measure pose consistency. Right: Qualitative visualizations of generated images across different levels of pose and texture complexity.

the conditions, forming a comprehensive representation for adjusting modulation strength of the control signals.

3. Qualitative Results on the Synthetic Datasets

As stated in the main paper, our training is performed on four subsets of the FoundHand-10M dataset: the real InterHand2.6M dataset and three synthetic datasets: DART, RenderIH, and ReInterHand. The InterHand2.6M dataset provides 1,361,062 frames covering both single hand and interacting hand scenarios. For evaluation, it offers 488,968 test frames, which serve as our primary real world benchmark. The synthetic subsets offer a wider range of poses and more appearance variations, which are beneficial for improving robustness during training.

For completeness, we also provide additional qualitative results on the three synthetic datasets. As shown in Fig. 7, our method preserves fine-grained appearance details and accurate hand geometry across diverse synthetic domains. Furthermore, when compared with existing baselines, our proposed approach significantly better maintains pose alignment and cross-identity appearance consistency, effectively mitigating common visual artifacts such as texture blurring or joint distortion.

4. More Results on Hand Pose Evaluation

The quality of synthesized images directly determines the reliability of feature extraction by the pose estimator, thereby influencing the pose estimation accuracy. The observations can be summarized as follows: (1) Pose distortions in synthesized images hinder geometric feature extraction, leading to substantial pose estimation errors. As shown in Fig. 8, existing methods often fail to adapt the modulation strength of control signals, which results in strong coupling between pose and appearance representa-

Table 2. Quantitative results of ablation studies about the number of reference appearance images.

Type	FID↓	LPIPS↓	SSIM↑	PSNR↑
N=1	10.608	0.4264	0.5003	14.194
N=2	9.982	0.4355	0.5290	14.383
Ours (N=3)	9.046	0.4078	0.5714	14.965

Figure 3. Qualitative results with different numbers of reference appearance images. We validate the effectiveness of our pose-invariant appearance modeling: even with a single reference (N=1), the model preserves stable appearance consistency.

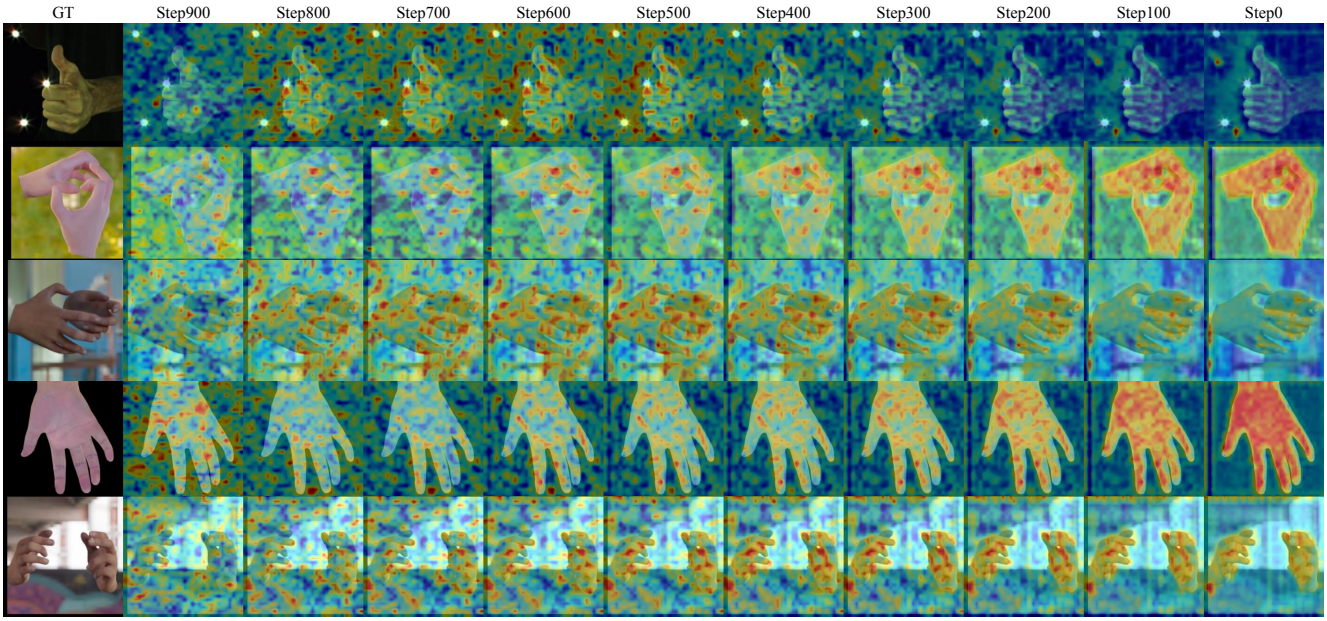

Figure 4. Visualizing the attention map of the context factor $\mathbf{Q}_c$ over the noisy latent in TCCA across denoising timesteps.

tions. Consequently, severe artifacts such as finger twisting, fingertip merging, and unnatural joint bending frequently occur (e.g., Row 1, 4). (2) Unnatural appearance further degrades feature extraction. Although HaMeR primarily focuses on pose, its feature extractor is still sensitive to color and texture distributions of the hand images. Coshand frequently exhibits severe color shifts or illumination inconsistencies (e.g., oversaturated or yellowish tones), which negatively affect pose estimation. Our method achieves temporal and content co-awareness, which enables the dynamic adjustment of pose and appearance injection strength, resulting in pose-appearance decoupling. This method ensures that the synthesized images exhibit pose consistency and superior appearance fidelity.

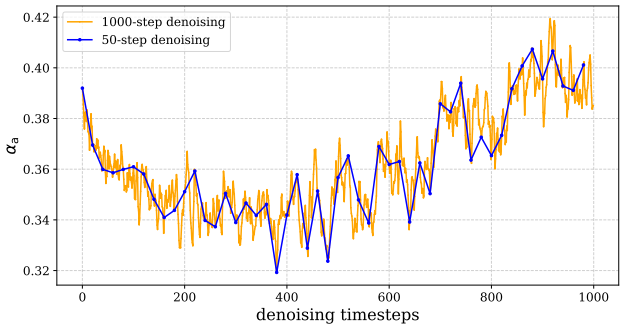

Figure 5. The appearance control injection strength α_a across all denoising timesteps. Reported values represent token-wise averages of the appearance modulation weights.

5. Ablation Studies about Each Component

Ablation studies (Tab. 1 and Fig. 1) demonstrate the contribution of each component of our method. Removing PIAE breaks pose-invariant appearance modeling, causing texture drift and inconsistent color under different poses. Eliminating the modulation guidance from Temporal-Aware (TA) factors disrupts temporal adaptivity, making pose constraints weak in early denoising steps and texture refinement insufficient in later denoising steps. Removing the modulation guidance from Content-Aware (CA) factors prevents the model from adapting modulation strength to pose and texture complexity, leading to geometric misalignment and artifacts in hard cases. For example, in Fig. 1, the sample in the fourth column shows blurred finger boundaries and over-smoothed skin textures without CA. When all components are removed, both structure and appearance degrade notably. These results confirm that **PIAE, TA, and CA provide complementary benefits**, ensuring robust appearance consistency, temporally aligned modulation, and content-aware adjustment, together maintaining the high-fidelity texture and pose consistency.

6. Ablation Studies about the Number of Reference Appearance Images

As shown in Fig. 3, our method maintains stable appearance fidelity even when only a single reference image is available. With a single reference appearance image, the generated hands already exhibit clear textures, stable color tones, and identity-consistent appearance. This demon-

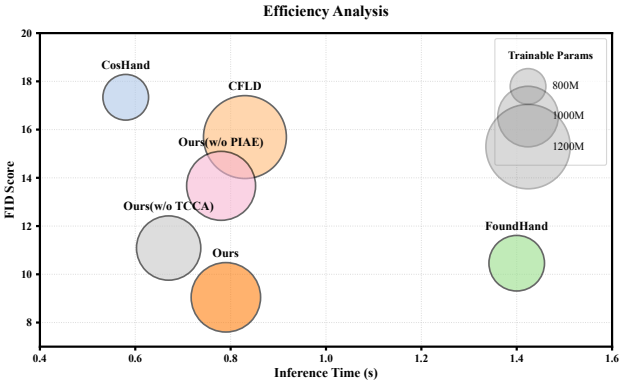

Figure 6. Inference latency and parameter comparison.

strates that the proposed PIAE effectively disentangles appearance from pose at the patch level, mitigating pose leakage into appearance subspaces. Using two or three reference images further enhances local details and reduces occlusion-related artifacts, and in our method we adopt N=3 as the default setting for optimal generation quality. This confirms that PIAE provides a robust pose-invariant appearance representation, enabling our framework to perform well even under minimal reference supervision.

7. Temporal Dynamics of the Context Factor

Fig. 4 visualizes the attention map of the context factor $\mathbf{Q}_c$ over the noisy latent across denoising timesteps. In the early, high-noise stages of denoising, the attention distribution remains broad and diffuse, enabling the model to capture low-frequency structural cues of the hand. As the denoising trajectory advances into intermediate timesteps, the attention progressively contracts toward the primary hand regions. In the later denoising timesteps, it sharpens further and focuses on fine-grained appearance details such as fingertips and knuckle contours, reflecting a transition from global structure modeling to detailed texture refinement. The global-to-local attention progression indicates that $\mathbf{Q}_c$ effectively infers the current denoising state and achieves temporal awareness to dynamically regulate the strength of the modulation signals throughout the diffusion process.

8. Study of Appearance Modulation Weight $\alpha_\mathbf{a}$

As shown in Fig. 5, the appearance modulation weight α_a exhibits a clear temporal evolution pattern across denoising timesteps. This indicates that the model relies less on appearance cues during the early noisy stages, when the global structure is still being formed, and progressively strengthens appearance constraints as the denoising process begins refining fine-grained texture details. Similar to the pose-modulation trend in the main paper, both the 50-step and 1000-step denoising schedules follow nearly identical trajectories, confirming the stability and generality of our adaptive modulation strategy.

9. Efficiency Analysis

Fig. 6 demonstrates that our method achieves a superior trade-off between generation quality and computational efficiency. While CosHand and FoundHand require fewer parameters via simple channel concatenation, they lack the adaptive capacity for pose-appearance injection, which leads to performance degradation. Notably, the inference latency of FoundHand is nearly $\mathbf{2}\times$ ours due to the heavy dense attention in its DiT-based architecture.

Figure 7. Additional qualitative results on synthetic datasets, showing that our method maintains consistent pose and appearance across diverse synthetic domains.

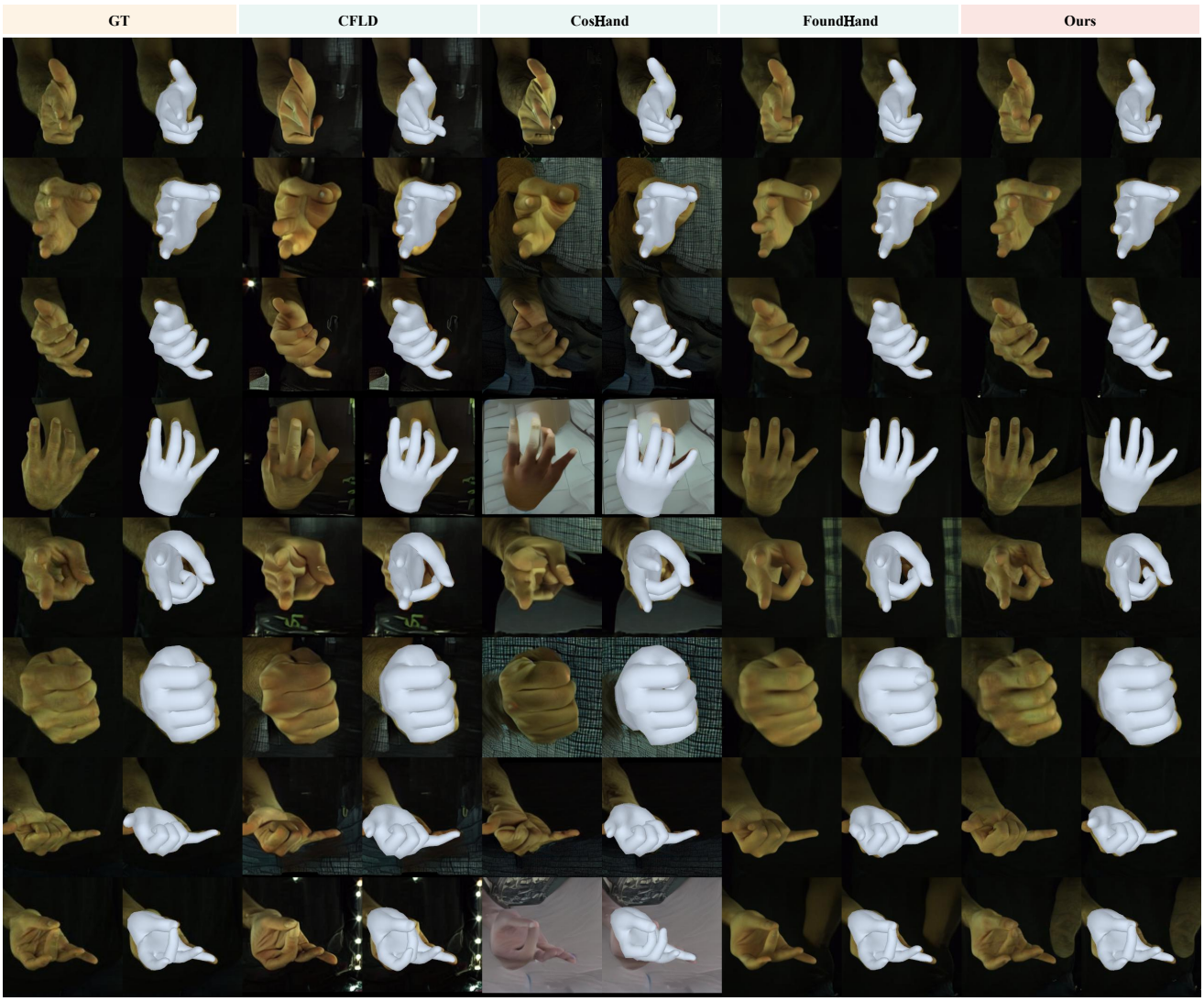

Figure 8. Qualitative comparisons with state-of-the-art approaches on hand pose evaluation.