

# X-Dyna: Expressive Dynamic Human Image Animation

## Supplementary Material

### 6. Video Results

In our offline webpage, we provide additional video results generated from X-Dyna. Please unzip the supplementary files and open the HTML file on your browser.

**Comparison of Different Appearance Reference Module Designs:** To demonstrate the effectiveness of our proposed Dynamics-Adapter, we provide visual comparisons with IP-Adapter and ReferenceNet. Please refer to the **Different Architecture Designs** section on the offline page for details.

**Comparison to Previous Works:** To evaluate the performance of X-Dyna in generating dynamic textures for human image animation, we present visual comparisons with previous state-of-the-art methods, including the ReferenceNet-based approach from [4] and the SVD-based method from [56]. Details can be found in the **Comparison to Previous Works** section.

**Ablation Study:** To highlight the contribution of Harmonic Data Fusion Training to our pipeline, we present a visualized ablation study. Please refer to the **Effectiveness of Mix data training** section of the attached page.

### 7. Quantitative Evaluation of Cross-Driving Reenactment

In this section, we present quantitative evaluations for cross-driving video generation. We generated 200 videos for X-Dyna and each baseline method using various in-the-wild driving motions and reference images. The overall quality of cross-driving generation is assessed using DTFVD and FID metrics, comparing the distribution of the generated videos with the training videos. To evaluate the control accuracy of facial expressions, we crop the face area of both generated and driving videos and calculate their mean difference of face landmarks by MediaPipe [28]. The numerical results are summarized in Tab. 5, where X-Dyna demonstrates superior face expression control accuracy (Face-Exp) and dynamics (DTFVD), and comparable perceptual quality (FID).

### 8. Details of User Study

In this section, we provide a comprehensive user study for qualitative comparison between X-Dyna and previous works [4, 17, 51, 56]. We generate 50 different human animation results from all baseline models and X-Dyna, where the results are anonymized and shuffled. On the online platform Prolific, we ask 100 users to rate these methods from 0(worst) - 5(best).

Table 5. **Quantitative comparisons of X-Dyna with recent state-of-the-art (SOTA) methods on cross-driving human animation.** A downward-pointing arrow indicates that lower values are better. **DTFVD** and **FID** are used to evaluate the overall quality of generated videos. **Face-Exp** denotes the absolute error of facial expressions between generated videos and driving videos.

Method	DTFVD ↓	FID ↓	Face-Exp ↓
MagicAnimate [51]	6.708	250.75	0.134
Animate-Anyone [17]	<u>6.820</u>	253.29	0.123
MagicPose [4]	7.062	<b>244.25</b>	0.121
MimicMotion [56]	6.823	258.91	<u>0.109</u>
X-Dyna	<b>5.923</b>	<u>246.16</u>	<b>0.105</b>

**Criteria for Judgment:** Since our paper focuses on the dynamics of texture generation and motion control with human reference, the criteria for evaluation are (1) dynamics quality of background nature (BG-Dyn), (2) dynamics quality of human foreground (FG-Dyn), (3) appearance and identity preservation ability (ID).

**Results and Statistical Analysis:** The result is presented in Tab. 3 of the main paper. In addition, we perform a one-way analysis of variance (ANOVA) test on the ratings. ANOVA tests whether the means of multiple groups of data (methods in this case) are significantly different. For each metric, we compare the ratings across all five methods. Specifically, **F-statistic** measures the ratio of variance between group averaged values to the variance within groups. A higher F-statistic indicates greater variability between group-averaged values relative to within-group variability. **P-value** tests the null hypothesis that all group means are equal. A small p-value (typically  $\leq 0.05$ ) indicates significant differences between groups. As reported in Tab. 6, all metrics (FG-Dyn, BG-Dyn, ID, Overall) have p-values  $\leq 0.05$ , indicating statistically significant differences between methods. The F-statistic for each metric shows the relative strength of these differences. X-Dyna consistently achieves the highest averaged ratings across all metrics (as seen in Tab. 3 of the main paper), and the differences are statistically significant.

Metric	F-statistic	p-value
FG-Dyn	7.495	0.000007
BG-Dyn	5.327	0.000331
ID	4.685	0.001016
Overall	5.617	0.000199

Table 6. ANOVA Test Results for Ratings from the User Study.

## 9. More Details on Prior Appearance Reference Control Designs

**ReferenceNet** was initially introduced by Animate-Anyone [17]. It adopts the same architecture as the Appearance Encoder in MagicAnimate [51] and the Appearance Control Model in MagicPose [4]. Building upon prior advancements in dense reference image conditioning, such as the manipulation of self-attention layers in the UNet demonstrated by MasaCtrl [3] and Reference-only ControlNet [54], ReferenceNet enhances identity and background preservation, significantly improving single-frame fidelity. The naive self-attention calculation in the transformer blocks of the diffusion UNet can be represented as:

$$A_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right) V_i, \quad (4)$$

However, ReferenceNet introduces a trainable duplicate of the base UNet, which computes conditional features from the reference image  $I_R$  for each frame  $I_i$ . Unlike ControlNet, which integrates conditions additively in a residual manner, ReferenceNet injects the features derived from  $I_R$  directly into the spatial self-attention layers of the UNet blocks. This is achieved by concatenating the reference features with the original UNet’s self-attention hidden states. The process can be expressed as:

$$A_i = \text{softmax}\left(\frac{Q_i K_i'^\top}{\sqrt{d}}\right) V_i', \quad (5)$$

$$Q_i = W^{Q_i} z_i, K_i' = W^{K_i} [z_i, z_r], V_i' = W^{V_i} [z_i, z_r], \quad (6)$$

where  $[\cdot]$  denotes concatenation operation and  $z_i, z_r$  denotes the self-attention hidden states from  $I_i, I_R$ . This self-attention mechanism strictly queries and preserves the information from the reference image in the denoising process, including human identity and background.

**IP-Adapter** [52] is composed of two key components: an image encoder that extracts features from the image prompt and adapted modules with decoupled cross-attention to integrate these features into the LDM UNet. A pretrained CLIP image encoder is employed to extract features from the reference image  $I_R$ .

To effectively decompose the extracted global image embedding, a lightweight trainable projection network—comprising a linear layer and Layer Normalization is utilized. This network projects the global image embedding into a sequence of features, ensuring that the dimensionality of the projected image features matches the dimensionality of the text features used in the UNet.

The integration of image features into the UNet is performed through adapted modules with decoupled cross-attention. In the original LDM, text features from the

CLIP text encoder are incorporated into the UNet via cross-attention layers. In this setup, given the query features  $z_r$  derived from  $I_R$ , the hidden states of the UNet for each frame  $I_i$ , and the text features  $z_t$ , the output of the cross-attention mechanism is defined as:

$$A_i' = \text{softmax}\left(\frac{Q_i' K_i'^\top}{\sqrt{d}}\right) V_i', \quad (7)$$

$$Q_i' = W^{Q_i'} z_i, K_i' = W^{K_i'} z_t, V_i' = W^{V_i'} z_t, \quad (8)$$

Then, another cross-attention layer for each original layer in the UNet is added to inject image features. Given the image features  $z_r$ , the output of this cross-attention is computed as follows:

$$A_i'' = \text{softmax}\left(\frac{Q_i' K_R'^\top}{\sqrt{d}}\right) V_R', \quad (9)$$

$$Q_i' = W^{Q_i'} z_i, K_R' = W^{K_R'} z_r, V_R' = W^{V_R'} z_r, \quad (10)$$

The same query  $Q_i'$  is shared between the image cross-attention and the text cross-attention mechanisms. As a result, only two additional trainable parameters,  $W^{K_R'}$  and  $W^{V_R'}$ , are introduced as linear layers for each cross-attention module. The output of the image cross-attention is then combined with the output of the text cross-attention through a simple addition operation. Accordingly, the final formulation of the decoupled cross-attention is denoted as:

$$Out_i' = A_i' + \lambda A_i'', \quad (11)$$

where  $\lambda$  is an adjustable parameter. When  $\lambda = 0$ , the model is the same as a frozen pre-trained LDM.

**Stable Video Diffusion (SVD)** [1] is a diffusion-based video generation model that extends the latent diffusion framework originally designed for 2D image synthesis to produce high-resolution, temporally consistent videos from text and image inputs. SVD UNet introduces two types of temporal layers: 3D convolution layers and temporal attention layers, and temporal layers are also incorporated into the VAE decoder. For training, the DDPM [13] noise scheduler used in Stable Diffusion [36] is replaced by the EDM [21] scheduler, alongside EDM’s sampling method. Unlike traditional DDPM models that rely on discrete timesteps  $t$  for denoising, EDM uses a continuous noise scale  $\sigma_t$ . By incorporating  $\sigma_t$  as input to the model, EDM enables more flexible and effective sampling, utilizing continuous noise strengths instead of discrete timesteps during the denoising process. This end-to-end training paradigm enhances temporal consistency in video generation. However, SVD faces challenges when dealing with cross-driving

cases. The reference image is concatenated with the noisy latent and directly input to the UNet, leading the model to deform the reference image into the first frame of the video rather than encoding the reference image and learning its semantic information implicitly, as achieved by ReferenceNet [17], IP-Adapter [52], and Dynamics-Adapter. While fine-tuning the UNet, as in MimicMotion [56], is a potential solution, it struggles to generalize to out-of-domain identities beyond the training data, as shown in Fig. 5 of our main paper and the supplementary videos.