IDOL: Unified Dual-Modal Latent Diffusion for Human-Centric Joint Video-Depth Generation

Yuanhao Zhai*[1], Kevin Lin[2], Linjie Li[2], Chung-Ching Lin[2], Jianfeng Wang[2], Zhengyuan Yang[2], David Doermann[1], Junsong Yuan[1], Zicheng Liu[3], and Lijuan Wang[2]

[1] State University of New York at Buffalo
[2] Microsoft
[3] Advanced Micro Devices
https://yhzhai.github.io/idol/

Abstract. Significant advances have been made in human-centric video generation, yet the joint video-depth generation problem remains underexplored. Most existing monocular depth estimation methods may not generalize well to synthesized images or videos, and multi-view-based methods have difficulty controlling the human appearance and motion. In this work, we present IDOL (unIfied Dual-mOdal Latent diffusion) for high-quality human-centric joint video-depth generation. Our IDOL consists of two novel designs. First, to enable dual-modal generation and maximize the information exchange between video and depth generation, we propose a unified dual-modal U-Net, a parameter-sharing framework for joint video and depth denoising, wherein a modality label guides the denoising target, and cross-modal attention enables the mutual information flow. Second, to ensure a precise video-depth spatial alignment, we propose a motion consistency loss that enforces consistency between the video and depth feature motion fields, leading to harmonized outputs. Additionally, a cross-attention map consistency loss is applied to align the cross-attention map of the video denoising with that of the depth denoising, further facilitating spatial alignment. Extensive experiments on the TikTok and NTU120 datasets show our superior performance, significantly surpassing existing methods in terms of video FVD and depth accuracy.

1 Introduction

The capacity to control and manipulate human-centric video content – modulating subjects' actions, altering appearance and background – has attracted continual study [29, 41, 62, 63, 68]. With the rapid evolution of generative models, from generative adversarial networks [6, 17] to recent latent diffusion models [56], the quality of the generated video has improved significantly. However, most of the existing research [29, 41, 62, 63, 68] has focused on the generation of 2D content. This imposes a natural limit on applications that require depth perception,

* Work done during an internship at Microsoft.

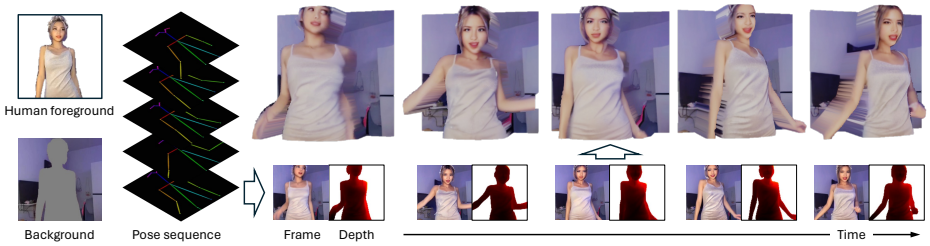

Fig. 1: Given a human foreground image, an arbitrary background image, and a defined pose sequence, our IDOL generates high-fidelity video and the corresponding depth maps, which can be rendered as realistic 2.5D video.

such as virtual and augmented reality, as well as advanced video games. In this paper, we explore joint video-depth generation for human actions (*e.g.*, dancing and daily activities), where the video and the corresponding depth map are simultaneously generated. By learning a holistic representation of the human, it not only enhances the visual fidelity of the synthesized content but also paves the way for applications that demand a deeper spatial understanding.

Existing methods confront several challenges in dealing with this task. First, discriminative monocular depth estimation methods [27, 54], which are typically trained on natural images, have been empirically observed to exhibit degraded performance when applied to generated images [2, 3, 80, 81]. On the other hand, while multi-view-based methods [18, 48, 52, 69, 73] can estimate depth, they primarily focus on the synthesis of individual object/scene. They often struggle with inference from single-view inputs or underperform in manipulating the object's appearance and motion. To address these problems, we propose to jointly generate the video and the corresponding depth.

However, joint video-depth generation presents nontrivial challenges for two primary reasons. First, video and depth are inherently two different modalities: the former is represented as a 3-channel RGB frame sequence and the latter as a scalar depth map sequence [53, 54]. Contrarily, prevailing diffusion models are pre-trained on the single-modal image generation task [56]. Thus, designing a dual-modal diffusion model for joint video-depth generation is challenging, not to mention harnessing the power of pre-trained latent diffusion models. Second, spatial layout control has been a long-standing problem in diffusion models [16, 22, 43, 66]. Even with human pose control, maintaining an accurate spatial alignment between the generated video and depth remains a challenge. This challenge can be more pronounced if the denoising process is conducted in the latent space, given the intricate mapping from the latent space to the final output. In addressing these problems, we propose IDOL (unIfied Dual-mOdal Latent diffusion), a framework that aims to generate a human-centric video and the corresponding depth jointly.

To tackle the problem of distinct video and depth representations, we propose to render depth maps as RGB images by applying a color map to them. This

conversion reframes the depth generation task as a stylized video generation problem. Besides, existing methods suggest that incorporating depth as input enhances structural understanding and boosts the generation quality [13]. Drawing inspiration from this, we hypothesize that a richer interplay between depth and video generation would be reciprocally beneficial. To this aim, we design a unified dual-modal U-Net that shares parameters across both video and depth denoising processes. Our model leverages a modality label to specify the denoising target, *i.e.*, video or depth. In this way, it enables joint learning of depth and video for better generation quality while remaining parameter-efficient. Furthermore, a cross-modal attention layer is added during the joint denoising process, enabling an explicit correlation between the joint video-depth learning.

To ensure a precise video-depth alignment, we propose to synchronize the motion pattern of the intermediate video and depth features. Specifically, the intermediate video and depth features within the U-Net contain semantic information, where regions with similar semantic meanings are represented with similar features [5,16,66], as shown in Fig. 3. Thus, by using the proposed motion consistency loss, we enforce a consistent motion between video and depth features, thereby promoting a precise video-depth spatial alignment. Additionally, we propose a cross-attention map consistency loss to further strengthen video-depth alignment. This is drawn inspiration from existing observations [22, 49] on the influence of cross-attention maps on the spatial layout. Different from [22, 49], we consider spatial alignment across video and depth modalities.

In summary, our contributions are as follows.

- We propose IDOL, a pioneering framework for human-centric joint video-depth generation. Our IDOL features a parameter-sharing unified dual-modal U-Net for video-depth denoising. Besides, cross-modal attention modules are used to enable the mutual information flow.
- We propose a motion consistency loss and a cross-attention map consistency loss to allow fine-grained video-depth spatial alignment.
- We conduct extensive experiments on two distinct datasets: TikTok and NTU120, whose depth maps were obtained using different methods [27, 53, 54]. The results not only show a significant improvement in video quality and depth accuracy over state-of-the-art methods, but also highlight the flexibility of our method in adapting to different depth maps. Furthermore, our experiments reveal that our IDOL can be easily adapted to different diffusion models.

2 Related Work

Controllable diffusion models. Evolving from early diffusion models [12, 24], recent latent diffusion models (LDM) [56] achieves high-quality image generation by conducting the denoising process in the latent space. ControlNet [78], GLIGEN [33] and T2I-Adapter [44] add trainable modules on pre-trained diffusion models, and achieve fine-grained control given additional inputs, such as sketch and depth. HumanSD [29] directly concatenates the pose feature and the noise in the input to avoid potential feature discrepancy problems. In the

video generation domain, controllable video diffusion models [11, 26, 41] leverage the depth/flow/edge sequence to generate a temporally consistent video. Our work builds upon these controllable models, utilizing plug-in modules to facilitate human pose control in IDOL. Our experiments demonstrate our adaptability across various frameworks (*e.g.*, ControlNet [78] and T2I-Adapter [44]).

Image animation. Image animation involves producing a video wherein an object from the source image moves in congruence with the motion in a target video. Traditional methods require specific knowledge about the target object, such as facial landmarks [21, 50, 77], gestures [65], or semantic segmentation maps [46]. Several methods learn a motion field [62, 70] from the driving videos. First order motion model (FOMM) [62] improves the animation quality by learning a local affine transformation. Motion representation for articulated animation (MRAA) [63] achieves better quality in representing articulated motions, such as the human body. Thin-plate spline (TPS) [79] estimates a more accurate motion, and improves previous methods over large-scale motions. Besides, there exists a line of methods [42, 55] exploiting end-to-end learning frameworks. Based on diffusion models, DreamPose [30] proposes an adapter module, such as to control the appearance of the generated human. DisCo [68] disentangles the human attribute and pose condition learning by employing a two-stage training scheme. Several concurrent works improve DisCo [68] using more advanced appearance and motion control [8,25,74]. Furthermore, several methods [31,41] inject motion/pose prior to pre-trained text-to-image models. Distinctively, our IDOL focuses on the joint video-depth generation, and utilizes depth generation to enhance the overall quality of the animated video, instead of video-only generation.

Multi-modality synthesis. There exists a line of diffusion models for view synthesis, such as diffusion models using multi-view images [18, 69, 73], point cloud [45], and text-to-3D models [48, 52]. There are also methods developed for human body mesh estimation [38–40]. However, they may struggle with human-centric joint video-depth synthesis, as they typically require multi-view input or lack the ability to precisely control the appearance and motion of the target object/scene. Except for the use of multi-view images, LDM3D [64] modifies the autoencoder in LDM, such that the latent can be decoded into an RGB image and a depth map. MM-Diffusion [58] focuses on the audio-video generation, and features a coupled U-Net structure to simultaneously denoise video and audio latents. Concurrently, HyperHuman [37] proposes structural expert branches to denoise latents of different modalities, such as RGB image, depth and surface normal. To the best of our knowledge, there is no existing method directly working on the joint video-depth generation task, and thus, we modify the backbones of existing multi-modal generation methods for comparison.

Diffusion models for dense prediction. Diffusion models have been used in various dense prediction tasks, such as segmentation tasks [1, 5, 10, 28, 71, 72], depth estimation [28, 59] and object detection [9]. For depth estimation methods [28, 59], they typically necessitate modifying the output from 3-channel RGB images to scalar depth maps. Such a transformation impedes their ability to harness the power of large-scale pre-trained diffusion models [56]. Concurrently,

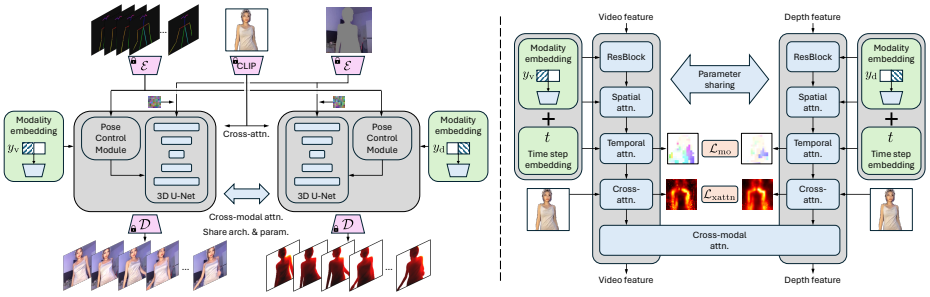

Fig. 2: Left: **Overall model architecture.** Our IDOL features a unified dual-modal U-Net (gray boxes), a parameter-sharing design for joint video-depth denoising, wherein the denoising target is controlled by a one-hot modality label (y_{v} for video and y_{d} for depth). Right: **Individual U-Net block structure.** Cross-modal attention is added to enable mutual information flow between video and depth features, with consistency loss terms $\mathcal{L}_{\mathrm{mo}}$ and $\mathcal{L}_{\mathrm{xattn}}$ ensuring the video-depth alignment. Skip connections are omitted for conciseness.

DepthAnything [75] address this problem by leveraging large-scale unlabeled data to improve the estimation quality. Additionally, when applied to synthesized content, existing depth estimation methods tend to underperform [2, 3, 80, 81], failing to generate high-quality depth maps from outputs of image animation methods. To mitigate this problem, we propose IDOL to directly synthesize the video and the corresponding depth, significantly improving the depth accuracy.

In contrast to existing methods, we reframe the depth synthesis task as a stylized image synthesis task. By rendering target depths as RGB images, we are positioned to directly use pre-trained image generation models with minimal modifications. Furthermore, we propose a unified dual-modal U-Net for improved joint video-depth generation, and enhance video-depth spatial alignment via the proposed consistency losses.

3 Method

Problem formulation. We begin by formally defining the task of human-centric joint video-depth generation. Given a human foreground image f, a background image b, and a pose sequence $p = \{p_1, p_2, ..., p_L\}$ of length L, our objective is to generate a video $v = \{v_1, v_2, ..., v_L\}$ and the associated depth map sequence $d = \{d_1, d_2, ..., d_L\}$. The video v should faithfully animate the human foreground f into the target pose p while integrating it with the background b. The depth map sequence d should correctly represent the depth within the video v. An illustration is shown in Fig. 1.

Preliminaries. Latent diffusion models (LDM) [56] recently show great success in image generation. It operates in the latent space of a pre-trained autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, where a time-conditioned U-Net [57] $\epsilon_\theta(\cdot)$ with learnable parameter θ is used for denoising the latent feature. Within the U-Net, the conditional signal

c (*e.g.*, textual prompt) is fed in the cross-attention module after CLIP [51] encoding. During training, the objective is the mean square error (MSE) between the predicted and ground truth noise:

$$\mathcal{L} = \mathbb{E}_{v,\epsilon,t,c}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right], \tag{1}$$

where $\epsilon \in \mathcal{N}(0, I)$ is the ground truth noise, and z_t is the noisy latent at the t-th reverse step, and can be obtained from $\mathcal{E}(v)$ and ϵ [24]. To enable spatial layout control, ControlNet [78] adds a copy of the U-Net down and middle blocks upon the LDM. The output of ControlNet is added to the original U-Net via skip connections. By manipulating the intermediate features, ControlNet trained with additional conditional input (poses in our case) can control the output image/video [68].

3.1 Unified dual-modal U-Net

Video and depth are different modalities and typically have distinct representations. Designing a model architecture that can jointly generate both video and depth, and potentially leverage pre-trained diffusion models, presents a significant challenge. In addressing this problem, we propose to reformulate depth synthesis as a stylized image generation task. By rendering the depth map as a heatmap, we bridge the gap between the video and depth modalities. To further harness this approach, we introduce a unified dual-modal U-Net, specially tailored for the joint video-depth denoising.

Video LDM baseline. We build our model upon 3D U-Net [13], which modifies the 2D U-Net [56] by adding a temporal convolutional layer in the ResBlock, and adding a temporal attention layer after the spatial attention within each block. For the human appearance and pose control, we follow DisCo [68] to feed the CLIP foreground latent CLIP(f) via cross-attention, and feed the pose latent $\mathcal{E}(p)$ via a ControlNet to the U-Net. We show in Sec. 4.1 that our method can be implemented on different pose control modules, such as T2I-Adapter [44]. For the background control, different from DisCo that leverages a ControlNet, we find that adding the background latent $\mathcal{E}(b)$ to the input noise of the U-Net not only yields comparable results but also reduces the number of parameters. The proposed model architecture is illustrated in Fig. 2.

Sharing U-Net for joint video-depth denoising. Existing methods [13] have shown that incorporating depth map as conditional input enhances the model's video structure awareness, improving the video generation quality. Building on this insight, we hypothesize that maximizing the information exchange between video and depth during generation can benefit the generation quality for both modalities. Thus, we propose a unified dual-modal U-Net for video and depth denoising, where the architecture and parameters of the U-Net and ControlNet are shared between the two modalities. To indicate the denoising modality, we add a learnable modality embedding to the time step embedding, which is fed to all blocks within the U-Net and the ControlNet, as shown in Fig. 2 right. In this way, given a one-hot modality label, the corresponding modality embedding will

be selected to feed into the model and further control the output modality. Apart from being parameter-efficient, we find that this unified architecture improves both video and depth generation quality (Sec. 4.2).

Cross-modal attention. Though the unified dual-modal U-Net enables implicit structural information learning, direct information communication during the video and depth denoising process is needed for better alignment. Thus, we add a multi-modality attention at the end of each block (Fig. 2 right). During the joint video-depth denoising process, the video and depth features are concatenated to conduct spatial self-attention. Note that self-attention is only performed in the spatial dimension, as it aims to promote spatial alignment between video and depth, and the preceding temporal layers already ensure the temporal smoothness.

The joint video-depth denoising objective $\mathcal{L}_{\text{denoise}}$ is:

$$\begin{aligned}\mathcal{L}_{\text{denoise}} = \mathbb{E}_{v,d,\epsilon_{\text{v}},\epsilon_{\text{d}},t,f,b,p}\big[\|\epsilon_{\text{v}} - \epsilon_\theta(z_{\text{v},t}, t, f, b, p; y_{\text{v}})\|_2^2 + \\ \|\epsilon_{\text{d}} - \epsilon_\theta(z_{\text{d},t}, t, f, b, p; y_{\text{d}})\|_2^2\big],\end{aligned} \tag{2}$$

where y_{v} and y_{d} are the modality labels for video and depth, respectively, $\epsilon_{\text{v}}, \epsilon_{\text{d}} \in \mathcal{N}(0, I)$ are independently sampled Gaussian noises, and f, b, p denote the human foreground image, the background image, and the target pose sequence, respectively. In Eq. (2), the first term is the video denoising loss and the second term is the depth denoising loss.

3.2 Learning video-depth consistency

Though the pose control ensures the coarse correspondence between the video and depth, we empirically find that only applying the joint denoising objective $\mathcal{L}_{\text{denoise}}$ may still lead to misalignment between the generated video and depth, especially when the training data is limited, as shown in the first column of Fig. 3. To promote a precise video-depth alignment, we propose a motion consistency loss $\mathcal{L}_{\text{mo}}$ and a cross-attention map consistency loss $\mathcal{L}_{\text{xattn}}$.

Motion consistency loss. In image diffusion models, the intermediate self-attention features in the U-Net are found to contain semantic information [5,16,66]. We found this finding also holds in our video-depth diffusion model. However, though video and depth features share similar layouts, they may differ in the temporal motion, as illustrated in Fig. 3. Thus, we attribute the video-depth misalignment to the unsynchronized motion between video and depth features. To address this problem, we propose to enforce a synchronized motion between the video and depth features via a motion consistency loss $\mathcal{L}_{\text{mo}}$.

Specifically, given intermediate video self-attention feature maps at frame l and $l+1$: $F_{\text{v},l}, F_{\text{v},l+1} \in \mathbb{R}^{H\times W\times D}$, their cost volume $C_{\text{v},l} \in \mathbb{R}^{H\times W\times H\times W}$ can be constructed by conducting cosine similarity between all point pairs on the two frames:

$$c_{\text{v},l,i,j,h,k} = f_{\text{v},l,i,j} \cdot f_{\text{v},l+1,h,k}, \tag{3}$$

where $f_{\text{v},l,i,j}$ is the normalized feature vector on $F_{\text{v},l}$ at spatial location i, j. The cost volume $C_{\text{d},r}$ for the depth feature maps can be computed in a similar way.

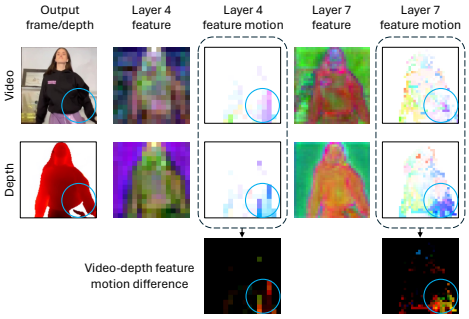

Fig. 3: Visualization of the video and depth feature maps and their motion fields without consistency losses. We attribute the inconsistent video-depth output (blue circle) to the inconsistent video-depth feature motions (the last row). This problem exists in multiples layers within the U-Net, and we randomly select two layers for visualization. We follow [66] to visualize the feature maps, and different color in the motion filed indicates different moving direction.

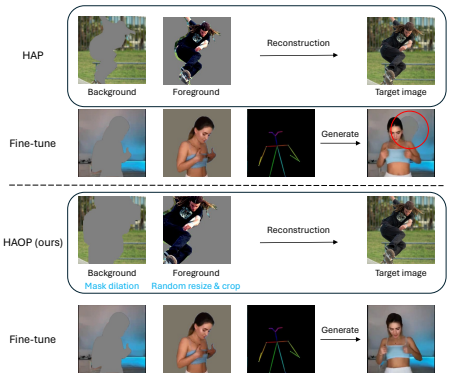

Fig. 4: Comparison between human attribute pre-training (HAP) [68] and our human attribute outpainting pre-training (HAOP). HAP may result in an apparent background mask when the target pose deviates from the original position, while our HAOP mitigates this problem.

As the video and depth features may distributed differently, we further normalize the cost volume via softmax, resulting in a motion field $U_{\mathrm{v},l}$:

$$u_{\mathrm{v},l,i,j,h,k} = \frac{\exp\left(c_{\mathrm{v},l,i,j,h,k}/\tau\right)}{\sum_{h'}\sum_{k'}\exp\left(c_{\mathrm{v},l,i,j,h',k'}/\tau\right)}, \tag{4}$$

where τ is a temperature hyper-parameter controlling the concentration of the distribution. A high value on the motion field indicates a movement between two points. The motion consistency is achieved by minimizing the MSE loss between the video and depth motion fields:

$$\mathcal{L}_{\mathrm{mo}} = \frac{1}{LHWHW}\sum_{l,i,j,h,k}\|u_{\mathrm{v},l,i,j,h,k} - u_{\mathrm{d},l,i,j,h,k}\|_2^2. \tag{5}$$

By enforcing similar motion of video and depth features, it promotes a consistent video and depth output.

Cross-attention map consistency loss. Except for leveraging the self-attention feature, the cross-attention maps between the foreground image and the input noise are also found to be critical in the image layout control [22, 49]. Thus, we additionally enforce the cross-attention maps from the video stream and the depth stream to be similar via an MSE loss $\mathcal{L}_{\mathrm{xattn}}$:

$$\mathcal{L}_{\mathrm{xattn}} = \|M_{\mathrm{v}} - M_{\mathrm{d}}\|_2^2, \tag{6}$$

where M_{v} and M_{d} represent the video and depth cross-attention map, respectively.

As the ControlNet fuses features in the U-Net up blocks, we apply the consistency loss terms only to the U-Net up blocks to efficiently learn the U-Net and the ControlNet. The overall training objective $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{denoise}} + \sum_{n=1}^{N} \left(w_{\text{mo}} \mathcal{L}_{\text{mo},n} + w_{\text{xattn}} \mathcal{L}_{\text{xattn},n} \right), \tag{7}$$

where w_{mo} and w_{xattn} are weighting hyper-parameters, N is the number of up blocks, and $\mathcal{L}_{\text{mo},n}$ and $\mathcal{L}_{\text{xattn},n}$ are the motion consistency loss and the cross-attention map consistency loss on the n-th up block, respectively.

3.3 Human attribute outpainting pre-training

To disentangle the pose and background control, DisCo [68] proposes a human attribute pre-training (HAP) task to learn human appearance. Specifically, HAP learns to reconstruct the whole image given human foreground and background images, with the pose ControlNet being removed, as shown in Fig. 4 top. However, we empirically find that such pre-training may lead to an apparent background mask when the target pose deviates from the original position.

To address this problem, we propose human attribute outpainting pre-training (HAOP). Our HAOP differs from HAP in two ways. First, HAOP extends the background mask through dilation, forcing the model to fill in the background around the foreground. Second, we apply a random crop and resize on the foreground image, prompting the model to extrapolate the partial human attributes. This process not only mitigates the background masking problem but also improves image synthesis, as demonstrated in Fig. 4 and validated in our ablation studies (Sec. 4.2). Formally, the pre-training learning objective is as follows:

$$\mathcal{L}_{\text{HAOP}} = \mathbb{E}_{v,\epsilon,t,f_{\text{aug}},b_{\text{aug}}} \left[\|\epsilon - \epsilon_\theta(z_t, t, f_{\text{aug}}, b_{\text{aug}})\|_2^2 \right], \tag{8}$$

where f_{aug} and b_{aug} represent augmented foreground and background, respectively. Note that the pre-training does not require depth denoising, which significantly lowers the data demand.

4 Experiments

Datasets. We train and evaluate on two public datasets: TikTok [27] and NTU120 [35,61]. The TikTok dataset consists of ∼350 human dancing videos. We follow DisCo [68] to use 335 videos for training, and use 10 videos for evaluation. The NTU120 dataset consists of daily activities videos, we select 588 videos for training and 72 videos for evaluation. The videos are cropped to center to the subject. We specifically ensure distinct subjects and backgrounds between training and evaluation, and select only certain subjects to better evaluate the generalization ability. Besides, we use Grounded-SAM [32, 36] for the human foreground mask estimation, and follow ControlNet [78] to use OpenPose [7]

Method	Motion control	TikTok				NTU120			
		Video		Depth	Image	Video		Depth	Image
		FID-FVD↓	FVD↓	L2↓	FID↓	FID-FVD↓	FVD↓	L2↓	FID↓
FOMM [62]	Target video	38.36	404.31	-	85.03	40.34	1439.50	-	80.29
MRAA [63]		24.11	306.49	-	54.47	58.19	1441.79	-	97.07
TPS [79]		29.20	337.79	-	53.78	37.42	1339.86	-	<u>61.75</u>
DreamPose [30]	DensePose [19]	52.62	614.07	-	75.08	80.11	791.25	-	116.23
DisCo [68]	OpenPose [7]	<u>20.75</u>	<u>257.90</u>	0.0975†	<u>39.02</u>	<u>26.21</u>	<u>458.92</u>	<u>0.0371</u>†	68.53
LDM3D [64]		45.30	553.03	0.0637	69.36	71.11	587.84	0.0650	120.74
MM-Diffusion [58]		48.92	771.32	<u>0.0367</u>	68.47	58.44	504.05	0.0404	102.77
IDOL	OpenPose [7]	**17.86**	**223.69**	**0.0336**	**36.04**	**20.23**	**314.82**	**0.0317**	**50.70**

Table 1: Quantitative results comparison between our IDOL and existing methods on the TikTok and NTU120 datasets. The ground truth depth is estimated via HDNet [27]. "-" indicates incapable of generating depth map, "†" indicates the HDNet-inferred depth from the synthesized image. The best and the second best results are denoted in **bold** and <u>underscored</u>, respectively.

for human pose estimation. For the depth estimation, we leverage two different methods: HDNet [27] for high-fidelity depth estimation on the human foreground, where the background depth is set to a constant; MiDaS [53, 54] for the whole-frame depth estimation, which lacks details on the human foreground. The HDNet depth map is rendered by applying the "hot" colormap; the MiDaS depth map is rendered as grayscale images. For pre-training, we follow DisCo [68] to use ∼700k images from a combined dataset of TikTok [27], COCO [34], SHHQ [14], DeepFashion2 [15], and LAION [60].

Evaluation metrics. We separately evaluate the quality of the generated video and depth. For video quality evaluation, we follow DisCo [68] to use FID-FVD [4] and FVD [67] over every 8-frame snippet. For the evaluation of depth synthesis, for simplicity, we first scale the depth in the range $[0, 1]$, and compute L2 distance between the ground truth depth[4] and the generated depth. In addition to evaluating the generated video and depth, we further use FID [23] on the frames to measure the image quality.

Comparison methods. We adopt state-of-the-art image animation methods FOMM [62], MRAA [63], TPS [79], DreamPose [30], and DisCo [68] for the video and image quality comparison. As a pioneering method in human-centric joint video-depth generation, to our knowledge, there is no direct prior work for the joint video-depth generation quality comparison. Thus, we adapt the most relevant multi-modal generation methods (LDM3D [64] for text to image-depth and MM-Diffusion [58] for text to video-audio), modify their backbones to the same VideoLDM baseline as ours (Sec. 3.1) to enable appearance and pose control, and compare the joint video-depth generation results with them.

[4] The depth evaluation is conducted on the synthesized depth and the depth estimated from the ground truth image instead of that estimated from the generated images. This is because we empirically find that the depth estimated from the generated image tends to be noisy (see Fig. 5 and Fig. 6).

Fig. 5: Qualitative results comparison between TPS [79], DisCo [68], LDM3D [64], MM-Diffusion [58] and our IDOL on the TikTok and NTU120 datasets with HDNet estimated depth [27]. Note that TPS [79] is unable to generate depth, and the depth map for DisCo is estimated via HDNet on its generated frames. Please find video comparison in the supplementary material.

4.1 Main results

In our experiments, we use the HDNet depth [27] by default for its high-fidelity human foreground depth estimation. We also train and evaluate on the MiDaS depth [53, 54] to demonstrate the generalization ability of our IDOL.

Comparison with the state-of-the-art. We compare IDOL with state-of-the-art methods on the TikTok and NTU120 datasets in Tab. 1, with the generated frame and depth visualized in Fig. 5. We make the following observations. (1) We outperform all competing motion transfer and video generation methods [30, 62, 63, 68, 79] in terms of both video and image quality. (2) When adapting multi-modal generation methods [58, 64] for human-centric joint video-depth generation, with the integration of the video LDM baseline (Sec. 3.1), IDOL consistently outperforms them across video, image, and depth metrics on both datasets, demonstrating the effectiveness of our method. (3) Applying HDNet depth estimation to the previous state-of-the-art DisCo generated frames leads to suboptimal results, with incomplete, fragmentary or overly simplistic estimations, as shown in Fig. 5. As the estimation result deviates from the ground truth, the HDNet results lead to a high depth L2, as listed in the DisCo row of Tab. 1. This finding indicates the limited generalization ability of monocular depth methods, and highlights the need for multi-modal generation approaches.

Comparison on synthesizing different types of depth map. Tab. 2 compares the results on using the MiDaS grayscale depth map [53, 54] for the joint video-depth generation. The results consistently show that our IDOL outperforms

	Method	Video		Depth	Image
		FID-FVD↓	FVD↓	L2↓	FID↓
TikTok	DisCo [68]	20.75	257.90	0.1758‡	**39.02**
	LDM3D [64]	42.09	529.43	0.0646	72.30
	MM-Diffusion [58]	52.37	715.28	0.1040	70.09
	IDOL	**19.01**	**216.96**	**0.0271**	39.76
NTU120	DisCo [68]	26.21	458.92	0.0695‡	68.53
	LDM3D [64]	77.04	591.03	0.0244	115.16
	MM-Diffusion [58]	73.47	503.53	0.0260	117.22
	IDOL	**20.56**	**439.63**	**0.0210**	**55.70**

Table 2: Quantitative comparison between our method and existing methods on the TikTok and NTU120 datasets with MiDaS estimated depth [53, 54]. "‡" indicates the MiDaS estimated depth from the synthesized image.

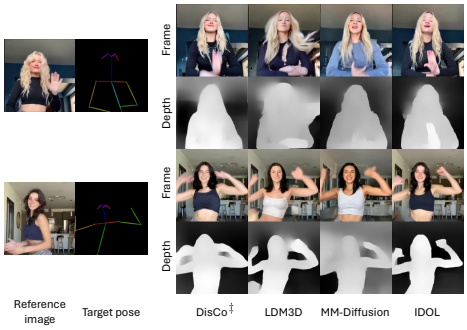

Fig. 6: Qualitative results comparison on MiDaS [53, 54] grayscale whole-frame depth map generation.

Motion repr.	Temporal modeling	Pose control	Video		Depth	Image
			FID-FVD↓	FVD↓	L2↓	FID↓
OpenPose [7]	Handcrafted	ControlNet [78]	17.86	223.69	0.0336	36.04
OpenPose [7]	AnimateDiff [20]	ControlNet [78]	19.58	201.83	0.0350	33.14
DWPose [76]	AnimateDiff [20]	ControlNet [78]	16.73	179.20	0.0245	31.06
DWPose [76]	AnimateDiff [20]	T2I-Adapter [44]	14.40	188.16	0.0195	30.06

Table 3: Quantitative result of our IDOL generalize to different motion representations, temporal modeling designs, and pose control module on TikTok [27] with HDNet depth [27]. The gray row indicates our default setting. Please find our supplementary material for additional implementation details.

existing multi-modal generation methods [58, 64] in both video and depth quality. This further highlights IDOL's effectiveness and its adaptability to different types of depth maps. Additionally, as shown in the DisCo row of Fig. 6, applying MiDaS on the DisCo generated frames also yields suboptimal results, blurring the areas surrounding the human subject, and leads to poorer depth accuracy (measured by L2 norm) than that achieved by generative approaches.

Generalization ability. Our IDOL follows DisCo [68] for setting the default motion representations (OpenPose [7]) and backbone architectures (ControlNet [78]). In Tab. 3 and Fig. 7, we further show that our IDOL can be easily adapted to other motion representations, temporal modeling modules, and pose control modules. First, IDOL can benefit from the motion prior from pre-trained AnimateDiff [20] to improve the video quality. Besides, our IDOL can be conditioned to more fine-grained DWPose [76], which provides additional facial and hand keypoints, to improve the video and depth quality. Impressively, even when using OpenPose, which lacks hand keypoint as condition, IDOL effectively generates plausible human hands, demonstrating its robust human attribute learning capabilities. Furthermore, our IDOL can be implemented with different pose control modules, such as ControlNet [78] and T2I-Adapter [44], achieving similar quantitative and qualitative results. Such results demonstrate that our IDOL can be generalized to different design options.

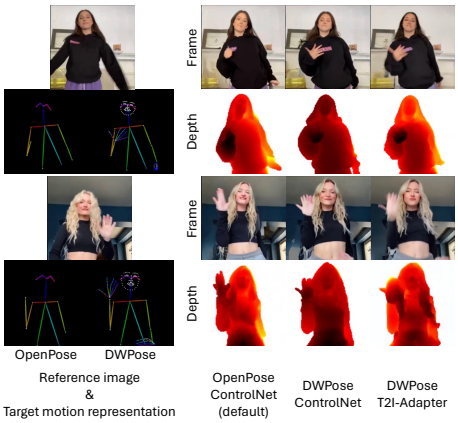

Fig. 7: Qualitative results of our IDOL generalize to different motion representations and pose control modules.

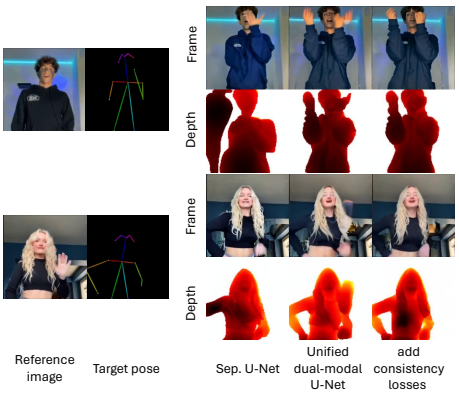

Fig. 8: Qualitative comparison between the baseline video LDM, unified dual-modal U-Net w/o and w/ consistency losses.

Settings	#Param.	Video FID-FVD↓	Video FVD↓	Depth L2↓	Image FID↓
Sep. U-Net for joint denoise	2×1.39B	24.28	282.50	0.0822	41.72
Share U-Net for joint denoise	1.39B	22.10	272.37	0.0369	39.43
+ Cross-modal attn.	1.41B	**19.28**	**260.65**	**0.0360**	**39.01**

Table 4: Ablation study on the joint video-depth learning and the unified dual-modal U-Net design.

$\mathcal{L}_{\text{xattn}}$	$\mathcal{L}_{\text{mo}}$	Video FID-FVD↓	Video FVD↓	Depth L2↓	Image FID↓
		19.28	260.65	0.0360	39.01
✓		19.99	244.58	0.0351	37.89
✓	✓	**17.86**	**223.69**	**0.0336**	**36.04**

Table 5: Ablation study on the video-depth consistency loss functions $\mathcal{L}_{\text{mo}}$ and $\mathcal{L}_{\text{xattn}}$.

4.2 Ablation studies

We conduct a set of ablation studies on the TikTok dataset [27] with HDNet depth [27] to demonstrate the effectiveness of our proposed method. These studies are conducted accumulatively, layering each component to assess its incremental impact on the overall performance.

Unified dual-modal U-Net for joint video-depth denoising. We analyze whether joint video-depth learning improves the performance, and the effectiveness of the designs proposed in the unified dual-modal U-Net in Tab. 4. The key observations are as follows. (1) Jointly video-depth learning with a shared U-Net is beneficial for both video and depth generation (the second row), while using only half the parameters compared to the separate counterpart. Such a result underscores the significance of our structurally-aware shared U-Net. (2) The explicit cross-modal information exchange between the video and depth denoising (the third row), *i.e.*, through cross-modal attention, further improves both the video and depth generation quality. Moreover, as shown in Fig. 8, our unified dual-modal U-Net markedly improves frame and depth quality compared to the separate U-Net baseline. Such results confirm the effectiveness of our design.

Learning video-depth consistency. In our IDOL, we introduce a motion consistency loss $\mathcal{L}_{\text{mo}}$ and a cross-attention map consistency loss $\mathcal{L}_{\text{xattn}}$ to enhance

Pre-training strategy	Video FID-FVD↓	Video FVD↓	Depth L2↓	Image FID↓
w/o pre-training	39.24	452.81	0.0363	56.80
HAP [68]	19.83	227.86	**0.0336**	38.71
HAOP (ours)	**17.86**	**223.69**	**0.0336**	**36.04**

Table 6: Ablation study on the pre-training.

Method	FLOPs (T)	#Param. (B)	GPU mem. (MB)	Infer. time (s)
LDM3D [64]	48.80	1.61	**7661**	42.02
MM-Diffusion [58]	41.02	2.78	14651	13.49
IDOL (ours)	**39.35**	**1.41**	10251	**12.23**

Table 7: Computational complexity comparison.

video-depth alignment. We analyze their impact in Tab. 5, where the results demonstrate that both of them improve the depth generation quality, and the motion consistency loss improves all video, depth, and image generation results. The final column of Fig. 8 further illustrates how these consistency loss terms effectively refine video-depth alignment in our model.

Human attribute outpainting pre-training. We analyze the effectiveness of our proposed human attribute outpainting pre-training (HAOP) in Tab. 6. The results reveal that HAOP surpasses human attribute pre-training (HAP) [68] across all metrics, confirming its efficacy. Notably, even without pre-training, our IDOL already outperforms other competing multi-modal generation methods [58, 64].

Complexity and computational requirements. We analyzed the complexity of generating an 8-frame video-depth sequence on a single V100 GPU, comparing our method with other multi-modal generation methods in Tab. 7. Thanks to our unified U-Net design, IDOL requires the fewest FLOPs, has the lowest number of trainable parameters, and the shortest inference time, demonstrating our efficiency.

5 Conclusion

In this paper, we propose IDOL, a framework tailed for human-centric joint video-depth generation. Our proposed unified dual-modal U-Net improves the video and depth synthesis by implicit video structure learning, with cross-modal attention explicitly bridging the joint video-depth denoising process. Our motion consistency loss and cross-attention map consistency loss promote spatial alignment between the generated video and depth. Extensive experiments on the TikTok and NTU120 datasets show our superior performance compared with existing methods, and the adaption ability to different kinds of depth maps. Our IDOL is also able to generalize to different motion representations and pose control modules.

Limitations. Despite the performance advantage of our IDOL, it faces several major limitations. First, the computational demands of processing dual-modal data, particularly at high resolutions, hinder its suitability for real-time applications, highlighting a need for further optimization. Additionally, the reliance on high-quality depth maps for training limits its applicability in scenarios where such data is limited or of low quality. Addressing this, future work may explore unsupervised methods or data augmentation strategies to mitigate the data quality constraint.

Negative societal impact. Our model raises ethical concerns, including the potential for creating deepfake videos, producing biased outputs, and threatening intellectual property rights. To mitigate these risks, we can incorporate invisible watermarks to ensure content authenticity.

Acknowledgements

This work is supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Atapour-Abarghouei, A., Breckon, T.P.: Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: CVPR. pp. 2800–2810 (2018)
3. Bae, J., Moon, S., Im, S.: Deep digging into the generalization of self-supervised monocular depth estimation. In: AAAI. vol. 37, pp. 187–196 (2023)
4. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional gan with discriminative filter generation for text-to-video synthesis. In: IJCAI. vol. 1, p. 2 (2019)
5. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
6. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
8. Chang, D., Shi, Y., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Yang, X., Soleymani, M.: Magicdance: Realistic human dance video generation with motions & facial expressions transfer. arXiv preprint arXiv:2311.12052 (2023)
9. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: ICCV. pp. 19830–19843 (2023)
10. Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. In: ICCV. pp. 909–919 (2023)
11. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS **34**, 8780–8794 (2021)
13. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: ICCV. pp. 7346–7356 (2023)

14. Fu, J., Li, S., Jiang, Y., Lin, K.Y., Qian, C., Loy, C.C., Wu, W., Liu, Z.: Stylegan-human: A data-centric odyssey of human generation. In: ECCV. pp. 1–19 (2022)
15. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: CVPR. pp. 5337–5345 (2019)
16. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
18. Gu, J., Trevithick, A., Lin, K.E., Susskind, J.M., Theobalt, C., Liu, L., Ramamoorthi, R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. pp. 11808–11826 (2023)
19. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR. pp. 7297–7306 (2018)
20. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
21. Ha, S., Kersner, M., Kim, B., Seo, S., Kim, D.: Marionette: Few-shot face reenactment preserving identity of unseen targets. In: AAAI. vol. 34, pp. 10893–10900 (2020)
22. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
25. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)
26. Hu, Z., Xu, D.: Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. arXiv preprint arXiv:2307.14073 (2023)
27. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: CVPR. pp. 12753–12762 (2021)
28. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. arXiv preprint arXiv:2303.17559 (2023)
29. Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., Xu, Q.: Humansd: A native skeleton-guided diffusion model for human image generation. arXiv preprint arXiv:2304.04269 (2023)
30. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. arXiv preprint arXiv:2304.06025 (2023)
31. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
32. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

33. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR. pp. 22511–22521 (2023)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
35. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE TPAMI **42**(10), 2684–2701 (2019)
36. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
37. Liu, X., Ren, J., Siarohin, A., Skorokhodov, I., Li, Y., Lin, D., Liu, X., Liu, Z., Tulyakov, S.: Hyperhuman: Hyper-realistic human generation with latent structural diffusion. arXiv preprint arXiv:2310.08579 (2023)
38. Luan, T., Li, Z., Chen, L., Gong, X., Chen, L., Xu, Y., Yuan, J.: Spectrum auc difference (saucd): Human-aligned 3d shape evaluation. In: CVPR. pp. 20155–20164 (2024)
39. Luan, T., Wang, Y., Zhang, J., Wang, Z., Zhou, Z., Qiao, Y.: Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In: AAAI (2021)
40. Luan, T., Zhai, Y., Meng, J., Li, Z., Chen, Z., Xu, Y., Yuan, J.: High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition. In: CVPR. pp. 16795–16804 (2023)
41. Ma, Y., He, Y., Cun, X., Wang, X., Shan, Y., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. arXiv preprint arXiv:2304.01186 (2023)
42. Mallya, A., Wang, T.C., Liu, M.Y.: Implicit warping for animation with image sets. NeurIPS **35**, 22438–22450 (2022)
43. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR. pp. 6038–6047 (2023)
44. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
45. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
46. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: ICCV. pp. 7184–7193 (2019)
47. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
48. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
49. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
50. Qian, S., Lin, K.Y., Wu, W., Liu, Y., Wang, Q., Shen, F., Qian, C., He, R.: Make a face: Towards arbitrary high fidelity face manipulation. In: ICCV. pp. 10033–10042 (2019)

51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. pp. 8748–8763 (2021)
52. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., et al.: Dreambooth3d: Subject-driven text-to-3d generation. arXiv preprint arXiv:2303.13508 (2023)
53. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
54. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE TPAMI **44**(3), 1623–1637 (2020)
55. Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., Curless, B.: Film: Frame interpolation for large motion. In: ECCV. pp. 250–266. Springer (2022)
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
57. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
58. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: CVPR. pp. 10219–10228 (2023)
59. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
60. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
61. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: CVPR. pp. 1010–1019 (2016)
62. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: CVPR (2019)
63. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: CVPR (2021)
64. Stan, G.B.M., Wofk, D., Fox, S., Redden, A., Saxton, W., Yu, J., Aflalo, E., Tseng, S.Y., Nonato, F., Muller, M., et al.: Ldm3d: Latent diffusion model for 3d. arXiv preprint arXiv:2305.10853 (2023)
65. Tang, H., Wang, W., Xu, D., Yan, Y., Sebe, N.: Gesturegan for hand gesture-to-gesture translation in the wild. In: ACM MM. pp. 774–782 (2018)
66. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR. pp. 1921–1930 (2023)
67. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
68. Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. In: CVPR (2024)
69. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. arXiv preprint arXiv:2210.04628 (2022)
70. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: ECCV. pp. 670–686 (2018)
71. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning. pp. 1336–1348. PMLR (2022)

72. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)
73. Wynn, J., Turmukhambetov, D.: Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In: CVPR. pp. 4180–4189 (2023)
74. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498 (2023)
75. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
76. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: ICCV. pp. 4210–4220 (2023)
77. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: ICCV. pp. 9459–9468 (2019)
78. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
79. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: CVPR (2022)
80. Zhao, S., Fu, H., Gong, M., Tao, D.: Geometry-aware symmetric domain adaptation for monocular depth estimation. In: CVPR. pp. 9788–9798 (2019)
81. Zheng, C., Cham, T.J., Cai, J.: T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In: ECCV. pp. 767–783 (2018)

A Additional experiments

Qualitative results. Please refer to the accompanying video for the qualitative comparison. In the video demonstration, we stitch each generated video snippets together to form a long video, which may lead to unnatural transitions between each snippet. We make the following observations. (1) Compared with DisCo [68], LDM3D [64], and MM-Diffusion [58], our generated videos and depth sequences exhibit more natural and smoother transitions, demonstrating the effectiveness of our method. A notable observation on the NTU120 dataset [35, 61] is LDM3D's limitation in generating diverse frames after fine-tuning the autoencoder, often resulting in repetitive "stuck" video sequences. This issue may stem from the subtle motions and predominant static content in NTU120 videos, suggesting that fine-tuning the autoencoder might necessitate a larger and more varied training dataset. (2) Our IDOL is able to composite different foreground and background, while simultaneously generating video and depth. This feature distinguishes it from concurrent methods [8, 25, 74], offering a unique capability in the area of video-depth synthesis. (3) Our method is able to generalize to different pose conditions, such as OpenPose [7] and DWPose [76].

Cross-attention map consistency. To enhance video-depth alignment in IDOL, we propose a cross-attention map consistency loss $\mathcal{L}_{\text{xattn}}$. This loss function encourages alignment of the video and depth cross-attention maps. We also explore alternate ways to align the cross-attention maps. One straightforward approach is to use a shared cross-attention map for both branches. We test two variations: replacing individual cross-attention maps with their average, and using the video stream's cross-attention map as a substitute for both. Our results in Tab. 8 reveal that sharing a cross-attention map significantly reduces performance, particularly impacting depth L2 accuracy (as seen in rows 2 and 3). These findings highlight the need for each stream to maintain diverse cross-attention maps to produce high-quality outputs. Our implementation of $\mathcal{L}_{\text{xattn}}$ effectively balances the need for consistency with the preservation of each map's unique characteristics, ultimately contributing to superior overall performance.

Setting	Video		Depth	Image
	FID-FVD↓	FVD↓	L2↓	FID↓
-	19.28	260.65	0.0360	39.01
Share cross-attention map (avg)	20.82	297.76	0.0706	49.66
Share cross-attention map (video)	20.00	253.73	0.0718	44.58
Apply $\mathcal{L}_{\text{xattn}}$	**19.99**	**244.58**	**0.0351**	**37.89**

Table 8: Ablation study on the cross-attention map operations on the TikTok dataset [27] with HDNet depth [27].

B Implementation details

Our code is developed based on diffusers [47]. We follow DisCo [68] to use Stable Diffusion v1.4 [56] as the backbone. For HAOP pre-training, we follow DisCo [68] to freeze the ResBlocks and train the model for 25k steps, with input image size 256×256 and learning rate $1e^{-3}$. For fine-tuning, we adopt a two-stage approach. In the first stage, the temporal layers are removed, and the framework is trained on joint image-depth denoising. In the second stage, the whole framework with temporal modules is trained for the joint video-depth denoising. Both the first and the second stages are trained for 15k steps with a learning rate of $1e^{-4}$. For the second stage, the model is trained on 8-frame sequences. Both the pre-training and fine-tuning are conducted on 32 V100 GPUs. The weight hyper-parameters are set via a grid search: $w_{\text{mo}} = w_{\text{xattn}} = 0.01$. We set the temperature term τ in the motion field computation to $1/\sqrt{D_n}$, where D_n is the number of channels in the n-th layer.

Comparison methods. We use DisCo [68], a recent diffusion-based human dance video generation method, as a strong human-centric video generation baseline. As a pioneering method directly tailed for human-centric joint video-depth generation, we compare our IDOL with the closest multi-modal generation counterparts. We choose MM-Diffusion [58], initially designed for text-to-video-audio synthesis, and LDM3D [64], aimed at text-to-image-depth generation. To facilitate a fair comparison, we align their backbones to the same video LDM baseline, adapting them for the human-centric video-depth task. For MM-Diffusion [58], we replace the audio U-Net with a duplicate of the video U-Net (without sharing parameters, unlike in our IDOL) and retain the rest of the structure unchanged. In the case of LDM3D [64], we inflate the 2D U-Net to a 3D U-Net to accommodate video generation. Both adapted methods employ ControlNet [78] for human pose control and process background and foreground inputs similarly to IDOL, ensuring consistency in our comparative evaluation.

Generalization to different designs. In our main manuscript, we evaluated the generalization ability of our IDOL to different designs, including DWPose [76], AnimateDiff [20], and T2I-Adapter [44]. For the adaptation of AnimateDiff [20], we remove the original temporal convolutional and attention layers in the 3D U-Net, and insert the AnimateDiff [20] pre-trained motion modules. For the T2I-Adapter [44], we replace the original pose ControlNet with a pre-trained OpenPose T2I-Adapter. Note that the video and depth streams share the same pose T2I-Adapter, similar to the pose ControlNet.