# MagicPose4D: Crafting Articulated Models with Appearance and Motion Control

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

With the success of 2D and 3D visual generative models, there is growing interest in generating 4D content. Existing methods primarily rely on text prompts to produce 4D content, but they often fall short of accurately defining complex or rare motions. To address this limitation, we propose **MagicPose4D**, a novel framework for refined control over both appearance and motion in 4D generation. Unlike traditional methods, MagicPose4D accepts monocular videos as motion prompts, enabling precise and customizable motion generation. MagicPose4D comprises two key modules: **(i) Dual-Phase 4D Reconstruction Module** which operates in two phases. The first phase focuses on capturing the model's shape using accurate 2D supervision and less accurate but geometrically informative 3D pseudo-supervision without imposing skeleton constraints. The second phase refines the model using more accurate pseudo-3D supervision, obtained in the first phase and introduces kinematic chain-based skeleton constraints to ensure physical plausibility. Additionally, we propose a Global-local Chamfer loss that aligns the overall distribution of predicted mesh vertices with the supervision while maintaining part-level alignment without extra annotations. **(ii) Cross-category Motion Transfer Module** leverages the predictions from the 4D reconstruction module and uses a kinematic-chain-based skeleton to achieve cross-category motion transfer. It ensures smooth transitions between frames through dynamic rigidity, facilitating robust generalization without additional training. Through extensive experiments, we demonstrate that MagicPose4D significantly improves the accuracy and consistency of 4D content generation, outperforming existing methods in various benchmarks.

## 1 Introduction

The 4D generation task involves creating a temporal sequence of 3D models of moving objects. Given the difficulty of the general problem, recent approaches have made use of pretrained models, selected using prompts to convey user's intentions. Recently, there has been a significant focus on the use of text/image-prompts to describe appearance and motion [16, 22, 40]. The general pipeline of such existing consists of two steps: (i) acquiring static geometry through a 3D generation (e.g., a text/image-to-3D) model [17], which generates 3D representation such as meshes/implicit fields according to text/image prompts, and (ii) obtaining motion information via a video generation model [7], which generates video according to text/image prompts. This approach has achieved impressive 4D content results.

However, challenges lie in facilitating users to freely and precisely specifyarticulated motion of a non-rigid 3D object, and the generation faithfully reflecting the prompts, in terms of accurately capturing desired object appearance, geometry and motion. The main issues with current methods are the following: **(i) Temporal inconsistency in 3D geometry:** Most current video generation models

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

fail to ensure temporal consistency of 3D geometry. This involves additional complexity when objects are articulated, because the 3D shape and movement must change mutually consistently during motion. For instance, while using the existing methods for 4D animal generation, we have observed unnatural and implausible configurations, such as the number of limbs of the object changing across different frames. **(ii) Limited to common motions:** Current video generation models perform well in generating simple, subtle, and common actions, such as walking or small motions like shaking, but they struggle to satisfactorily generate more complex motions, e.g., involving large or uncommon movements such as *"King Kong dancing Hip-Hop"* and *"pig running like a rabbit"* as shown in Fig.6. **(iii) Text is inadequate for accurately describing the details of motion:** As shown in Tab.4, most existing methods use text as prompts to describe the desired motion, with some efforts [3] allowing trajectory (root body transformation) control. However, they are unable to specify detailed motions, e.g., by showing a real-world animal's movement (say, via a monocular short video).

To address these issues, we propose **MagicPose4D**, which enables detailed control over both appearance and motion, and temporal consistency. Unlike existing methods, which rely mostly on text descriptions as motion prompts and for which it is difficult to convey complex or rare motions, our approach accepts monocular videos or dynamic mesh sequences as motion prompts. This allows for more precise supervision and faithful 4D generation. MagicPose4D introduces two key modules: **Dual-Phase 4D Reconstruction Module** and **Cross-category Motion Transfer Module**. The first module estimates a sequence of 3D models of the object while also simultaneously estimating motion parameters from the motion prompts, and the second module transfers this motion to the target object, which is generated by a 3D generation model controlled by appearance prompts.

**Dual-Phase 4D Reconstruction Module**: Given the complexity, particularly due to articulated objects, we divide this module into two phases. First, we use image descriptors (e.g., segments, flow) as 2D supervision to estimately relatively less accurate but geometrically informative 3D models to serve as geometric priors (otherwise absent in the video) for pseudo supervision of 3D reconstruction. This phase does not impose constraints emanating from the articulated nature of the object, e.g., captured in the underlying skeletal structure and plausible changes in it during motion, allowing each part to learn arbitrary rotations ($\mathbf{R}$) and translations ($\mathbf{t}$), while focusing on learning the model's shape. In the second phase, we ensure physical plausibility of the motion by enforcing kinematic chain-based skeletal movement constraints. Additionally, we propose a Global-local Chamfer loss, which ensures that the overall distribution of predicted mesh vertices aligns with the supervision while maintaining part-level alignment, without requiring additional annotations. **Cross-category Motion Transfer Module**: This module achieves cross-species motion transfer by mapping the skeleton of one species to another by establishing joint and limb correspondences through a kinematic-chain-based representation of the skeleton. Our motion transfer module is non-training-based. It helps improve generalization and prevents the poor performance seen in many existing methods when tested on data with significant gaps from the training set. Also, we leverage dynamic rigidity [38] to guarantee smoothness between frames, unlike the existing approaches which perform frame-independent pose transfer.

The following are the main contributions of this paper: **(i) A Novel 4D Generation Framework:** Our new framework leverages monocular videos as motion prompts, providing more accurate and more precisely specifiable 4D action generation. **(ii) Skeleton Based 4D Representation:** By using skeleton-mediated geometric and 3D prior, we achieve more accurate motion estimation and 3D reconstruction, improving the physical plausibility of the generation. **(iii) Global-local Chamfer Loss**: We introduce a novel loss function to better align estimated mesh vertices with the supervisory 3D model overall while maintaining part-level alignment, without additional annotations. **(iv) Cross-category Motion Transfer**: Our cross-category mapping in terms of skeleton based dynamic rigidity representation enables smooth transitions between frames and robust generalization without additional training. **(v) Outperforms SOTA:** Through extensive experiments, we demonstrate that MagicPose4D provides highly accurate 4D content and significantly outperforms existing methods for 4D reconstruction and pose transfer across all three benchmarks that we experiment with.

## 2 Related Work and Motivation

**4D/3D Reconstruction.** Significant advancements in 4D/3D reconstruction include the development of specialized parametric models such as SCAPE [1] and SMPL [18] for human bodies, MANO [23] for hand movements, FLAME [14] and EMOCA [9] for facial expressions, and SMAL [42] for

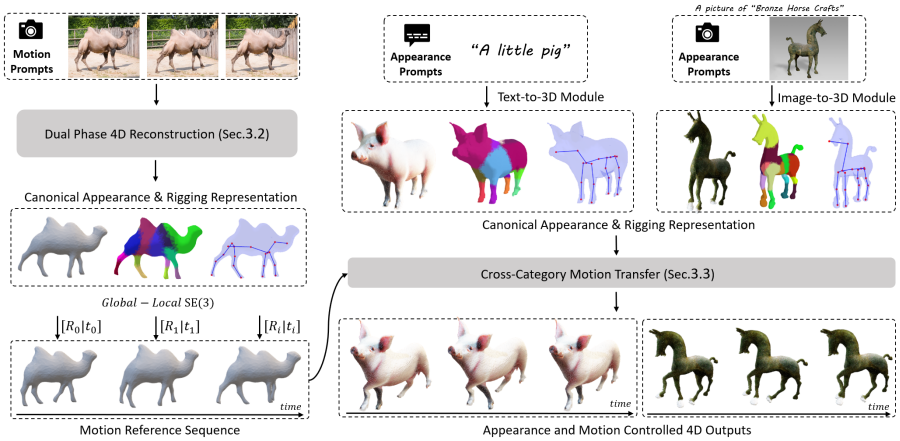



Figure 1: Overview of **MagicPose4D**, which takes motion prompts (monocular video or dynamic mesh sequence) and appearance prompts (text or image) to control the 4D content generation.

quadruped animals. However, these methods require predefined parametric models that include skeletons and skinning weights, limiting their generalization to uncommon species and out-of-distribution data. Recent neural implicit-based methods [20, 34, 11, 38, 31] offer promising alternatives by jointly learning static 3D meshes and time-varying parameters without predefined templates. However, they often fail with monocular videos containing sparse views, limiting their practical applicability. To address this, MagicPose4D introduces an innovative Dual-Phase 4D Reconstruction module, which reduces the requirements from multi-view videos to single-view videos and achieves robust reconstruction.

**Diffusion-based 4D Generation and Pose Transfer.** Recent diffusion-based 4D generation methods have shown promising results by leveraging text-to-3D models followed by text-to-video supervision. These techniques [24, 22, 16, 4, 41, 10, 36, 12, 40] have improved the geometry and appearance of generated models but are generally limited to minor movements within a fixed location and rely on text prompts, making precise motion control difficult. Our goal is to create dynamic 4D animations that closely transfer the motion of any given reference real-world object. For pose transfer, recent skeleton-based frameworks [25, 15] have explored the use of rigging points and key points. However, learning-based methods often struggle with in-the-wild data and significant domain gaps between identities. MagicPose4D introduces a cross-category motion transfer module that supports cross-species transfer while ensuring generalization and maintaining temporal smoothness.

# 3 MagicPose4D

MagicPose4D accepts two distinct types of input prompts: (i) appearance prompts and (ii) motion prompts. Consistent with recent methods [8, 27], both images and textual descriptions can function as appearance prompts, delineating the desired object and its visual characteristics. In a departure from existing approaches, MagicPose4D enables users to specify precise motions and trajectories by providing a video/mesh sequence that represents the anticipated movement.

As illustrated in Fig 1, MagicPose4D comprises three critical components: (i) the 4D Reconstruction module (Sec.3.2), (ii) the Cross-Category Motion Transfer module (Sec.3.3), and (iii) the Image-to-3D Generation module. Each module is tailored to facilitate distinct aspects of dynamic modeling, enabling adaptive 4D reconstructions that align with user-defined specifications.

## 3.1 Terminology and Overview

To represent an animated 3D model, our method learns static representations, such as the visible canonical shape $\mathbf{S} \in \mathbb{R}^{N\times 3}$, the underlying skeleton $\mathbf{S_k} = \{\mathbf{J} \in \mathbb{R}^{J\times 3}, \mathbf{B_s} \in \mathbb{R}^{B}, \mathbf{P_i} \in \mathbb{R}^{B}\}$, and the skinning weights $\mathbf{W} \in \mathbb{R}^{N\times B}$. Additionally, it captures time-varying parameters, including the global-local (root body-bone) transformations $\tau = \{\tau_0^t, \tau_1^t, \ldots, \tau_B^t\}$ and the camera parameters $\mathbf{P_c}$. Here, $B$, $N$, and $J$ represent the number of bones, vertices on the mesh surface, and joints, respectively. The transformations $\tau_i^t \in SE(3)$ include $\tau_0^t$ for the root body, with the remaining $\tau_i^t$

for the bones. The skeleton topology is described by $\mathbf{J}$ (joint coordinates), $\mathbf{B_s}$ (bone scale), and $\mathbf{P_i}$ (parent indices of joints).

As illustrated in Fig 1, utilizing a short monocular video as a motion prompt, we introduce a dual-phase 4D reconstruction module (Sec.3.2) designed to predict a sequence of 4D meshes as motion references. Appearance prompts may consist of text descriptions, images, or a monocular video. Depending on the type of input, a corresponding module—text-to-3D, image-to-3D, or dual-phase 4D reconstruction is employed to generate the static 3D representation of the desired object. Subsequently, this static representation along with the motion reference are fed into the cross-category motion transfer module (Sec.3.3). This module adeptly transfers motions from the reference to the target object while ensuring temporal smoothness and consistency.

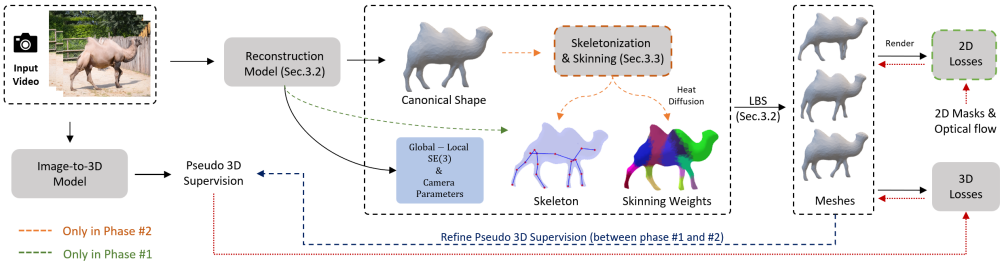



Figure 2: Overview of **Dual-Phase 4D Reconstruction Module**.

### 3.2 Dual-Phase 4D Reconstruction from video

Reconstructing 4D models from short monocular videos is a challenging task as it requires jointly learning numerous parameters, resulting in an extensive optimization space. Erroneous predictions of any parameter can trigger cascading failures. To address these challenges, we propose a Dual-Phase 4D Reconstruction module, which employs differentiated supervision across two distinct phases and focuses on learning diverse representations.

In the first phase, the primary focus is on accurately capturing the external appearance (shape) of the model. The underlying skeleton serves merely as an intermediary variable, utilizing non-kinematic chains skeleton and learnable skinning weights to afford the skeleton greater deformation freedom. This approach accelerates the learning of correct shapes. In contrast, the second phase aims to derive a more physically plausible motion reference for effective motion transfer. Therefore, we adopt kinematic chain skeletons and heat diffusion-based skinning weights, which narrow the deformation space of the skeleton, thereby ensuring the plausibility of the internal structure. (Sec.3.2.1)

From a supervision perspective, the first phase blends 2D and pseudo-3D supervision, updating the pseudo-3D supervision at the end of this phase. In the second phase, the 2D loss is removed, relying solely on the updated 3D pseudo-supervision to guide the learning process. (Sec.3.2.2)

#### 3.2.1 Model Articulation

**Skinning Weights** $\mathbf{W}$ is designed to represent the probabilities that each vertex corresponds to $B$ semi-rigid parts. **In the first phase**, following [31, 34], the skinning weights are modeled by the mixture of $B$ Gaussian ellipsoids as: $\mathbf{W}_{n,b} = \mathcal{F}e^{-\frac{1}{2}(\mathbf{X}_n-\mathbf{C}_b)^T\mathbf{Q}_b(\mathbf{X}_n-\mathbf{C}_b)}$, where $\mathcal{F}$ is the factor of normalization and precision matrices $\mathbf{Q}_b = \mathbf{V}_b^T\mathbf{\Lambda}_b\mathbf{V}_b$. In each Gaussian ellipsoid, $\mathbf{C} \in \mathbb{R}^{B\times 3}$ denotes Gaussian centers, $\mathbf{V} \in \mathbb{R}^{B\times 3\times 3}$ defines the orientation and $\mathbf{\Lambda} \in \mathbb{R}^{B\times 3\times 3}$ denotes the diagonal scale matrix. $\mathbf{X}_n$ is the 3D location of vertex n. **In the second phase**, as shown in Fig.3, a skeletonization module, is leveraged to obtain a skeleton for the canonical mesh, and following [29, 5], skinning weights $\mathbf{W}$ are obtained by a heat diffusion process. This approach guarantees a more natural assignment of skinning weights to the bones, enhancing the realism of the skeletal animations.

**Blend Skinning.** The mapping from surface vertex $\mathbf{X}_n^0$ in canonical space (time 0 by default) to $\mathbf{X}_n^t$ at time t in camera space is designed by blend skinning. The forward blend skinning is shown below:

$$\mathbf{X}_n^t = \tau_0^t(\sum_{b=1}^{B}\mathbf{W}_{n,b}\tau_b^t)\mathbf{X}_n^0, \tag{1}$$

including the number of bones $B$, skinning weights $\mathbf{W}$ and the transformation $\tau^t = \{\tau_b^t\}_{b=0}^B$ at time $t$, where $\tau_0^t$ represents the root body transformation and $\tau_{b>0}^t$ describes the bone transformation. Each $\mathbf{X}_n^0$ is first transformed by the weighted sum of each bone transformation $\mathbf{T}_{b>0}^t$ and then transformed by root body transformation $\mathbf{T}_0^t$ to achieve $\mathbf{X}_n^t$.

**Skeleton Articulation.** As described above, bone and root body transformations $\tau$ are crucial in the process of blending skinning. **In the first phase**, we define these transformations as independently learnable $SE(3)$ transformations. Practically, this is implemented by initializing learnable quaternions to compute the rotation matrices, along with defining learnable translations to achieve affine transformations. During this phase, we impose no constraints on the skeleton; each bone can rotate and translate freely. These bones serve merely as intermediate variables, with the ultimate objective of accurately determining the correct shape. **In the second phase**, we define per-frame joint angles $Q$, each describing the pose between a bone and its parent with three degrees of freedom. Instead of directly learning bone transformations, we determine $Q$ and compute the bone transformations $\tau_{b>0}^t$ using forward kinematics. The deformed mesh is then obtained by blend skinning.

**Extendable Bones.** It is noteworthy that kinematic-chain-skeleton-based methods rest on a fundamental assumption that the structure of the utilized skeleton closely mirrors the natural skeletal architecture of animals. This allows for the desired deformed mesh to be achieved by manipulating the skeleton's pose and employing blend skinning techniques. However, in practice, this assumption often does not hold completely. Animal bones are not solely rigid hinge connections; they include extendable cartilaginous tissues that lead to non-rigid deformations among joints. To address this, we introduce the concept of extendable bones. We allow for slight variations in bone length between frames, achieved by learning a time-varying scale parameter for each bone $\mathbf{B_s} \in \mathbb{R}^B$. This modification enhances the flexibility and realism of our skeletal model.

### 3.2.2 Supervision and Losses

As shown in Fig. 2, given the input from a monocular video, we utilize segmentation models [13, 37] and optical flow prediction models [28] to obtain 2D supervision. Furthermore, we employ an image-to-3D model to independently predict meshes for each frame, serving as pseudo-3D supervision. A question arises: *"Why not directly use an image-to-3D model to independently predict 3D meshes for each frame?"* As illustrated in Fig.7 (c), objects always exhibit self-occlusion in the video, under which circumstances image-to-3D models typically fail to produce accurate results. For instance, some frames might only depict a camel with two visible legs, resulting in a 3D mesh sequence that does not effectively capture the action information portrayed in the video. Moreover, as demonstrated in Video 3 of the supplementary materials, meshes generated directly from the image-to-3D module are independent of one another, thus lacking temporal continuity and smoothness. Although the initial pseudo-3D supervision may not be very accurate, it still provides valuable geometric priors that help address the information loss caused by insufficient perspectives of the object in the video.

In the first phase, we blend both 2D and 3D supervision to optimize the mesh shape and leverage a reconstruction loss, which consists of silhouette loss, optical flow loss, texture loss, perceptual loss, smooth, motion, and symmetric regularizations, and global-local chamfer (GLC) Loss. In the second phase, we only leverage GLC Loss and regularization terms without 2D losses. The details of the 2D loss functions and regularizations will be described in the Appendix.6.5.

**Global-Local Chamfer Loss (GLC).** The objective of GLC loss is to ensure that the predicted mesh closely resembles the expected mesh in **(i) overall shape** and in terms of their **(ii) respective poses**. This dual focus helps achieve high fidelity in both the structural and positional accuracy of the meshes. The GLC loss is the sum of chamfer distances across two levels. Initially, it involves the computation of the chamfer distance for the entire predicted mesh $\mathbf{S}$ and Pseudo mesh $\hat{\mathbf{S}}$ as follows:

$$\mathcal{L}_{\text{global}}(\mathbf{S}, \hat{\mathbf{S}}) = \frac{1}{|\mathbf{S}|} \sum_{x \in \mathbf{S}} \min_{y \in \hat{\mathbf{S}}} \|x - y\|^2 + \frac{1}{|\hat{\mathbf{S}}|} \sum_{y \in \hat{\mathbf{S}}} \min_{x \in \mathbf{S}} \|x - y\|^2. \tag{2}$$

The second level involves computing the weighted chamfer distances between $B$ parts. First, we calculate skinning weights $\mathbf{W}, \hat{\mathbf{W}} \in \mathbb{R}^{N \times B}$ for $\mathbf{S}$ and $\hat{\mathbf{S}}$ via the heat diffusion process [29, 5]. Then, we perform an *argmax* across the $B$ dimension to achieve a part-wise decomposition of the whole body into $K$ parts, where $K$ is less than or equal to $B$. In practice, $K$ often equals $B$, but in cases where $K$ is less than $B$, the loss computation simply omits the non-existent parts.

Subsequently, we calculate the part-level Chamfer distances between the $K$ pairs from predicted mesh $\mathbf{P}_k, k \in \{0, ..., K-1\}$ and Pseudo mesh $\hat{\mathbf{P}}_k$, as shown in the following equation:

$$\mathcal{L}_{\text{local}}(\mathbf{S}, \hat{\mathbf{S}}) = \frac{1}{|K|} \sum_{k=0}^{K-1} \frac{1}{|\mathbf{P}_k|} \sum_{x \in \mathbf{P}_k} \mathcal{W} \min_{y \in \hat{\mathbf{P}}_k} \|x - y\|^2 + \frac{1}{|\hat{\mathbf{P}}_k|} \sum_{y \in \hat{\mathbf{P}}_k} \mathcal{W} \min_{x \in \mathbf{P}_k} \|x - y\|^2, \quad (3)$$

where $\mathcal{W} = \mathbf{W}_{x,k} \times \hat{\mathbf{W}}_{y,k}$ represents the multiplication of skinning weights between vertices $x$ and $y$ for bone $k$. $\mathcal{W}$ serves as a rough key points estimator based on skinning confidence. It assigns greater weight to vertices at the central part of a segment and lesser weight to vertices at the junctions of multiple parts. This approach effectively addresses inconsistencies at the edges of part decompositions when calculating skinning weights using heat diffusion for models in various poses.

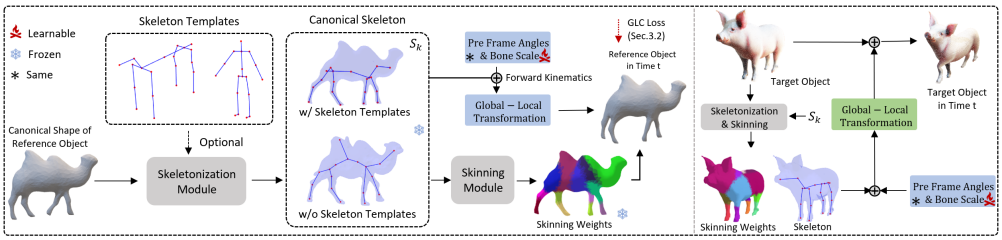



Figure 3: Overview of **Cross-Category Motion Transfer Module**.

### 3.3 Cross-Category Motion Transfer

In the second phase of 4D Reconstruction, as shown in Fig. 3(a), we extract the underlying skeletal motion of the reference meshes. Given a sequence of meshes for the reference object, the skeletonization module extracts the skeleton of the canonical shape, which is defined as the shape corresponding to time $= 0$. We fix the canonical skeleton and the skinning weights, thus controlling the skeleton's pose solely by learning the pre-frame angles and bone scales. Subsequently, by employing blend skinning, we obtain the deformed mesh.

The input to the skeletonization module is a mesh, along with an optional skeleton template. When the reference object is a commonly recognized form, such as a quadruped or a human, we utilize the corresponding skeleton template and embed it into the mesh. Following the methodology described in [5], we construct a coarse discrete representation to locate the approximate position of the skeleton within the internal space of the mesh. This process involves embedding the skeleton into a graph derived from the character's internal volume and determining an optimal solution using the A* algorithm across all possible matches. When a template of the object is not available, we employ skeleton extraction methods [2, 38] to extract the canonical skeleton of the reference object. After learning the reference motion represented by the pre-frame angles and the bone scale of the skeleton, we can readily compute the global-local transformation of the target object using forward kinematics, as illustrated in Fig. 3(b). We extract the canonical skeleton for the target object by inputting the canonical skeleton of the reference object, adhering to the method previously described following [5]. Subsequently, blend skinning is employed to generate the deformed meshes.

## 4 Experiments

This section primarily covers the following parts: (i) 4D reconstruction results and comparisons; (ii) motion transfer results and comparisons; and (iii) our framework v.s. 4D generation. Detailed descriptions of the benchmarks (Sec.6.2), implementation (Sec.6.3), and as shown in Tab.5, **more than** 60 **video results** (Sec.6.1) are provided in the Appendix.

### 4.1 Qualitative and Quantitative Comparison

**4D Reconstruction.** As depicted in Fig.5, we first compare the results of MagicPose4D with LASR [32] and BANMo [35]. These methods fall short of achieving ideal reconstruction, due to a lack of structural information about the objects, since they learn from only a single monocular video,


Figure 4: **Appearance and Motion Controlled 4D Generation.**

which contains sparse views. NeRF-based methods such as BANMo [34] and MagicPony [30] require particularly dense views from multiple cameras or long videos to achieve decent results. However, MagicPose4D achieves good reconstructions from only sparse views.

Furthermore, for a comprehensive quantitative comparison, we examine the performance of MagicPose4D against the best existing methods, LASR [32] and ViSER [33], which perform best (in terms of 2D keypoint transfer accuracy) among existing methods using a single monocular video as input. As shown in Tab.1, MagicPose4D consistently outperforms both LASR and ViSER for all animal subjects with a notable margin. When we run the author-provided codes for LASR, and ViSER on DAVIS, and PlanetZoo, we find that their results are highly variable across different runs. The results we report for these cases are therefore the mean accuracies we have obtained over multiple runs of all three methods.

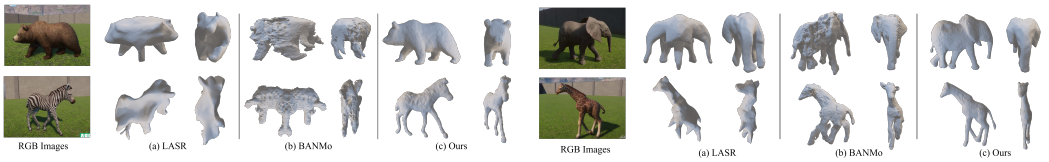
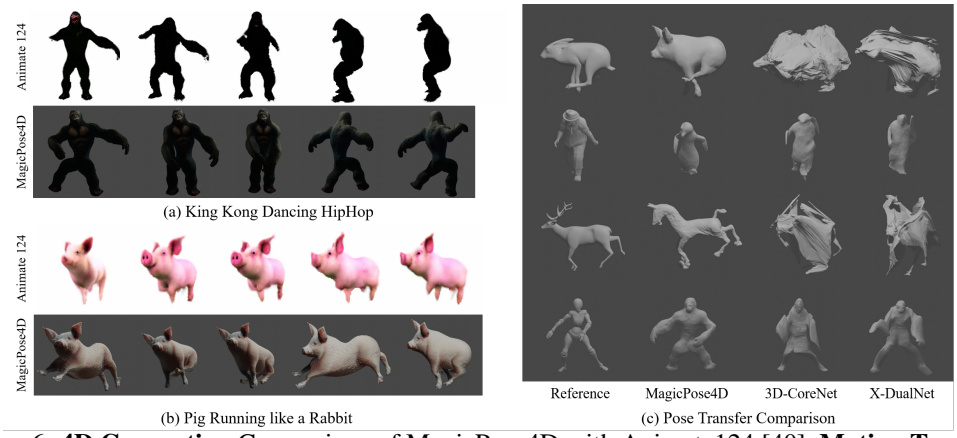



Figure 5: **4D Reconstruction Results.** We show the mesh reconstruction results of (a) LASR, (b) BANMo, and (c) Ours in the PlanetZoo's `bear, zebra, elephant, and giraffe`.

**Cross-Category Motion Transfer.** We present motion transfer results from MagicPose4D in Fig. 4. The target identities and reference motion sequences can either be humanoid or animal subjects. MagicPose4D is able to learn and retarget the motion from dynamic mesh prompts, e.g., Hip-Hop Dancing, or monocular video prompts, e.g., Hands Up.

We further compare the motion transfer ability for mesh generation quality with recent methods 3D-CoreNet [25] and X-DualNet [26] in Fig. 6 and Tab.2 (c). Both methods use a deep neural network to learn latent shape codes to retarget the motions and are trained/evaluated on SMPL [19](humanoid) and SMAL [42](animal). Since they do not include disentangled components or shape deformation




(a) King Kong Dancing HipHop (b) Pig Running like a Rabbit (c) Pose Transfer Comparison

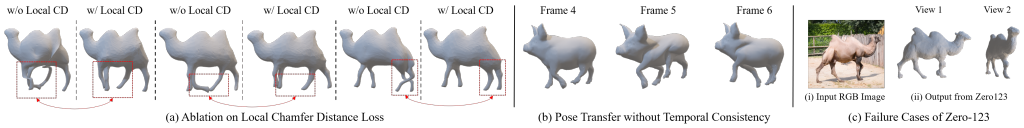
Figure 6: **4D Generation** Comparison of MagicPose4D with Animate124 [40], **Motion Transfer** Comparison of MagicPose4D with 3D-CoreNet [25] and X-DualNet [26]. Videos are in Sec.6.1.

Table 1: **2D Keypoint Transfer Accuracy.** This table presents the 2D keypoint transfer accuracy on DAVIS and PlanetZoo datasets. For classes from DAVIS (`camel`) and PlanetZoo (`giraffe, tiger`), we carry out multiple executions and report the mean of the observed accuracy in each case. Note that the results for LASR and ViSER here are from our executions of the codes provided by the authors.

| Method | DAVIS | | | | | PlanetZoo | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | camel | cow | dog | bear | dance-twirl | giraffe | tiger | elephant | bear | zebra | deer | Ave. |
| ViSER | 71.7 | 73.7 | 65.1 | 72.7 | 78.3 | 51.2 | 68.4 | 68.9 | 60.3 | 55.7 | 57.3 | 65.8 |
| LASR | 75.2 | 80.3 | 60.3 | 83.1 | 55.3 | 56.3 | 70.4 | 69.5 | 63.1 | 57.4 | 60.3 | 66.5 |
| MagicPose4D | **78.9** | **83.1** | **67.9** | **86.2** | **84.3** | **62.9** | **73.3** | **70.6** | **67.1** | **60.2** | **61.9** | **72.4** |

modules to represent the structural information of 3D meshes, these methods rely heavily on pre-processed training data with high-quality mesh annotations. Due to these constraints, these baselines cannot generalize well to in-the-wild target identities or reference motions. In contrast, MagicPose4D generates temporally consistent and smooth motions, while strictly preserving the identity and appearance of the target mesh. It is also worth noting that because previous methods focus only on pose transfer without considering the motion trajectory (pose sequence), the generated object always stays at the same position. This can be observed in side-by-side video comparisons in the supplementary materials Sec.6.1.

**Our Framework v.s. Existing 4D Generation.** Text/image-prompts-based 4D generation has been a popular trend. To compare with a representative recent work Animate124 [40], we feed the reference image, which provides the identity, and text prompt, which describes the motion of the generated object, into their model and optimize the NeRF with SDS-loss. In our case, we use the reference image to generate the target mesh with an Image-to-3D model and transfer the motion of the reference mesh sequence. We visualize the rendered video from different viewpoints from Animate124 and compare it with MagicPose4D in Fig. 6(a) and Fig. 6(b). More video comparisons are in the supplementary material. MagicPose4D provides more temporally consistent and smooth generation with the learning of motion deformation. Since there are no ground truth mesh sequences to evaluate the generation, we provide a comprehensive user study for Motion Transfering and 4D Generation for a qualitative evaluation in Sec. 6.4 and Tab.3 to conclude our findings.




(a) Ablation on Local Chamfer Distance Loss (b) Pose Transfer without Temporal Consistency (c) Failure Cases of Zero-123 (i) Input RGB Image (ii) Output from Zero123

Figure 7: **Ablation Experiments.**

## 4.2 Diagnostic

**Effectiveness of Global-Local Chamfer Loss.** Relying solely on the geometric information contained in a short monocular video often makes it challenging to achieve satisfactory 3D reconstruction results. This difficulty arises because most videos typically offer sparse viewpoints and the presence of non-rigid object deformations increases the optimization space. Consequently, many methods that

Table 2: User study of MagicPose4D on Motion Transfer. We collect the rating results from 50 participants of eleven mesh-sequence comparison experiments. The scale of rating is 0 (low) - 5 (high). The participants found that MagicPose4D generates the best motion transfer results. Criteria for judgment: **1)** The generated motion should match the reference pose mesh sequence. **2)** The identity of the transferred mesh should match the identity reference. **3)** The generated mesh sequence should be consistent and smooth. More details can be found in Appendix Sec.6.4

| Method | **Exp-1** | **Exp-2** | **Exp-3** | **Exp-4** | **Exp-5** | **Exp-6** | **Exp-7** | **Exp-8** | **Exp-9** | **Exp-10** | **Exp-11** | **Ave.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3D-CoreNet [25] | 1.85 | 1.73 | 2.10 | 2.31 | 1.90 | 1.73 | 2.29 | 2.73 | 2.65 | 1.88 | 1.79 | 2.09 |
| X-DualNet [26] | 2.88 | 1.94 | 1.53 | 1.88 | 2.22 | 2.18 | 2.14 | 2.86 | 2.71 | 2.22 | 2.31 | 2.26 |
| MagicPose4D | **4.92** | **4.47** | **4.69** | **4.39** | **4.45** | **4.45** | **3.90** | **4.12** | **4.02** | **4.10** | **4.20** | **4.34** |

Table 3: User study of MagicPose4D on 4D Generation compared to Animate124 [40]. The settings are the same as the study on Motion Transfer. Criteria for judgment: **1)** The generation's identity should match the reference image. **2)** The video should be temporal-consistent and smooth.

| Method | **Exp-1** | **Exp-2** | **Exp-3** | **Exp-4** | **Exp-5** | **Exp-6** | **Ave.** |
|---|---|---|---|---|---|---|---|
| Animate124 [40] | 2.60 | 1.98 | 1.90 | 1.92 | 2.28 | 2.24 | 2.15 |
| MagicPose4D | **4.53** | **4.37** | **3.91** | **4.00** | **4.60** | **4.62** | **4.34** |

rely only on 2D supervision [31, 34, 30] struggle to ensure effectiveness on in-the-wild videos. To enhance the generalization capability of the reconstruction module, we utilize existing image-to-3D model methods to obtain geometric priors. By using these models, we generate pseudo-3D meshes for each frame as 3D supervision. We employ the Chamfer distance as the 3D loss objective to align the predicted 3D mesh closely with the pseudo-3D mesh in terms of point distribution. **However**, global Chamfer distance alone can only ensure similar point distributions between the two meshes and does not guarantee the correspondence of points as expected. For instance, as shown in Fig.7(a), without using the local Chamfer loss, the overall point distribution might be similar, but the mesh shape could be entirely incorrect, such as the left and right legs being swapped or vertices originally on the right front leg being moved near the left hind leg. By incorporating the local Chamfer loss, we enforce consistency between each part of the predicted mesh and the pseudo mesh, thus ensuring pose consistency and significantly mitigating the previously mentioned issues.

**Temporal Consistency** is crucial for 4D generation. As shown in Fig.7 (b), performing pose transfer independently for each frame results in noticeable discontinuities when viewing the entire video. To address this, we first define the canonical space using the initial frame, where all frames share the same canonical mesh, skeleton, and skinning weights. Each subsequent frame is derived from the deformation of the first frame using global-local transformations and blending skinning. This approach ensures the consistency of the static model and rigging. Additionally, we introduce a dynamic rigidity regularization term between consecutive frames, which minimizes the deformation of the object between successive frames. This ensures temporal smoothness of the 4D content.

# 5 Conclusion & Limitations

We introduce MagicPose4D, a novel framework for 4D generation providing more accurate and customizable 4D motion transfer. We propose a dual-phase reconstruction process that initially uses accurate 2D and pseudo 3D supervision without skeleton constraints and subsequently refines the model with skeleton constraints to ensure physical plausibility. We incorporate a novel loss function that aligns the overall distribution of mesh vertices with the supervision and maintains part-level alignment without additional annotations. MagicPose4D enables cross-category motion transfer using a kinematic-chain-based skeleton, ensuring smooth transitions between frames through dynamic rigidity and achieving robust generalization without the need for additional training.

The main **limitations** of our method and existing works are: **(i)** Deformation based on blend skinning relies on accurate and robust skeletons and skinning weights predictions, facing a trade-off between generalization and accuracy. Learning-based methods have limited generalization due to restricted training datasets, while non-learning methods suffer from inductive bias, leading to suboptimal results. **(ii)** MagicPose4D can infer poses quickly for pose transfer without training, but 4D reconstruction requires significant training (10 hours on an L40S). **(iii)** Our method struggles with detailed motion control, such as fingers and facial features, due to the challenge of capturing fine-grain details during 4D reconstruction. These issues represent future research directions.

# References


[1] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG)*, 22(3):587–594, 2003.

[2] Andreas Bærentzen and Eva Rotenberg. Skeletonization via local separators. *ACM Transactions on Graphics (TOG)*, 40(5):1–18, 2021.

[3] Sherwin Bahmani, Xian Liu, Yifan Wang, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. *arXiv preprint arXiv:2403.17920*, 2024.

[4] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023.

[5] Ilya Baran and Jovan Popović. Automatic rigging and animation of 3d characters. *ACM Transactions on graphics (TOG)*, 26(3):72–es, 2007.

[6] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video, 2018.

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. CVPR*, 2023.

[9] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

[10] Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365*, 2024.

[11] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21003–21012, 2023.

[12] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[14] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[15] Zhouyingcheng Liao, Jimei Yang, Jun Saito, Gerard Pons-Moll, and Yang Zhou. Skeleton-free pose transfer for stylized 3d characters. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022.

[16] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023.

[17] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proc. ICCV*, 2023.

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[20] Pablo Palafox, Aljavz Bovzivc, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 12675–12685, 2021.

[21] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.

[22] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. DreamGaussian4D: Generative 4D Gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

[23] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[24] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. In *Proc. ICML*, 2023.

[25] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. 3d pose transfer with correspondence learning and mesh refinement. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[26] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. Unsupervised 3d pose transfer with cross consistency and dual reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13, 2023.

[27] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[28] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[29] Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan. Scanning 3d full human bodies using kinects. *IEEE transactions on visualization and computer graphics*, 18(4):643–650, 2012.

[30] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023.

[31] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15980–15989, 2021.

[32] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15980–15989, 2021.

[33] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. *Advances in Neural Information Processing Systems*, 34:19326–19338, 2021.

[34] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022.

[35] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022.

[36] Qitong Yang, Mingtao Feng, Zijie Wu, Shijie Sun, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Beyond skeletons: Integrative latent mapping for coherent 4d sequence generation. *arXiv preprint arXiv:2403.13238*, 2024.

[37] Hao Zhang, Fang Li, Lu Qi, Ming-Hsuan Yang, and Narendra Ahuja. Csl: Class-agnostic structure-constrained learning for segmentation including the unseen. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7078–7086, 2024.

[38] Hao Zhang, Fang Li, Samyak Rawlekar, and Narendra Ahuja. Learning implicit representation for reconstructing articulated objects. *arXiv preprint arXiv:2401.08809*, 2024.

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

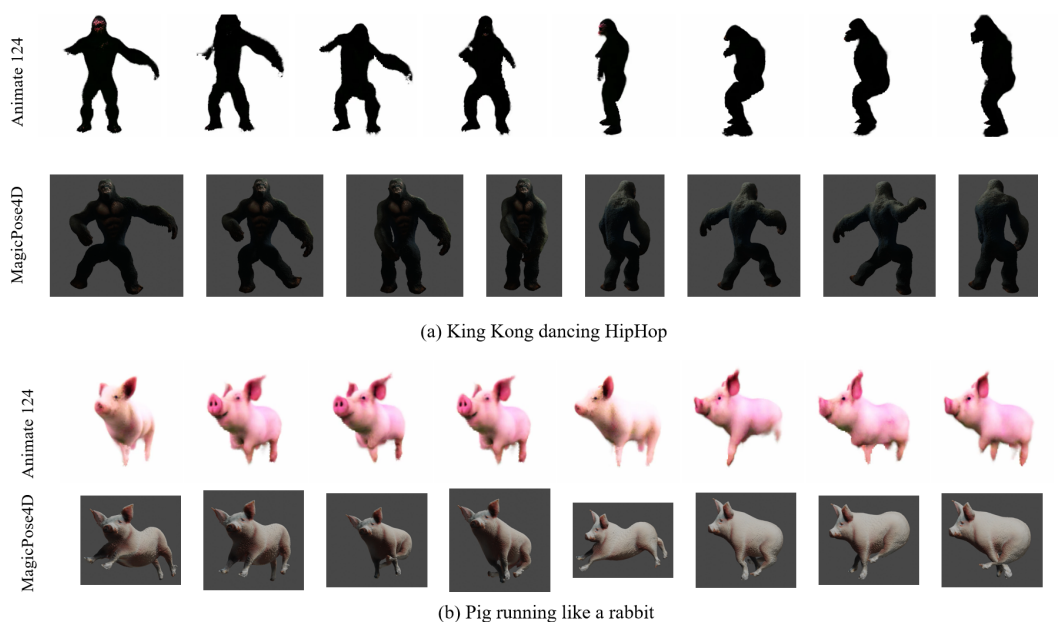
[40] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

[41] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*, 2023.

[42] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017.

Table 4: Support Prompts Comparison. Random represents no control of motion.

| Method | Appearance | | Motion | | | |
|---|---|---|---|---|---|---|
| | Text | Image | Text | Video | Trajectory | Random |
| Animate123 [40] | ✓ | ✓ | ✓ | | | |
| 4D-fy [4] | ✓ | | ✓ | | | |
| DreamGaussian4D [22] | ✓ | ✓ | | | | ✓ |
| AYG [16] | ✓ | | ✓ | | | |
| Dream-in-4D [41] | ✓ | ✓ | ✓ | | | |
| TC4D [3] | ✓ | ✓ | ✓ | | ✓ | |
| MagicPose4D | ✓ | ✓ | ✓ | ✓ | ✓ | |




(a) King Kong dancing HipHop

(b) Pig running like a rabbit

Figure 8: **4D Generation Comparison with Animate124**

# 6 Appendix

In this section, we aim to provide additional information not included in the main text due to length constraints. This supplemental content includes:

1. Support prompts comparison with existing 4D generation methods in Tab.4.
2. Video results (Sec.6.1)
3. Implementation details (Sec.6.3)
4. Information about the dataset used for evaluation (Sec.6.2)
5. In-depth explanation of loss and regularization terms (Sec.6.5)
6. Further experimental results: (i) comparison with existing 4D generation method in Fig.8, (ii) comparison with existing pose transfer methods in Fig.9, (iii) skeleton and skinning weights results in Fig.10
7. User study (Sec.6.4)
8. Broader social impacts (Sec.7)

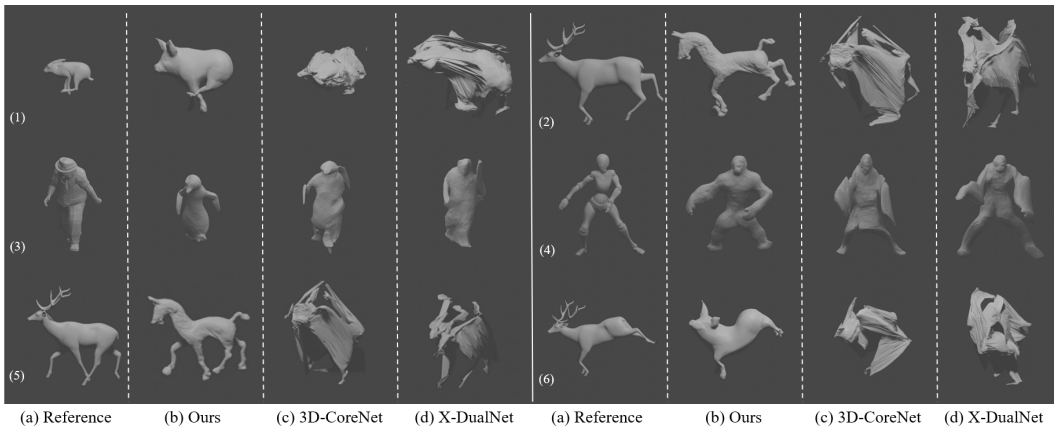



Figure 9: **Motion Transfer Comparison with 3D-CoreNet [25] and X-DualNet [26]**

### 6.1 Video Results

As shown in Tab.5, we include an extensive set of 4D generation results in video format in Google Drive: https://drive.google.com/drive/folders/123zX75-mRM4Yi49Nw8AlPpoiWXlC18pE?usp=sharing. There, we demonstrate 4D generation results across different species with diverse motions and compare the performance of pose transfer with 3D-CoreNet [25], X-DualNet [26], and Animate124 [40]. Since the video files exceed 500 MB, we are unable to upload them directly as a zip file. Therefore, we have used an anonymous Google Drive link.

### 6.2 Dataset

#### 6.2.1 Animal

**Davis-Camel** provides a real animal video in BADJA [6] with 2D keypoints and mask annotations, derived from the DAVIS video segmentation dataset [21] and online stock footage. We extract reference motion from the reconstructed mesh sequence and transfer it to other identities.

**PlanetZoo** includes RGB synthetic videos of different animals with around 100 frames each. PlanetZoo covers a 180-degree visual field captured by a moving camera to allow better evaluation of 3D reconstruction when imaging parameters must also be dynamically estimated due to the moving camera, and over a large visual field. In addition, following BADJA [6], we also provide 2D key point annotations.

**DeformingThings4D** is a synthetic dataset containing 1,972 animation sequences containing 31 categories of both humanoids and animals. Each sequence consists of 40 to 120 frames of motion animation. In this dataset, the first frame is the canonical frame, and its triangle mesh is given. From the 2nd to the last frame, the 3D offsets of the mesh vertices are provided, and we export the triangle meshes for all these frames. We use the motions of these animal mesh sequences as pose references and transfer the pose to different identities.

#### 6.2.2 Human

**EverybodyDanceNow** consists of **full-body** videos of five human subjects. We use these monocular videos to generate human motions and transfer them to other identities.

**DeformingThings4D** also contains humanoid examples of dynamic mesh, as mentioned before. We use the motions of these sequences as pose references and transfer the pose to different identities.

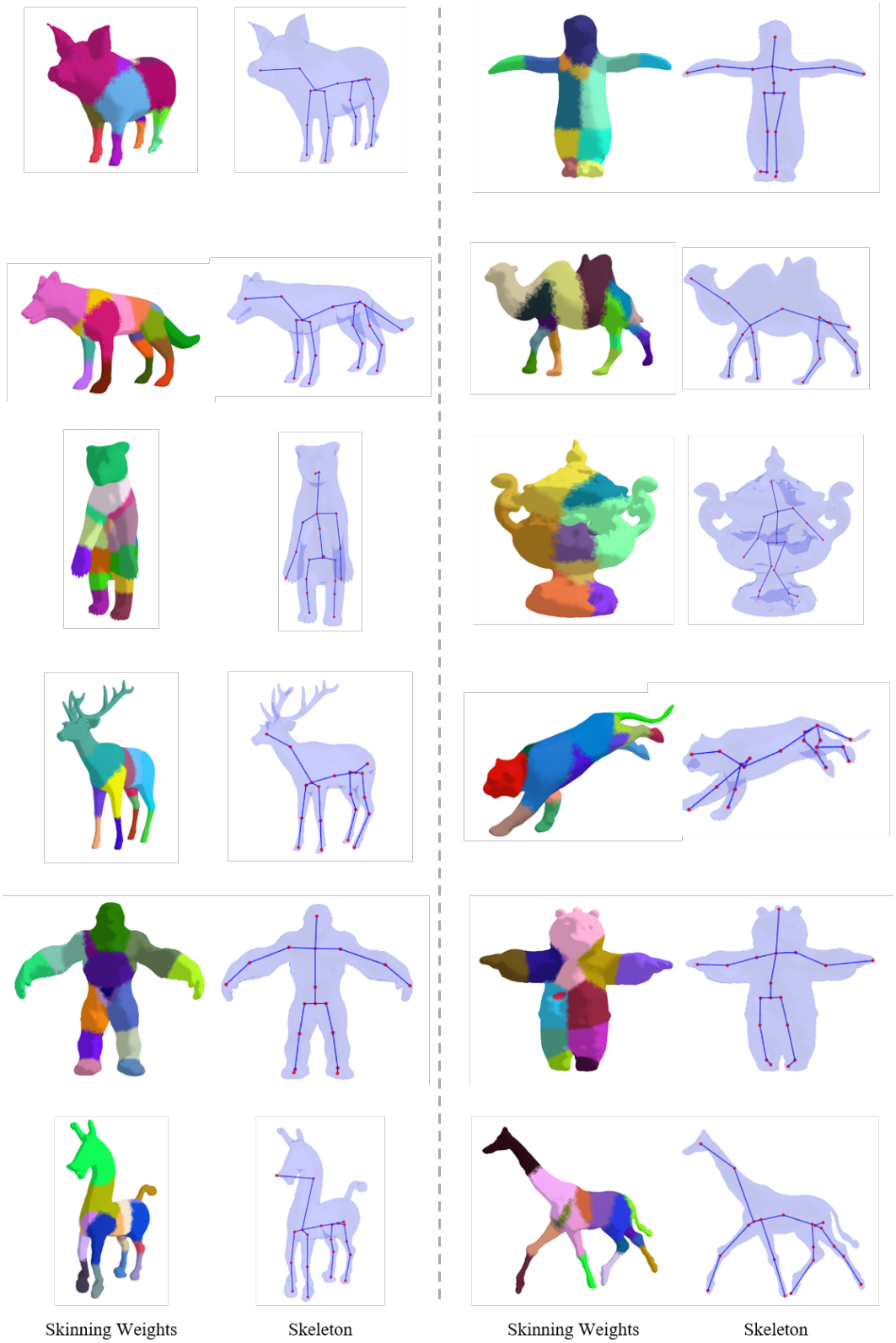



Figure 10: **Skinning Weights** and **Skeleton** Results from Our Method.

Table 5: **List of Video Results.**

| | reference motion | target object | our result | 3D-CoreNet | X-DualNet | Animate124 |
|---|---|---|---|---|---|---|
| 1 | bear_death | horse | ✓ | ✓ | ✓ | |
| 2 | bear_death | pig | ✓ | ✓ | ✓ | |
| 3 | canie_jump | horse | ✓ | ✓ | ✓ | |
| 4 | canie_jump | pig | ✓ | ✓ | ✓ | |
| 5 | cattle_walkback | horse | ✓ | ✓ | ✓ | |
| 6 | cattle_walkback | pig | ✓ | ✓ | ✓ | |
| 7 | deer_attack1 | horse | ✓ | ✓ | ✓ | |
| 8 | deer_attack1 | pig | ✓ | ✓ | ✓ | ✓ |
| 9 | deer_attack2 | horse | ✓ | ✓ | ✓ | |
| 10 | deer_attack2 | pig | ✓ | ✓ | ✓ | |
| 11 | deer_attack3 | horse | ✓ | ✓ | ✓ | |
| 12 | deer_attack3 | pig | ✓ | ✓ | ✓ | |
| 13 | deer_jumptort | horse | ✓ | ✓ | ✓ | |
| 14 | deer_jumptort | pig | ✓ | ✓ | ✓ | |
| 15 | drunkwalk | panda | ✓ | ✓ | ✓ | |
| 16 | drunkwalk | kingkong | ✓ | ✓ | ✓ | |
| 17 | drunkwalk | penguin | ✓ | ✓ | ✓ | |
| 18 | drunkwalk | smpl | ✓ | ✓ | ✓ | |
| 19 | hiphop | panda | ✓ | ✓ | ✓ | |
| 20 | hiphop | kingkong | ✓ | ✓ | ✓ | ✓ |
| 21 | hiphop | penguin | ✓ | ✓ | ✓ | |
| 22 | rabbit_run | horse | ✓ | ✓ | ✓ | |
| 23 | rabbit_run | pig | ✓ | ✓ | ✓ | ✓ |
| 24 | bear_attack | horse | ✓ | | | |
| 25 | bear_attack | pig | ✓ | | | |
| 26 | bear_drink | horse | ✓ | | | |
| 27 | bear_drink | pig | ✓ | | | |
| 28 | bear_run | horse | ✓ | | | |
| 29 | bear_run | pig | ✓ | | | |
| 30 | cattle_attack2 | horse | ✓ | | | |
| 31 | cattle_attack2 | pig | ✓ | | | |
| 32 | deer_jump | horse | ✓ | | | |
| 33 | deer_jump | pig | ✓ | | | |
| 34 | deerFEL_tort | horse | ✓ | | | |
| 35 | deerFEL_tort | pig | ✓ | | | |
| 36 | dog_jumpup | horse | ✓ | | | |
| 37 | dog_jumpup | pig | ✓ | | | |
| 38 | tiger_run | horse | ✓ | | | |
| 39 | tiger_run | pig | ✓ | | | |
| 40 | hands_up | panda | ✓ | | | |
| 41 | hands_up | kingkong | ✓ | | | |
| 42 | hands_up | penguin | ✓ | | | |
| 43 | hiphop | kingkong_superman | ✓ | | | |
| 44 | hiphop | Chinese pot | ✓ | | | |
| 45 | drunkwalk | Chinese pot | ✓ | | | |

Figure 11: **Post Texture Editing.**

#### 6.2.3 Self-collected in-the-wild data

These in-the-wild data are collected from online resources. We use unique objects such as Chinese pot, and others as appearance identities to demonstrate the generalization ability of our method.

### 6.3 Implementation Details

For Davis-Camel, PlanetZoo, EverybodyDanceNow and those self-collected data without ground truth mesh, we first train the Dual-Phase 4D Reconstruction Module on 2 NVIDIA L40S GPUs with batch size 16 for 10 epochs with a learning rate of $0.0001$. We then transfer the per-frame motion from Phase 2 of 4D Reconstruction to the target object with the Cross-Category Motion Transfer Module. The motion transfer is not learning-based and does not require any trainable parameters, which makes MagicPose4D generalize well to unseen identities and reference motions. For all other data with ground truth mesh available, we directly train the second phase of the 4D Reconstruction Module and then transfer the motions.

### 6.4 User Study

We provide a user study for comparison between MagicPose4D and previous works [25, 26] on motion transfer. We asked 50 lay participants from **Prolific**, an online platform for user studies, to rate the quality of **eleven** in-the-wild retargeted mesh sequences from 3D-CoreNet [25], X-DualNet [26] and MagicPose4D on a scale of 0(low) to 5(high). The participants are paid with an hourly rate of 16 USD. In each mesh sequence comparison, we collect target identity mesh, reference motion mesh sequence, and retargeting results. We visualize them side-by-side. The retargeting results from different methods are anonymized as A, B, C, and the order is randomized. We provide video visualization of these comparisons (**For each video, from left to right: Reference Motion; MagicPose4D; 3D-CoreNet; X-DualNet**) in the Google Drive. Criteria for judgment: **1)** The generated motion should match the reference pose mesh sequence. **2)** The identity of the transferred mesh should match the identity reference. **3)** The generated mesh sequence should be consistent and smooth. From the results presented in Tab. 2, we conclude that MagicPose4D provides the most satisfying generation of mesh sequences.

Similarly, we provide a user study for comparison with Animate124 [40], a representative work using the diffusion model and SDS-loss to optimize a neural representation. The settings for participants are the same. In each comparison experiment, we collect reference images, text descriptions of motion, mesh reference motion sequences, and generated videos from both methods. We feed the reference image, which provides the identity, and text prompt, which describes the motion of the generated object, into Animate124 and optimize the NeRF with SDS-loss. For MagicPose4D, we use the reference image to generate the target mesh with an Image-to-3D model and transfer the motion of the reference mesh sequence. We compared the rendered videos from Animate124 to videos of generated mesh from MagicPose4D in different viewpoints. We provide video visualization of these comparisons (**Method 1: Animate124; Method 2: MagicPose4D**) in the Google Drive. Criteria for judgment: **1)** The generation's identity should match the reference image. **2)** The generated mesh sequence should be consistent and smooth. From the results presented in Tab. 3, we conclude that MagicPose4D provides the most satisfying visualizations.

### 6.5 2D Losses and Regularization

The 3D loss is described in the main article. Here we introduce the 2D losses and regularization terms. The 2D losses are similar to those in existing differentiable rendering pipelines [31, 34, 38]. We define $\mathbf{S^t}$,$\mathbf{I^t}$, $\mathbf{F}^{2\mathrm{D},t}$ as the silhouette, input image, and optical flow of the input image, and their corresponding rendered counterparts as $\tilde{\mathbf{S}}^{\mathbf{t}}$,$\tilde{\mathbf{I}}^{\mathbf{t}}$, $\tilde{\mathbf{F}}^{2\mathrm{D},t}$. The following losses ensure the fitting between rendered and original:

$$\mathcal{L}_{\text{silhouette}} = \sum_{\mathbf{x}^t} \left\| \mathbf{s}(\mathbf{x}^t) - \hat{\mathbf{s}}(\mathbf{x}^t) \right\|_2, \tag{4}$$

$$\mathcal{L}_{\text{optical flow}} = \sigma \left\| (\tilde{\mathbf{F}}^{\mathbf{2D,t}}) - (\mathbf{F}^{\mathrm{2D,t}}) \right\|_2^2, \tag{5}$$

$$\mathcal{L}_{\text{texture}} = \left\| \tilde{\mathbf{I}}^{\mathbf{t}} - \mathbf{I}^{\mathbf{t}} \right\|_1, \tag{6}$$

$$\mathcal{L}_{\text{perceptual}} = \text{pdist}(\tilde{\mathbf{I}}^{\mathbf{t}}, \mathbf{I}^t), \tag{7}$$

where pdist(,) is the perceptual distance [39]. Also, we leverage three **regularization terms:** (1) Dynamic Rigidity term, which is introduced in paper.[38]. (2) Symmetry term, which encourages the canonical mesh to be symmetric.

$$\mathcal{L}_{\text{symm}} = \sum_i \min_j \left\| v_j - \phi\left(v_i{}^{symm}\right) \right\|^2, \tag{8}$$

where, $v_j$ is the vertex j in the one side of the symmetry plane, and $v_i{}^{symm}$ is the vertex from the other side. $\phi$ is the reflection operation w.r.t to the symmetry plane. (3) Laplacian smoothing: We apply laplacian smoothing to generate smooth mesh surfaces.

$$\mathcal{L}_{\text{shape}} = \left\| \mathbf{X}_i^0 - \frac{1}{|N_i|} \sum_{j \in N_i} \mathbf{X}_j^0 \right\|^2, \tag{9}$$

where, $\mathbf{X}_i^0$ is coordinates of vertex $i$ in canonical space. And $N$ is the number of vertices

## 7 Broader Social Impacts

The proposed MagicPose4D for motion transfer offers extensive applications, enhancing communication in digital environments by enabling more effective self-expression through avatars or digital characters. Additionally, MagicPose4D has the potential to revolutionize the entertainment and media production industries by facilitating the creation of more lifelike and expressive characters in movies, video games, and animations.

However, this technology also presents potential negative social impacts. Privacy concerns arise from the unauthorized creation of realistic animations of individuals, and the technology could facilitate the spread of misinformation through deepfakes. Risks include job displacement in fields like acting and modeling, psychological effects from the blurring of reality and virtual experiences, and the exploitation of the technology for unethical purposes. Furthermore, cultural insensitivity, security threats, and the misuse of realistic animal animations could have broader societal implications. Addressing these issues requires robust ethical guidelines, legal frameworks, and technological safeguards to ensure responsible use and mitigate harm.

# NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the `Abstract` section, we briefly illustrate our contributions at a high level. At the end of the `Introduction` section, we claim our contributions in detail.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: In the `Contribution & Limitation` section of the main paper, we discuss the limitations of our proposed approach. We take these constraints as promising directions for subsequent research endeavors.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not have any theory results in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In the `Experiments` section in the main paper and the `Implementation Details` section in the appendix, we illustrate all the details required to reproduce the main experimental results of our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We did not release the code of our method till now, because the codes are not well packaged. However, we plan to release our code in the future, e.g. upon acceptance, to let other researchers implement our methods into their research.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the `Datasets` section and the `Implementation Details` section of the appendix, we show all the experimental settings and details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the `Experiments` section of the main paper and the `Evaluation Metrics` section of the appendix, we discuss the error bars and the meanings of each evaluation metric in great detail.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the `Implementation Details` section of the appendix, we provide information on the computation resource we used in all experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We read the NeurIPS Code of Ethics, and believe our research is conducted in the paper conform.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss all the potential impacts in the `Contribution & Limitation` section at the end of the main paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper presents no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original owners of all assets while mentioning them in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We document all references in the `Reference` section.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.