JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 4, pp. 55-66, April 2026
https://doi.org/10.9708/jksci.2026.31.04.055

Text-Controlled 4D Human Generation

Chanwoo Kim*, Sanghun Kim**, Hwasup Lim***

*Student, Dept. of AI-Robotics, KIST School, University of Science and Technology (UST), Daejeon, Korea
**Researcher, Broz, Seoul, Korea
***Professor, Dept. of AI-Robotics, KIST School, University of Science and Technology (UST), Daejeon, Korea

[Abstract]

Generating 4D humans from textual descriptions has become an important problem in various applications such as the metaverse and virtual reality. However, previous Text-to-4D generation methods typically generate appearance and motion jointly, which leads to limited controllability and high computational cost. In this paper, we propose a novel text-driven 4D human generation pipeline that integrates separately generated appearance and motion. First, from the given appearance and motion descriptions, a human appearance image and a motion sequence are generated using Stable Diffusion and Motion Diffusion Model, respectively. Next, MusePose combines the generated appearance and motion into a frontal-view video, which is then extended into multi-view videos using SV4D. Finally, Grid4D is employed to learn 4D representation from the synthesized multi-view videos. To validate the proposed pipeline, we construct a dataset for 4D human generation and conduct quantitative and qualitative evaluations on rendered videos. Experimental results show that the proposed method achieves 77.5% in Dynamic Degree, 58.3% in Aesthetic Quality, and 24.8% in Overall Consistency, indicating that while trade-offs exist among some metrics, the method maintains a balance between dynamic expressiveness and visual quality.

▸ **Key words**: Text-to-4D Generation, Dynamic Human Synthesis, Gaussian Splatting, Generative Model, Deep Learning

[요 약]

문장으로부터 4차원 인간을 생성하는 문제는 메타버스 및 가상현실과 같은 다양한 응용 분야에서 중요한 주제로 인식된다. 그러나 기존의 Text-to-4D 생성 방법들은 외형과 동작을 동시에 생성하는 방식을 주로 사용하고 이는 제어 가능성 저하와 높은 비용의 한계를 가진다. 본 논문에서는 외형과 동작을 분리하여 생성한 뒤 이를 통합하는 텍스트 기반의 4차원 인간 생성 파이프라인을 제안한다. 먼저 주어진 외형, 동작 문장으로부터 Stable Diffusion으로 인간의 외형 이미지를 생성하고, Motion Diffusion Model로 동작을 생성한다. 이후 MusePose로 외형과 동작을 결합한 정면 시점 비디오를 생성하고, SV4D를 통해 다시점 비디오로 확장한 뒤, Grid4D를 통해 4차원 표현을 학습한다. 제안한 파이프라인의 검증을 위해 4차원 인간 생성용 데이터 세트를 구축하고, 렌더링 비디오에 정량적, 정성적 평가를 수행하였다. 실험 결과, 제안한 방법은 77.5% Dynamic Degree, 58.3%의 Aesthetic Quality, 그리고 24.8%의 Overall Consistency를 나타내 일부 지표에서 trade-off가 있으나 전반적으로 동작성 개선과 품질 간 균형을 보인다.

▸ **주제어**: 문장 기반 4D 생성, 동적 사람 합성, 가우시안 스플래팅, 생성 모델, 딥러닝

• First Author: Chanwoo Kim, Corresponding Author: Hwasup Lim
*Chanwoo Kim (sky9739a@gmail.com), Dept. of AI-Robotics, KIST School, University of Science and Technology (UST)
**Sanghun Kim (powerkei@naver.com), Broz
***Hwasup Lim (hslim@kist.re.kr), Dept. of AI-Robotics, KIST School, University of Science and Technology (UST)
• Received: 2026. 02. 20, Revised: 2026. 03. 29, Accepted: 2026. 04. 01.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 텍스트 기반 생성 모델의 발전을 통해, 언어를 입력으로 사용해 사용자의 의도를 반영한 다양한 형태의 시각적 콘텐츠를 생성하는 기술이 빠르게 발전되고 있다. 특히, Diffusion Model[1]의 등장 이후 텍스트에서 이미지를 생성하는 Text-to-Image 모델은 높은 표현력과 사실적인 이미지를 생성하며 연구 성장을 크게 이끌었다. 이러한 기술의 확장은 정적인 2D 이미지 생성에 그치지 않고, 3D와 4D 영역으로까지 빠르게 확장되며 새로운 생성 패러다임을 형성하고 있다.

Text-to-4D 생성 연구들은 주로 Text-to-Image 또는 Text-to-Video 모델의 prior를 활용하는 방식으로 발전해 왔다. 대표적으로 Score Distillation Sampling (SDS) [2] 기반 방법들은 텍스트 조건을 반영해 3D/4D 표현을 직접 최적화하는 방식으로 문제를 다루었으며, 4D-fy[3]와 Dream-in-4D[4]는 대표적인 Image diffusion model과 Video diffusion model의 prior를 SDS 방식을 통해 학습한 방법이다. 이러한 방식들은 NeRF[5] 기반의 표현 방법을 사용했는데, 최근에는 빠른 렌더링과 높은 표현력의 장점을 가진 3D Gaussian Splatting[6](3DGS)의 등장으로 인해 장면 표현 방법이 NeRF 계열에서 Gaussian splatting으로 확장되고 있다. 예를 들면, 4Real[7], Free4D[8]는 시간적 변형을 학습하기 위해 videodiffusion model의 prior와 Gaussian 기반의 표현을 결합하여 시공간적 일관성과 렌더링 효율을 높이는 방향으로 발전하였다.

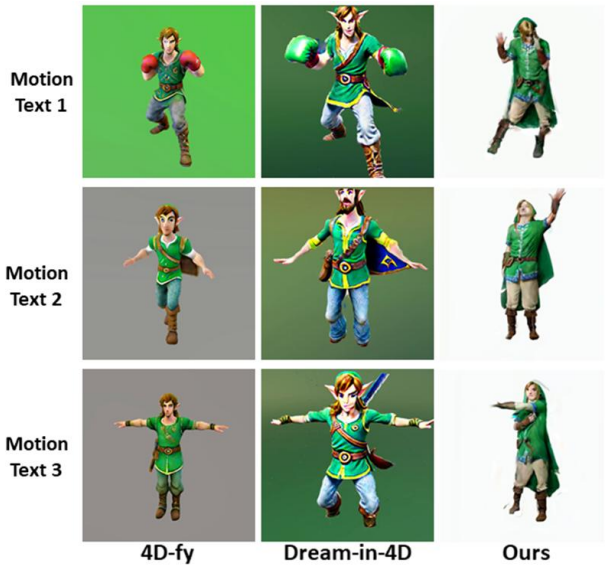

Fig. 1. Given the same appearance text and different motion texts, previous methods fail to preserve consistent appearance across motions, but our method maintains stable appearance while generating diverse motions.

그러나 이러한 기존 방법들은 외형과 동작을 결합한 문장을 기반으로 하나의 생성 과정에서 동시에 처리하거나, 정적 표현과 동적 변형을 단계적으로 학습하더라도 외형과 동작에 대한 입력 조건을 독립적으로 생성 및 제어하는 구조와는 다르다. 특히 4D 인간 생성 분야에서는 사람의 외형과 동작을 각각 완전히 독립적으로 생성하고 제어할 수 있는 능력이 중요하다. 이는 동일한 동작에도 다양한 외형을 결합할 수 있고, 동일한 외형에도 서로 다른 동작이 부여될 수 있기 때문이다. 그러나 기존 방식들은 외형과 동작을 하나의 생성 과정에서 함께 처리하기 때문에, 두 요소를 명확히 제어하기 어렵고, 다양한 외형-동작 조합 생성에 제약이 있다. 또한, 이전의 방식은 약간의 동작 편집이 필요한 경우에도 전체 과정을 처음부터 다시 수행해야 하므로, 재생성에 따른 비용이 요구되며 기존에 생성된 영상과의 일관성을 유지하기 어렵다. Fig. 1에서는 이전 방법들[3, 4]과 제안한 방법에 대해 동일한 외형 문장 조건에서 서로 다른 동작 문장을 입력으로 사용했을 때의 생성 결과를 시각화하였다. 그리고 기존에 제시된 방법들은 SDS 기반의 최적화 방식을 사용하여 4D 표현을 생성한다. 이러한 방식은 diffusion 모델의 학습된 prior를 활용해 정적인 구조를 생성하는 데에는 효과적이지만, 시간축에 따른 큰 동작 변화를 생성하는 데에는 한계가 있다. 따라서 외형과 동작을 분리하여 개별적으로 생성하고 이를 유연하게 조합할 수 있는 접근 방식은 다양한 4D 인간 표현을 생성하기 위한 중요한 연구 방향이라 할 수 있다.

본 연구에서는 이러한 한계를 극복하기 위해 사람의 외형과 동작을 텍스트로부터 독립적으로 생성하고, 이를 통합하여 4D 사람 모델로 확장하는 새로운 Text-to-4D Human 생성 파이프라인을 제안한다. 먼저 diffusion model 기반의 텍스트 기반의 이미지 생성 모델을 사용해 텍스트 조건에 맞는 사람 외형 이미지를 생성한다. 그리고 별도의 텍스트 기반의 동작 생성 모델을 통해 사람 동작 시퀀스를 생성한다. 이후 사람 이미지와 사람 동작을 결합하는 모델을 통해 정면 시점의 비디오를 생성하고, 단안 시점 비디오를 다시점으로 확장하기 위한 모델을 파이프라인에 통합한다. 마지막으로, 여기서 얻은 다시점 비디오를 3D 표현 방법인 Gaussian splatting 기반의 변형 네트워크를 학습하는 데 활용한다. 이를 통해 시공간적으로 일관된 4D 표현을 구성함으로써 다양한 시점에서 동작하는 사람을 렌더링할 수 있게 된다. 이러한 모듈화된 접근 방식은 사용자가 원하는 외형과 원하는 동작을 자유롭게 조합할 수 있게 하며, 높은 제어성과 다양성을 제공할 수 있게 한다.

본 연구의 핵심 기여는 다음과 같다.

1. 외형과 동작을 분리하여 생성한 뒤 통합하는 구조를 통해, 문장 기반 4D 인간 생성에서 외형-동작 조합의 유연성과 제어 가능성을 높이는 실용적 접근 방법을 제안하였다.

2. 기존의 외형 생성, 동작 생성, 외형-동작 결합, 다시점 영상 생성, 4D 복원 단계를 하나의 흐름으로 연결하는 4D 인간 생성을 위한 단계적 파이프라인을 제안하였다.

3. 외형과 동작 조합이 가능한 평가 데이터를 직접 구성하고, 이를 바탕으로 정량적, 정성적 평가를 수행하여 제안한 파이프라인이 동적 표현 능력을 크게 향상하면서 시각적 품질과의 균형을 유지할 수 있음을 보인다.

또한, 본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 선행 연구를 소개하고, 3장에서는 본 연구에서 제안하는 파이프라인과 각 모듈을 구체적으로 설명한다. 4장에서는 본 연구를 위해 구축한 데이터와 실험 결과를 서술하며, 5장에서는 결론과 한계점 및 향후 연구 방향을 제시한다.

II. Related Work

1. Human Motion Generation

인간 동작 생성 기술은 텍스트 또는 조건 입력으로부터 자연스러운 사람의 움직임을 생성하는 기술로, 최근 딥러닝 기반 학습 방식을 통해 크게 발전되었다. 이러한 기술은 사람을 스켈레톤 형태의 관절 정보로 표현하거나, 사람을 메쉬 형태로 표현한 SMPL[9] 모델을 사용하여 학습된다.

최근에는 CLIP model[10]을 사용하여 텍스트와 이미지 사이의 의미적 연관성을 학습하고, 이를 기반으로 텍스트 조건에 정교하게 대응하는 인간 동작 생성을 수행하는 연구 TMR[11]이 제안되었다. 또한, 생성 모델 VAE[12]와 Codebook을 활용하여 모션 시퀀스를 생성하는 연구인 T2M-GPT[13]도 제시되었다. 그러나, 이러한 방법들은 근본적으로 검색 기반 혹은 토큰 재조합 방식에 가까워, 복잡한 확률 분포를 표현하거나 다양한 형태의 동작을 생성하는 데에는 한계가 존재한다.

이와 달리, diffusion model의 강력한 생성 능력을 활용하여 더욱 자연스럽고 다양한 사람의 동작을 생성하는 논문들이 제시되고 있다. 대표적으로 Motion Diffusion Model(MDM)[14]은 diffusion model을 통해 복잡한 움직임을 안정적으로 생성하였다. 이러한 연구들은 대규모 모션 캡처 데이터를 기반으로 텍스트 조건에 따른 고품질 모션 시퀀스를 생성한다. 본 논문에서는 MDM을 사용하여 주어진 문장을 기반으로 자연스러운 사람의 동작을 메쉬 형태로 생성하였다.

2. Text-to-3D/4D Generation

Text-to-3D 연구는 텍스트 조건을 통해 정적 3D 모델을 생성하는 기술이다. 최근에는 diffusion model의 prior를 사용하는 방법인 Score Distillation Sampling (SDS)[2]를 통해서 텍스트-이미지 생성 모델의 score를 3D 표현에 직접 전달함으로써, NeRF[5]나 3D Gaussian Splatting(3D-GS)[6]을 최적화하는 방법들인 DreamFusion[2], DreamGaussian[15]이 제안되었다. 또한, Hunyuan3D[16]와 같이 대규모 멀티모달 모델을 활용해서 고해상도의 기하와 정교한 질감을 동시에 생성하는 방법들도 활발히 연구되고 있다.

이를 기반으로 Text-to-4D 연구 또한 활발히 진행되고 있는데, 초기 연구들인 4D-Fy[3], Dream-in-4D[4]는 정적 3D 생성을 위한 SDS 기반의 최적화 방식을 시간 축으로 확장하여 동적인 객체의 4D 표현을 생성하는 방식을 제시하였다. 최근에는 Gaussian splatting의 빠른 렌더링과 표현력을 기반으로 텍스트로부터 동영상을 생성해 정적 Gaussian의 offset을 학습하는 변형 네트워크를 통해 4D를 생성하는 Free4D[8]와 같은 연구가 제안되었다.

3. Dynamic Novel View Synthesis

Dynamic Novel View Synthesis는 시간 축으로 변화하는 동적 장면을 관찰된 영상으로부터 재구성하고, 이를 새로운 시점에서 렌더링하는 기술이다. 이는 3D를 표현하는 2가지의 방식인 NeRF와 Gaussian splatting을 사용하여 학습되었다. 암시적 모델인 NeRF 기반의 방식들은 D-NeRF[17], HyperNeRF[18]와 같이 시간 축을 포함한 장면의 변형을 암시적으로 모델링함으로써, 동적 장면을 효과적으로 표현하였다.

최근에는 빠른 렌더링을 위한 명시적인 표현 방식인 Gaussian splatting 기반의 방식이 도입되었으며, Dynamic-3DGS[18], 4D-GS[19], Grid4D[20] 등의 방법들이 제안되었다. 본 논문에서는 Grid4D 모델을 사용하여 동적 장면에 대해 표현하였다. Grid4D는 해시 인코딩 기반의 방식을 활용해 각 정적 Gaussian에 대한 시간적 offset을 추정함으로써, 연속적인 변형을 갖는 동적 장면을 표현할 수 있다. 따라서 본 논문은 Grid4D의 동적 장면 표현력을 활용하여 텍스트 기반 인간 4D 생성의 시간적 일관성을 보완하고 새로운 카메라 각도에서의 렌더링도 수행한다.

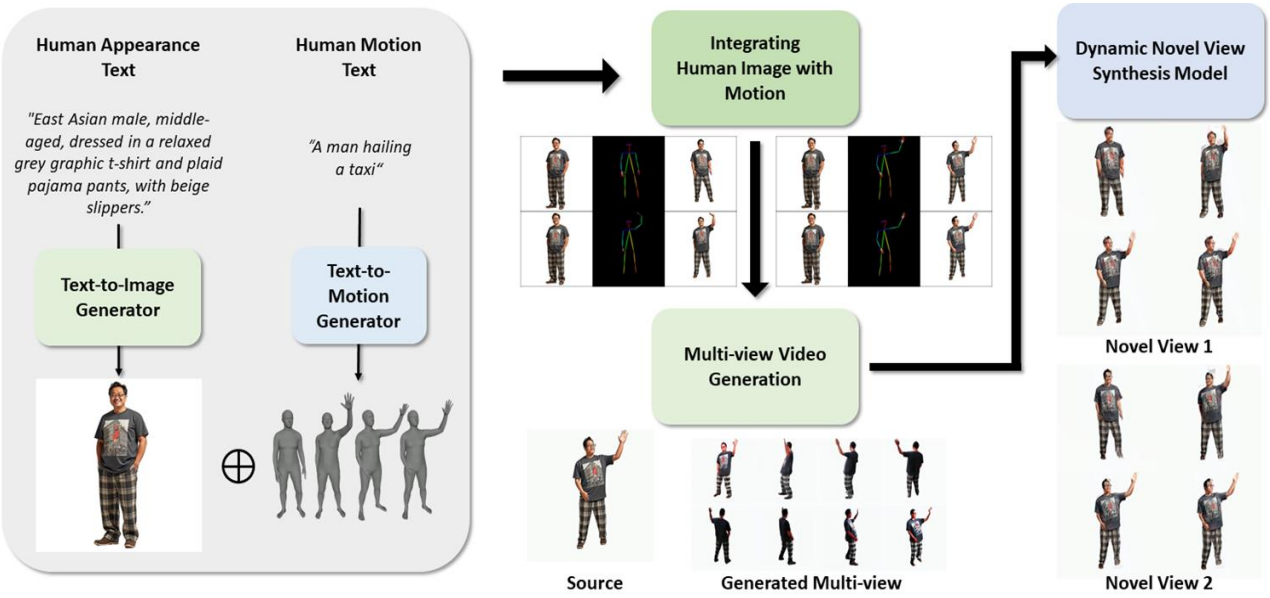

Fig. 2. Overview of our proposed pipeline.

III. Methods

본 연구는 텍스트 입력만으로 4D 인간(Human) 시퀀스를 생성하기 위해서 사람 외형 이미지 생성, 동작 생성, 사람 외형과 동작의 결합, 다시점 기반 영상 생성, 그리고 4D 재구성 단계를 하나의 파이프라인으로 통합한 프레임워크를 제안한다. 전체 과정은 Fig. 2에 요약되어 있으며, 각 단계에 대한 설명은 아래와 같다.

1. Preliminaries

본 연구에서 사용되는 3D 표현은 3D Gaussian Splatting[6]을 기반으로 한다. Gaussian splatting은 장면을 3D 공간 내의 연속적인 Gaussian primitives의 집합으로 나타내고, N개의 Gaussian으로 이루어진 Gaussian 집합은 위치 $\mu \in R^{N\times 3}$, 회전 $r \in R^{N\times 4}$, 크기 $s \in R^{N\times 3}$, 색상 $c \in R^{N\times 3}$, 불투명도 $o \in R^{N\times 1}$ 파라미터로 구성되어 실시간 렌더링이 가능한 표현 방식이다. 3D Gaussian은 다음과 같은 밀도 함수로 정의된다.

$$G(X) = \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)), \quad (1)$$

여기서, 공분산 $\Sigma = RSS^TR^T$ 은 회전 및 크기를 통해 구성된다. 그 후, 각 Gaussian은 카메라 투영 함수의 Jacobian J를 통해 $\Sigma' = J\Sigma J$ 형태의 2D 영상 평면에서의 공분산으로 매핑된다.

Gaussian splatting의 렌더링을 위해, 투영된 2D Gaussian은 깊이 순으로 정렬된 Gaussian의 가중 합으로 계산된다. 각 픽셀 색깔의 계산 방식은 아래와 같다.

$$C = \sum_{i\in N} c_i o_i \prod_{j=1}^{i-1}(1-o_j). \quad (2)$$

여기서 c_i와 o_i 는 i번째 Gaussian의 색상, 불투명도를 뜻하고, $\prod$ 항은 누적 투명도를 뜻한다. 결과적으로 이 식은 각 Gaussian이 해당 픽셀에 미치는 가중치를 통해 최종 RGB 값을 결정하게 되고, 가까운 깊이의 Gaussian이 더 큰 기여를 가지도록 보장한다. 이러한 합성 과정을 통해 Gaussian splatting 방식은 빠른 속도로 높은 품질의 이미지를 렌더링할 수 있다.

본 연구에서는 이러한 Gaussian splatting의 높은 표현력과 빠른 렌더링 속도를 기반으로, 시간에 따라 변화하는 3D 인간의 형태와 동작을 효율적으로 재구성할 수 있도록 하였다. 아래에서는 전체 과정을 단계별로 설명한다.

2. Human Image and Motion Generation

본 절에서는 주어진 텍스트 입력으로부터 인간의 정적인 외형과 변화하는 동작을 생성하는 과정을 설명한다. 본 연구에서는 이를 위해 텍스트 기반 이미지 생성기와 텍스트 기반 동작 생성기, 두 가지 생성 모듈을 독립적으로 활용한다.

인간 외형 생성을 위해 본 연구에서는 Stable Diffusion[22] 기반의 텍스트-이미지 생성 모델을 활용하였다. Stable Diffusion은 Latent Diffusion Model(LDM) 구조를 기반으로 하며, CLIP 텍스트 인코더에서 추출된 텍스트 임베딩을 조건으로 사용해 잠재 공간에서 이미지 denoising 과정을 수행한다. 텍스트 임베딩은 U-Net의

denoising 단계에 결합 되어 Diffusion Model이 텍스트의 의미를 반영하도록 유도하며 이를 통해 입력 문장에 부합하는 인간 외형 이미지를 효율적으로 생성할 수 있다. 이 모델은 얼굴, 의상 및 전반적인 스타일 등에 대한 문장을 받아 인간 이미지를 생성한다.

인간 동작 생성에는 Motion Diffusion Model(MDM) [14]을 사용하였다. MDM은 3D 스켈레톤 시퀀스를 Diffusion model을 통해 생성하며, 텍스트-동작 페어 데이터를 통해 학습되어 자연스럽고 일관된 동작을 생성할 수 있다. 생성 과정에서 텍스트 임베딩은 denoising 네트워크의 조건으로 사용되어 입력된 텍스트의 의미 (예: “사람이 달리고 있다.”, “사람이 가볍게 운동하고 있다.”)를 기반으로 시간 축에서 연속적인 동작을 생성한다. 최종적으로, MDM의 스켈레톤은 SMPL과 매핑되어 3D 메쉬 형태의 인간 동작을 얻을 수 있다. 실제로, 3D 메쉬 형태는 시각화를 위해서만 사용되고 다음 단계에서는 3D 스켈레톤을 사용한다.

3. Integrating human image with motion

생성된 인간 동작과 외형 이미지를 기반으로, 동적 영상을 생성하기 위해 본 논문에서는 MusePose[23]를 활용하였다. MusePose는 diffusion 기반 생성 구조를 사용하여 pose-driven human video를 생성하는 오픈 소스 프레임워크이다. 이 모델은 사람 외형 이미지에서 추출된 외형 특징과 OpenPose[24] 형식의 2D 스켈레톤 시퀀스를 입력으로 받아 각 시간 프레임에 대응되는 사람 영상을 생성한다. 이 과정에서 참조 이미지는 모든 프레임에서 공유되므로, 전체적으로 외형이 일관된 움직이는 사람 비디오를 생성한다. 본 연구에서는 MusePose의 입력으로 사용하기 위해, MDM으로부터 생성된 3D 관절 시퀀스를 OpenPose 형식의 2D 스켈레톤 시퀀스로 변환하여 사용하였다. 그 후 생성된 사람 외형 이미지가 생성된 동작을 따라서 움직이는 정면 시점 영상을 출력으로 얻는다.

4. Multi-view Video Generation

MusePose로부터 생성된 단일 시점 영상은 자연스러운 동작을 포함하지만, 4D 복원을 위해 필요한 다양한 시점 정보를 제공하지 못한다. 이를 해결하기 위해 본 연구에서는 SV4D[25] 모델을 사용하여 단일 시점 비디오를 다시점 비디오로 확장하였다. SV4D는 다시점 확장에 특화된 Video Diffusion 모델로, 입력 영상 시퀀스와 원하는 카메라 파라미터를 조건으로 받아 각 시간 프레임을 새로운 시점으로 변환한 영상을 생성한다.

생성된 다시점 영상의 수식은 아래와 같다.

$$f_{SV4D}(V, c^k) = V^k = \{I_i^k\}_{t=1}^{T}. \quad (3)$$

여기서 V는 MusePose를 통해 생성한 참조 비디오, c^k는 입력 카메라 파라미터 그리고 V^k는 해당 카메라를 기준으로 생성한 다른 시점의 비디오이다. 이 과정을 통해 4D 재구성 단계에서 필요한 다시점 비디오 정보를 보완할 수 있게 된다.

5. Training Dynamic View Synthesis Model

본 논문의 마지막 단계에서는 SV4D로 생성된 다시점 영상들을 활용하여 시간 축에서 변화하는 사람의 4D 표현을 복원하기 위해서 Grid4D[21]를 사용하였다. Grid4D는 동적 장면의 공간-시간적 정보를 4D decomposed hash encoding과 Gaussian splatting을 기반으로 재구성하는 모델이다. Grid4D는 공간 좌표 (x,y,z)와 시간 t를 해석하기 위해 (x,y,z)를 이용해 공간 정보를, 그리고 (x,y,t), (y,z,t), (x,z,t)를 이용해 시간 정보를 인코딩한다. 학습 과정에서 Grid4D는 정적 Gaussian G와 시간 $t \in [0,1]$를 입력으로 받아 Gaussian의 오프셋을 파이프라인 f_θ를 통해 아래와 같이 예측한다.

$$f_\theta(G,t) = (\triangle x, \triangle r, \triangle s, \triangle c, \triangle o), \quad (4)$$

이를 통해 해당 시점에서의 Gaussian 집합 G_t를 얻는다.

Grid4D의 렌더링 방식은 Gaussian splatting에서 제시된 미분 가능한 렌더링을 사용하며, 아래와 같이 예측 영상 $\widehat{I^k}$와 생성 영상 I^k 사이의 L1 손실 및 Structural Similarity Index Measure (SSIM) 손실, 그리고 grid feature를 부드럽게 정규화하는 손실 L_{reg} 를 결합한 함수를 통해 학습된다.

$$L = |\widehat{I^k} - I^k| + \lambda_1 \times SSIM(\widehat{I^k}, I^k) + \lambda_2 \times L_{reg}, \quad (5)$$

여기서 L_{reg}는 Grid4D에서 제안된 방식으로, 인접한 시공간 좌표에서의 grid feature가 유사하도록 유도하기 위한 손실함수이다. 자세하게는 입력 (x,y,z,t)에 대해 작은 랜덤 노이즈 값인 $(\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_t)$을 더한 인접 좌표와의 feature가 크게 변하지 않도록 하는 손실함수이다. 이를 정의하는 정규화 손실의 수식은 아래와 같다.

$$L_{reg} = \|g(x,y,z,t) - g(x+\epsilon_x, y+\epsilon_y, z+\epsilon_z, t+\epsilon_t)\|_2. \quad (6)$$

여기서, g는 입력된 시공간 좌표에 대해 hash encoder에서 얻은 feature를 의미한다. 결과적으로 이 손실함수는 인접한 Gaussian이 유사한 변형을 갖도록 유도하며, 이러한 정규화 손실은 프레임 간의 temporal

jittering과 3D 상에서의 distortion을 완화하는 데 도움을 준다. 이 과정을 통해 Grid4D는 여러 시점에서 관측된 시간적 변화를 통합하여 임의의 시점 k와 시간 t에서의 이미지를 예측할 수 있는 dynamic novel view synthesis 모델로 학습된다.

결과적으로, 본 연구는 텍스트 입력만으로 동적 사람의 4D 구조를 생성하기 위해 여러 모델을 모듈 형태로 통합한 파이프라인을 제시하였다. 특히, 본 논문에서 제안한 방법은 동작에 관한 문장과 외형에 관한 문장을 분리하여 처리함으로써, 사용자가 정의한 텍스트 조건에 따라 원하는 동작과 외형을 각각 생성하고 이를 유연하게 통합할 수 있는 파이프라인을 구성하였다. 다음 절에서는 제안한 파이프라인의 정량적·정성적 성능을 제시함으로써, 본 방법의 효과성과 실용성을 검증한다.

IV. Experiments

1. Experimental Settings

Table 1. System Environment

Item	Value
OS	Ubuntu 20.04 LTS
CPU	AMD EPYC 7713
GPU	NVIDIA A100 (80GB)
CUDA Version	11.8
Pytorch Version	2.1.1

본 연구의 실험을 위해, Table 1에 제시된 컴퓨팅 자원을 사용하였다. 본 논문에서는 먼저 텍스트로부터 인간 이미지 생성을 위해 Online Stable Diffusion[26]을 사용하였다. 다음으로 MDM을 통해서 120프레임의 인간 동작 시퀀스를 생성하였다. 생성된 이미지와 동작 정보는 MusePose를 통해 융합되어 정면 시점 영상을 생성하였다. 이후, SV4D를 이용한 다시점 확장 단계에서 SV4D에 긴 시퀀스를 그대로 입력할 때 시간적 일관성이 저하되는 경향이 있었기 때문에 120프레임의 동작 시퀀스를 60프레임으로 균등 샘플링하여 다시점 이미지를 생성하였다. 생성되는 다시점 비디오는 총 8개로, 정면 시점 카메라를 기준으로 40도씩 회전하여 카메라를 배치하였다. 마지막으로, Grid4D를 학습하여 시간 축 방향의 선형 보간을 수행함으로써 최종적으로 120프레임의 비디오를 렌더링하였다.

제안한 파이프라인에서 사용된 각 모듈은 공개된 기본 설정을 바탕으로 하였으며 일부 하이퍼파라미터를 조정하였다. 구체적으로, MDM에서 비교 평가용 동작 시퀀스로 32프레임, 전체 평가용 동작 시퀀스로 120프레임을 사용하였다. SV4D에서는 출력 해상도를 512×512로 설정하였고, Grid4D와 4D-GS에서는 시간 축 방향의 그리드 해상도를 기본 설정의 절반으로 조정하였다. 또한 densification은 25,000 iteration까지 수행하였으며, densify를 위한 gradient threshold를 0.0005로 설정하였다. 마지막으로, 학습 손실의 가중치로 λ_1=0.1, λ_2=0.5를 사용하였다.

비교 평가의 공정성을 위해, 베이스라인 방법들은 공개된 기본 설정을 그대로 사용하였다. 다만 평가를 위해 모든 방법의 출력 비디오를 동일한 해상도 512×512, 프레임 수를 32로 통일하였다. 또한 카메라 설정으로는 4D 생성 결과의 novel-view 평가를 위해 정면 시점에서 -10도, +10도씩 회전된 2개의 렌더링 비디오를 사용하였다. 그리고, 베이스라인 방법들은 한 문장을 사용해 4D를 생성하기 때문에, 외형과 동작 문장을 합쳐서 입력으로 사용하였다. 모든 실험의 랜덤 시드는 22로 고정하여 사용하였다.

2. Dataset and Measurements

본 연구에서는 제안한 파이프라인을 검증하기 위해 총 100개의 사람 외형 문장과 25개의 동작 문장을 구성하였다. 실험 수행 시간과 계산 비용을 고려하여 전체 평가 데이터 세트는 총 100개의 외형-동작 페어 데이터로 구성하였다. 구체적으로는 각 동작 문장에 대해서 서로 다른 4개의 외형 문장을 중복 없이 무작위로 매핑하는 방식으로 데이터를 구성하였으며, 이를 통해 각 동작 유형이 균등하게 포함되도록 하였다. Fig. 3은 외형 문장과 동작 문장에 대한 통계를 제시한다. 외형 문장에서는 age, appearance category, ethnicity, gender 측면에서 특정 범주에 과도하게 치우치지 않도록 구성하였다. 그리고 동작 문장 역시 low(“A man is stretching”, “A man does light exercise” 등), medium(“A woman is climbing the stairs”, “A man is throwing defensive punches” 등), high(“A man is running away from something afraid”, “A man is taking a soccer shot” 등) 에 해당하는 강도의 동작이 모두 포함되도록 설계하였다.

일반적으로 4D 생성 작업에서 사용되는 텍스트 설명은 비교적 단순한 문장으로 구성되는 경우가 많으며, 기존 연구들 또한 제한된 수의 텍스트 샘플을 사용하여 정성적, 정량적 평가를 수행한다. 이러한 점을 고려할 때, 본 연구에서 구성한 100개의 외형-동작 페어는 평가에 충분한 규모를 가진다. 한편, 기존 방법론과의 비교 실험에서는 해

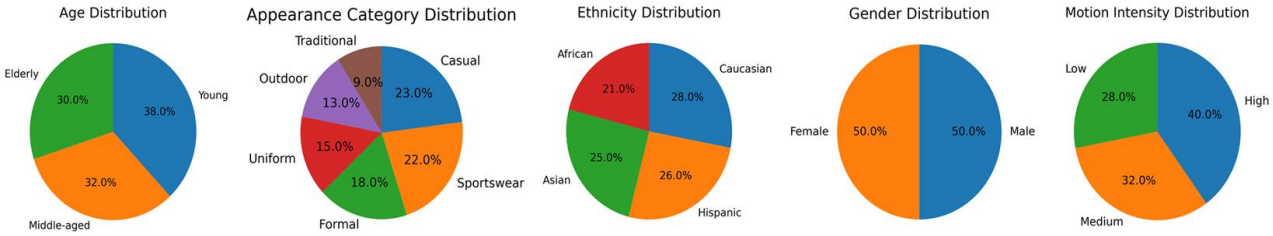

Fig. 3. Attribute distributions of the constructed text dataset.

당 방법들이 단일 샘플 생성에 많은 시간이 소요되는 특성을 고려하여, 전체 데이터 중 20개의 외형-동작 페어를 무작위로 선정한 비교 평가 데이터로 평가를 수행하였다.

Table 2. Examples of text prompts from the constructed dataset.

Appearance	Motion
Caucasian male, young, wearing a chunky mustard-colored turtleneck sweater and blue jeans, paired with brown ankle boots.	A man is running away from something in fear.
African American male, middle-aged, dressed in a red T-shirt and blue jeans with a confident stance.	A man is throwing defensive punches.
Elderly African American male sporting a lightweight, unbuttoned white shirt over a casual tee, matched with light beige pants and brown sandals.	A man is throwing a ball.
Middle-aged East Asian male attired in an orange tank top and gray shorts, finished with black running shoes.	A man is taking a soccer shot.
Hispanic female, middle-aged, geared up for winter sports in a blue ski jacket, teal pants, gloves, and ski boots.	A woman is climbing the stairs.

주어진 문장으로부터 생성된 4D 사람을 평가하기 위해, 본 연구에서는 아래 네 가지 측정 방법을 사용한다. 비디오 생성 작업에서의 성능 평가가 어려운 점을 고려하여, 생성된 비디오의 품질을 측정하기 위해 측정 지표는 VBench[27]에서 제안된 평가 지표들을 활용하였다. VBench는 생성된 비디오를 다양한 측면에서 평가하기 위한 프레임워크로, 동적 표현, 시각적 품질, 그리고 텍스트 조건과의 일치성 등을 정량적으로 측정할 수 있는 지표들을 제공한다. 본 연구에서는 이러한 지표 중 Dynamic Degree, Aesthetic Quality, 그리고 Overall Consistency를 사용하여 생성된 4D 인간 비디오의 성능을 평가하였다. 또한, 각 방법의 생성 시간에 대해 측정하여, 어떤 방법이 효율적으로 비디오를 생성하는지 측정하였다. 측정 지표에 대한 설명과 본 논문에서 수집한 데이터 중 5개의 샘플은 Table 2에 나타냈다.

(1) Dynamic Degree

Dynamic Degree는 생성된 비디오에서 얼마나 큰 움직임이 존재하는지를 평가하는 지표이다. 정적인 비디오의 경우 다른 시간적 품질 지표에서 높은 점수를 얻을 수 있기 때문에, 비디오가 실제로 충분한 동적 변화를 포함하고 있는지를 측정하는 것이 중요하다. 이를 위해 RAFT[28]를 기반으로 optical flow를 이용해 프레임 간의 움직임의 크기를 추정하고, 이를 기반으로 비디오의 동적 정도를 평가한다. 수치가 높을수록 생성된 비디오에 더 큰 움직임이 포함되어 있음을 의미한다.

(2) Aesthetic Quality

Aesthetic Quality 지표는 생성된 비디오의 미적 수준을 평가한다. 이를 위해 LAION Aesthetic Predictor[29]를 사용하여 각 프레임의 미적 점수를 추정한다. 해당 모델은 사람이 평가한 미적 점수를 기반으로 학습되어 색상 조화, 사실주의, 전반적인 예술적 인상 등 각 프레임에 대해 사람이 인지하는 미적 품질을 반영한 점수를 예측할 수 있다. 수치가 높을수록 생성된 비디오의 미적인 수준이 높음을 뜻한다.

(3) Overall Consistency

Overall Consistency는 생성된 비디오와 입력 텍스트 프롬프트 간의 의미론적 대응성을 정량화하기 위해 최근 공개된 ViClip[30]을 사용해 비디오-텍스트 임베딩 유사도를 측정한다. 이는 입력된 문장이 생성된 인간 비디오에 얼마나 정확하게 반영되었는지를 평가한다.

(4) Generation Time

Generation Time은 논문에 제시된 동일한 컴퓨팅 리

Table 3. Quantitative Comparison on Comparative Evaluation Set (20 texts)

비교 평가 데이터 (20개, 32-frame)	Dynamic Degree	Aesthetic Quality	Overall Consistency	Generation Time (전처리 포함/ 샘플당/동일자원)
Dream-in-4D[4]	0%	59.7%	23.6%	13h
4D-fy[3]	5%	60.5%	27.6%	12h
Ours	77.5%	58.4%	24.8%	3h

Table 4. Quantitative Results on Full Evaluation Set (100 texts)

전체 평가 데이터 (100개, 120-frame)	Dynamic Degree	Aesthetic Quality	Overall Consistency
Ours	57.0%	53.2%	24.1%
Ours with 4D-GS[20]	40.4%	38.7%	18.6%

소스 환경에서 각 방법이 주어진 하나의 텍스트 입력으로부터 하나의 결과 비디오를 생성하는 데 소요되는 전처리를 포함한 전체 생성 시간을 기준으로 측정한다. 생성에 걸리는 시간이 짧을수록 효율성이 높음을 의미한다.

3. Quantitative and Qualitative Results

본 논문에서는 제안한 파이프라인의 성능을 정량적으로 평가하기 위하여 앞 절에서 설명된 주요 지표를 활용하였다. Table 3은 비교 평가 데이터에서 테스트한 결과를 보여준다. Dream-in-4D[4]와 4D-fy[3]는 대부분 정적인 결과를 생성해 Dynamic Degree가 각각 0%와 5%로 측정되었으며, 이는 생성된 결과가 상대적으로 정적인 경향을 보이고 있음을 나타낸다. 반면, 제안한 방법은 77.5%로 높은 수준의 시간적 변화를 포함하는 결과를 생성함을 확인하였다. 이는 기존 방법들이 주로 Score Distillation Sampling (SDS) 기반 최적화 방식을 사용해 diffusion prior로부터 guidance를 받아 동작을 생성하는 방법이기 때문에 큰 움직임을 생성하는 데에 제한이 있지만, 본 연구는 동작을 명시적으로 생성한 뒤 이를 외형과 결합하는 구조를 사용하기 때문에 큰 움직임을 효과적으로 반영할 수 있기 때문이다.

한편, 제안한 방법은 동적 표현 측면의 향상과 함께 렌더링 된 비디오의 시각적 품질에 대한 평가인 Aesthetic Quality, Overall Consistency 측면에서는 trade-off가 존재하나 전반적으로 동적 표현을 개선하면서 시각적 품질 간의 균형을 보였다. 생성 시간은 동일한 비디오 길이, 해상도를 고려할 때 논문에서 제안한 방법이 기존 방법 대비 더 높은 생성 효율을 보이고 있음을 확인할 수 있다.

Table 3에서 볼 수 있듯이, 기존 방법들은 SDS를 기반으로 하기에 하나의 샘플을 생성하는 데 많은 시간이 소요된다. 따라서 본 논문에서는 제안한 파이프라인에 대해 전체 평가 데이터 세트를 사용하여 추가 실험을 수행하였다. Table 4는 전체 평가 데이터에 관한 결과를 보여주며, 다양한 텍스트 조건에서도 유사한 경향의 정량적 특성을 관찰하였다. 전반적으로, 제안한 파이프라인은 동작과 외형을 분리하여 생성한 후 결합하는 구조를 통해 기존 방법에 비해 동적 표현을 강화하면서도 외형의 일관성을 유지함을 확인하였다.

Fig. 4에서는 제안한 파이프라인의 단계별 결과와 기존 방법들과의 비교 결과를 함께 제시한다. 위에서부터 생성된 사람 이미지, 동작 생성 결과, 외형-동작 결합 결과, 다시점 생성 그리고 재구성된 4D로부터 새로운 시점 카메라를 이용한 렌더링 결과를 나타낸다. 마지막으로, 기존 방법인 Dream-in-4D와 4D-fy와의 비교 결과인데, novel view 결과를 비교하기 위해서 정면 시점 카메라에서 -10, 10도 회전하여 렌더링하였다. 결과를 비교하였을 때, 기존 방법들은 시간에 따른 객체의 변화가 상대적으로 제한적인 경향을 보이며, 외형 및 동작 문장을 동시에 입력으로 사용하는 구조로 인해 전반적으로 정적인 결과가 생성된다. 이는 정량적 평가에서 Dynamic Degree가 낮게 측정된 결과와도 일치하는 경향이다. 반면, 제안한 방법은 이전 방법들에 비해 시간적인 동작 변화가 뚜렷하게 나타나며, 다양한 동작 조건에서도 안정적인 외형을 유지하는 결과를 확인할 수 있다.

4. Ablation Study

본 연구에서는 4D 복원 단계에서 사용하는 모델에 대한 ablation study를 수행하였다. 공정한 비교를 위해, 동일한 파이프라인에서 마지막 4D 모듈만 4D-GS[20]로 교체하여 비교 실험을 수행하였다. Table 4의 “Ours with 4D-GS”와 같이, 4D 복원에 Grid4D를 사용할 때 더 높은 성능을 보였다. 이는 Grid4D에서의 3D grid 기반의 hash encoding이 2D grid 기반 인코딩 방식인 4D-GS보다 시-

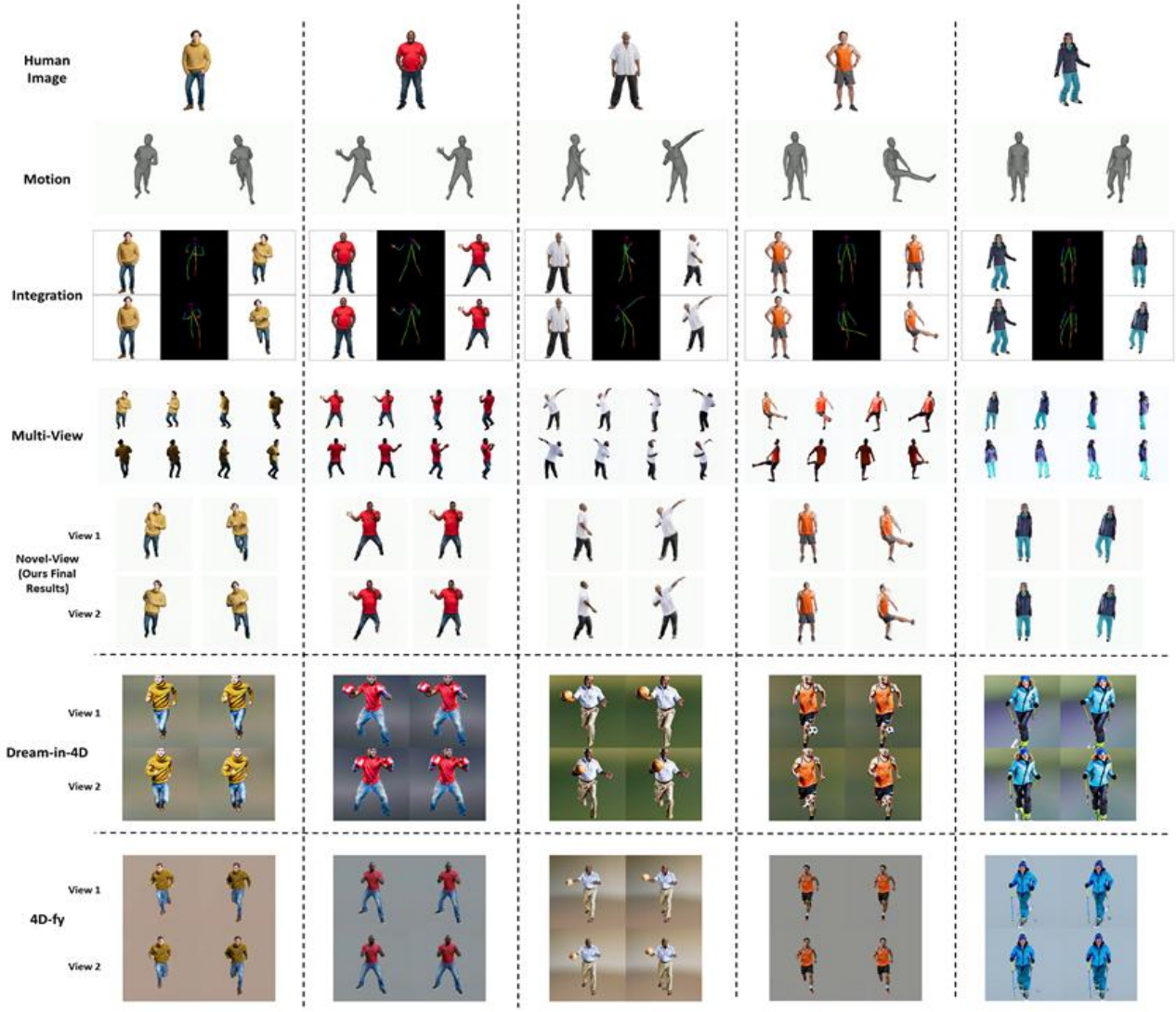

Fig. 4. Qualitative Comparison with previous methods.

공간적인 정보를 더 잘 분리할 수 있어 구조적인 정보와 시간적 변화를 안정적으로 반영하는 데 유리하기 때문이라고 해석할 수 있다. 이를 통해 본 연구에서는 4D 복원 모듈로서 Grid4D가 적합함을 확인하였다.

5. Failure Case

본 연구는 텍스트 기반의 동적 사람 생성을 위한 통합된 구조를 제시했지만, 여전히 몇 가지 한계가 존재한다. 먼저, 본 파이프라인의 중간 결과물은 모두 개별 생성 모델에 의해 생성되므로, 최종 4D 생성 품질은 각 모듈의 성능과 중간 출력물에 직접적으로 영향을 받는다. 그리고, 각 단계가 diffusion model을 포함하는 추론 과정과 optimization 단계로 구성되어 여전히 전체 계산 비용이 많이 든다. 마지막으로, SV4D 단계에서 생성되는 다시점 영상이 정확한 카메라 포즈를 기반으로 하지는 않기 때문에, 4D 재구성에서 공간적인 일관성이 완전하게 유지되지 않는 경우가 있다. 예를 들면, 빠른 전신의 움직임이나 복잡한 동작의 경우 self-occlusion의 영향으로 일부 프레임에서 불안정한 결과를 생성되는 경우가 있었다. 또한, 다시점 영상 생성 결과가 카메라와 정확하게 정렬되지 않는 경우 형상 드리프트가 관찰되기도 한다. 그러나 이러한 한계에도 불구하고 제안한 파이프라인은 텍스트 기반 4D 인간 생성의 가능성과 확장성을 보여주었고, 향후 강한 다시점 일관성 제약이나 pose prior를 통해 보완될 수 있다.

V. Conclusions

본 논문에서는 문장 기반의 4D 인간 생성을 위해 외형과 동작을 분리하여 생성한 후 이를 통합하는 단계적 파이

프라인을 제안하였다. 제안된 방법은 외형 이미지 생성, 동작 생성, 단일 시점 영상 합성, 다시점 확장 및 4D 재구성으로 이어지는 구조를 통해 문장 입력만으로 동적인 인간 모델링이 가능하도록 설계되었다. 특히 외형과 동작을 독립적으로 생성한 뒤 결합함으로써 보다 큰 동작 변화를 안정적으로 반영할 수 있도록 하였다. 이후 단일 시점 영상을 다시점으로 확장하고, 이를 기반으로 Gaussian splatting 기반 변형 네트워크를 학습하여 시공간적으로 일관된 4D 표현을 구성한다. 정성적 및 정량적 평가 결과, 제안한 방법은 일부 품질 지표와 trade-off가 존재하더라도 전반적인 동적 표현 능력의 향상과 시각적 품질 간의 균형을 확인하였다. 이는 외형과 동작을 분리하여 생성하는 구조가 텍스트 기반 4D 인간 생성에서 효과적인 접근 방식임을 보여준다.

향후 연구에서는 기술적 보완을 통해 동작 생성 단계에서 더 복잡하고 장기적인 동작을 생성하는 동시에, 겹침과 같은 문제를 최소화하여 자연스럽고 안정적인 동작 시퀀스를 생성할 수 있을 것이다. 또한 외형과 동작을 분리하여 생성하는 구조의 장점을 활용해 4D 사람의 부분적인 편집이나 조건 변경에 대한 세밀한 제어 기능을 추가하는 방향으로 확장할 수 있을 것이다. 궁극적으로 본 연구는 텍스트 기반 동적 인간 모델링의 새로운 연구 방향을 제시하였으며, 관련 분야의 후속 연구를 촉진함과 동시에 가상 콘텐츠 제작 등 다양한 응용 분야에서 활용될 수 있는 기술로 발전할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

This work was supported by the Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 2E33611).

REFERENCES

[1] J. Ho, A. Jain and P. Abbeel, "Denoising Diffusion Probabilistic Models", Advances in Neural Information Processing Systems (NeurIPS), pp. 6840-6851, 2020. DOI:10.48550/arXiv.2006.11239

[2] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D Using 2D Diffusion", International Conference on Learning Representations (ICLR), 2022. DOI:10.48550/arXiv.2209.14988

[3] S. Bahmani et al., "4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 7996-8006, 2024. DOI: 10.1109/CVPR52733.2024.00764.

[4] Y. Zheng, X. Li, K. Nagano, S. Liu, O. Hilliges and S. De Mello, "A Unified Approach for Text-and Image-Guided 4D Scene Generation," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 7300-7309, 2024. DOI: 10.1109/CVPR52733.2024.00697.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis", Proc. European Conference on Computer Vision (ECCV), pp. 405–421, 2020. DOI:10.1007/978-3-030-58452-8_24

[6] B. Kerbl, G. Kopanas, T. Leimkühler and M. P. Zollhöfer, "3D Gaussian Splatting for Real-Time Radiance Field Rendering", ACM Transactions on Graphics (TOG), Vol. 42, No. 4, pp. 1-14, 2023. DOI:10.1145/3592433

[7] Yu, Heng, et al., "4Real: Towards photorealistic 4d scene generation via video diffusion models." Advances in Neural Information Processing Systems (NeurIPS), pp. 45256-45280, 2024, https://doi.org/10.48550/arXiv.2406.07472

[8] T. Liu, Z. Huang, Z. Chen, G. Wang, S. Hu, L. Shen, H. Sun, Z. Cao, W. Li, and Z. Liu, "Free4D: Tuning-free 4D Scene Generation with Spatial-Temporal Consistency", IEEE/CVF International Conference on Computer Vision (ICCV), 2025. DOI:10.48550/arXiv.2503.20785.

[9] M. M. Loper, A. Mahmood, J. Romero, G. Pons-Moll and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model", ACM Transactions on Graphics (TOG), Vol. 34, No. 6, Article 248, 2015. DOI:10.1145/2816795

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Models From Natural Language Supervision", International Conference on Machine Learning (ICML), PMLR, pp. 8748-8763, 2021. DOI:10.48550/arXiv.2103.00020

[11] M. Petrovich, M. J. Black and G. Varol, "TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, pp. 9454-9463, 2023. DOI: 10.1109/ICCV51070.2023.00870.

[12] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes", International Conference on Learning Representations (ICLR), 2014. DOI:10.48550/arXiv.1312.6114,

[13] J. Zhang et al., "Generating Human Motion from Textual Descriptions with Discrete Representations," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 14730-14740, 2023. DOI:

10.1109/CVPR52729.2023.01415.

[14] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or and A. H. Bermano, "Human Motion Diffusion Model", International Conference on Learning Representations (ICLR), 2023. DOI:10.48550/arXiv.2209.14916

[15] Z. Tang, Y. Bai, Z. Yang, X. Chen, Y. Zhao, J. Zhang, C. Ma, and Z. Wang, "DreamGaussian: Generative Gaussian Splatting for 3D Content Creation", Proc. International Conference on Learning Representations (ICLR), 2024. DOI:10.48550/arXiv.2309.16653

[16] X. Yang, H. Shi, B. Zhang, F. Yang, J. Wang, H. Zhao, X. Liu, X. Wang, Q. Lin, J. Yu, L. Wang, J. Xu, Z. He, Z. Chen, S. Liu, J. Wu, Y. Lian, S. Yang, Y. Liu, Y. Yang, D. Wang, J. Jiang, and C. Guo, "Hunyuan3D 1.0: A Unified Framework for Text-to-3D and Image-to-3D Generation", arXiv preprint arXiv:2411.02293, 2024. DOI:10.48550/arXiv.2411.02293

[17] A. Pumarola, E. Corona, G. Pons-Moll and F. Moreno-Noguer, "D-NeRF: Neural Radiance Fields for Dynamic Scenes," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 10313-10322, 2021. DOI: 10.1109/CVPR46437.2021.01018.

[18] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, "HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields", ACM Transactions on Graphics (TOG), Vol. 40, No. 6, pp. 1-12, Article 238, 2021. DOI:10.1145/3478513.3480487

[19] J. Luiten, G. Kopanas, B. Leibe and D. Ramanan, "Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis," 2024 International Conference on 3D Vision (3DV), Davos, Switzerland, pp. 800-809, 2024. DOI: 10.1109/3DV62453.2024.00044.

[20] G. Wu et al., "4D Gaussian Splatting for Real-Time Dynamic Scene Rendering," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 20310-20320, 2024. DOI: 10.1109/CVPR52733.2024.01920.

[21] J. Xu, Z. Fan, J. Yang, and J. Xie, "Grid4D: 4D Decomposed Hash Encoding for High-Fidelity Dynamic Gaussian Splatting", Advances in Neural Information Processing Systems (NeurIPS), pp. 123787-123811, 2024. DOI:10.52202/079017-3934

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 10674-10685, 2022. DOI: 10.1109/CVPR52688.2022.01042.

[23] Z. Tong, C. Li, Z. Chen, B. Wu and W. Zhou, "MusePose: A Pose-Driven Image-to-Video Framework for Virtual Human Generation", 2024. https://github.com/TMElyralab/MusePose

[24] Z. Cao, T. Simon, S. -E. Wei and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1302-1310, 2017. DOI: 10.1109/CVPR.2017.143.

[25] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang and V. Jampani, "SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency", arXiv preprint arXiv:2407.17470, 2024, DOI:10.48550/arXiv.2407.17470

[26] Online Stable Diffusion, https://stablediffusionweb.com/ko

[27] Z. Huang et al., "VBench: Comprehensive Benchmark Suite for Video Generative Models," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 21807-21818, 2024. DOI: 10.1109/CVPR52733.2024.02060.

[28] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," European Conference on Computer Vision (ECCV), pp. 402–419, 2020, DOI:10.1007/978-3-030-58536-5_24.

[29] LAION-AI, "LAION Aesthetic Predictor V1", 2022. https://github.com/LAION-AI/aesthetic-predictor

[30] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, Z. Huang, W. Zhang, J. Zhu, H. Zhang, J. Gao, and L. Wang, "InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation", International Conference on Learning Representations (ICLR), 2024. DOI:10.48550/arXiv.2307.06942

Authors

Chanwoo Kim received his B.S. degree in Computer Science from Kyungpook National University in 2022. He is currently a M.S. student in AI-Robotics at University of Science and Technology – Korea Institute of Science and Technology, Republic of Korea. His research interests lie in Computer Vision and 4D Generation.

Sanghun Kim received his B.S. degree from Kyung Hee University, where he also completed his M.S. degree with a focus on computer vision and artificial intelligence. He is currently a researcher at Broz. His research interests include computer vision and generative AI.

Hwasup Lim received his Ph.D. in Electrical Engineering from The Pennsylvania State University in 2007. From 2007 to 2011, he was an R&D staff member at the Advanced Media Lab, Samsung Advanced Institute of Technology (SAIT), where he focused on time-of-flight (TOF) depth image processing and 3D human modeling & tracking. He is currently a Principal Researcher at the Center for Artificial Intelligence Research, Korea Institute of Science and Technology (KIST), and a Professor in the AI-Robotics program at the KIST School, University of Science and Technology (UST). His research interests include 3D human modeling, generative AI, and vision-language-action frameworks for AI-driven robotic systems.