단일 이미지 내 다중 피사체의 동적 행동 합성을 위한 개념 기반 개인화 기법연구

박찬훈, 백승준*

고려대학교

cksgns2010@korea.ac.kr, *sjbaek@korea.ac.kr

Concept Driven Multi-Subject Personalization for Dynamic Action Synthesisfrom a Single Image

Chan Hun Park, Seung Jun Baek*

Korea Univ.

요 약

본 논문은 단일 참조 이미지 내 다중 피사체의 정체성(Identity)을 유지하면서도, 텍스트 지시에 따른 새로운 자세와 상호작용을 생성하는 개인화 기법을 다룬다. 기존의 마스크 기반 정렬 학습은 피사체의 형상과 윤곽에 과도하게 과적합(Overfitting)되는 경향이 있으며, 이는 동적 표현의 다양성을 저해하는 '구조적 경직성' 문제를 야기한다. 이를 해결하기 위해 본 연구는 사전지식 기반 어텐션 정렬(Prior-Guided Attention Alignment)을 제안하며, 이 기법은 피사체 토큰의 주의(Attention) 분포를 의미론적으로 안정화하는 역할을 한다. 또한 데이터 희소성을 극복하고자 개념 주도 생성적 증강(Concept-Driven Generative Augmentation) 및 SDE 기반 의미론적 편집(SDE-based Semantic Editing) 파이프라인을 구축하였다. 이를 통해 정체성 유지와 텍스트 정합성을 동시에 강화한 학습 데이터를 확보하였으며, 결과적으로 외형 보존과 동적 행동 합성 간의 균형을 달성하였다.

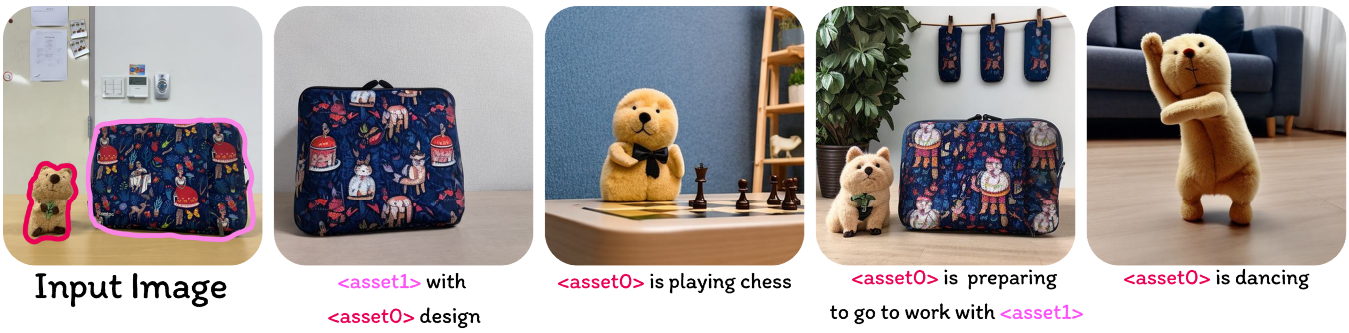

그림 1 본 연구의 생성 결과 예시. 단일 참조 이미지(좌측)로부터 학습된 피사체들이 텍스트 프롬프트에 따라 정체성을 유지하며 다양한 동적 행동을 수행하는 모습을 보여준다.

Ⅰ. 서 론

본 논문은 텍스트-조건부 확산(Diffusion) 모델의 발전과 함께 개인화(Personalization) 기술이 실사용 단계로 확장되는 흐름 속에서, 단 한 장의 이미지에 포함된 다중 피사체를 대상으로 하는 동적 합성 문제를 다룬다[1]. 기존 개인화 연구는 데이터가 제한된 환경에서 피사체의 외형을 복원하기 위해 참조 이미지에 모델을 과적합(Overfitting)시키는 방식에 의존해 왔다[2, 3]. 이로 인해 생성된 결과물이 원본 이미지의 정적인 구도와 배경에 고착되는 한계가 빈번히 발생한다. 특히 다중 피사체 분리를 위해 사용되는 마스크 기반 학습은 피사체의 의미보다 픽셀 단위 형상에 집중하도록 유도하는데[4], 이러한 방식은 텍스트 프롬프트가 요구하는 역동적인 변화를 수용하지 못하게 만든다.

이에 본 연구는 구조적 유연성을 확보하고 과적합을 방지하기 위해 다음과 같은 세 가지 핵심 전략을 제안한다.

(1) 사전지식 기반 어텐션 정렬(Prior-Guided Attention Alignment)을 도입한다. 이는 일반 클래스(Coarse Class)가 갖는 구조적 사전지식을 정규화 신호로 활용한다. 이를 통해 피사체 토큰이 외형적 디테일은 유지하되, 형태적 변형 가능성을 확보하도록 유도한다.

(2) 개념 주도 생성적 증강(Concept-Driven Generative Augmentation)을 수행한다. 비전-언어 모델(VLM)과 거대 언어 모델(LLM)을 연계하여 피사체의 핵심 개념을 추출하고, 단일 이미지에 부재한 다양한 행동 및 상호작용 시나리오를 확장한다[5].

(3) SDE 기반 의미론적 편집(SDE-based Semantic Editing)을 적용한다. 제로샷 분할과 인페인팅 기술을 결합하여 생성된 초안을 정제하는 과정이다[6, 7]. 이를 통해 원본의 정체성과 텍스트의 동적 지시가 양립하는 고품질 증강 데이터를 구축하고 학습에 활용한다.

Ⅱ. 본론

본 논문에서는 단일 참조 이미지 기반 개인화 생성에서 발생하는 정보

불완전성과, 정체성 보존과 동적 표현력 사이의 구조적 균형을 동시에 달성하는 통합 프레임워크를 제안한다. 핵심 가정은 피사체를 단순한 픽셀 집합이 아니라, 특정 의미론적 클래스(Semantic Class) 내부에 존재하는 특수 변형체로 해석하는 것이다. 이 관점은 모델이 참조 이미지의 특정 포즈·배경에 과적합되는 현상을 억제하고, 잠재 공간(latent space)에서 피사체가 취할 수 있는 동작 분포를 확장하도록 유도한다.

단일 이미지 학습에서 자주 관찰되는 '구조적 경직성(structural rigidity)'은, 모델이 피사체의 기하학적 형상 자체를 정체성의 필수 요소로 과도하게 동일시하는 데서 비롯된다. 이를 해결하기 위해 본 연구는 사전지식 기반 어텐션 정렬(Prior-Guided Attention Alignment)을 도입한다. 구체적으로, 사전학습된 생성 모델이 이미 학습한 일반 클래스(예: "panda") 토큰의 어텐션 맵을 구조적 앵커(structural anchor)로 사용하고, 학습 과정에서 피사체 식별자(placeholder) 토큰의 어텐션 분포가 해당 클래스 토큰의 분포와 위상적으로 정렬되도록 목적함수를 설계한다. 이 정렬은 피사체의 텍스처·스타일 단서는 유지하면서도, 특정 포즈에 결박되는 기하학적 제약을 완화한다. 결과적으로, 픽셀 수준 복원(reconstruction)에 치우쳐 포즈 전이가 어려웠던 기존 방식과 달리, 본 방법은 의미론적 정렬을 통해 다양한 자세·행동으로의 전이를 가능하게 한다.

또한 단일 이미지로는 확보하기 어려운 동작·상호작용 데이터를 보강하기 위해, 멀티모달 모델의 생성 능력을 활용한 개념 주도 생성적 증강(Concept-Driven Generative Augmentation) 파이프라인을 제안한다. 먼저 LLM/VLM[5]을 이용해 피사체의 핵심 속성을 추출하고, 이를 조건으로 상호작용이 포함된 다양한 시나리오 프롬프트를 생성한다. 그러나 텍스트 증강만으로는 시각적 일관성과 정체성 보존을 동시에 보장하기 어려우므로, 본 연구는 SDE 기반 의미론적 편집(SDE-based Semantic Editing)을 결합한다. 이는 시나리오에 맞춰 인페인팅된 초안 이미지에 노이즈를 주입한 뒤, 증강 프롬프트로 다시 디노이징하는 과정을 통해 원본 정체성을 유지하면서도 프롬프트가 요구하는 행동이 자연스럽게 융합된 고품질 합성 데이터를 생성함으로써 구현되었다.. 이로써 모델은 확장된 데이터 분포를 학습하여, 학습에서 관찰하지 못한 동작·상호작용을 제로샷(Zero-shot)으로 합성할 수 있는 능력을 갖춘다.

제안 방법의 효과를 검증하기 위해 Break-a-Scene과 COCO 등 총 15개 벤치마크에서 실험을 수행하였다. 정량 평가는 정체성 유지(CLIP-I)와 텍스트 정합성(CLIP-T, ImageReward[8])을 중심으로 진행했으며, 비교 기법으로 Textual Inversion[2], DreamBooth[3], Break-a-Scene[4]을 사용하였다. 결과적으로 DreamBooth는 참조 이미지 과적합으로 CLIP-I가 높았지만, 새로운 행동 지시를 충분히 반영하지 못하고 원본 포즈를 복제하는 경향이 있었다. Textual Inversion은 텍스트 정합성에서 일부 이점을 보였으나, 시각적 디테일 보존에는 한계를 보였다.
또한 다중 피사체 조합에 강점이 있는 Break-a-Scene과 비교해도, 본 방법은 '행동(Action)' 및 '상호작용(Interaction)' 프롬프트에서 유의미한 우위를 확인하였다. 정량적으로 ImageReward와 CLIP-T에서 최고 성능을 달성했으며, 이는 사전지식 기반 어텐션 정렬과 생성적 증강이 정체성 보존과 동적 표현력의 균형을 효과적으로 달성했음을 시사한다.

Ⅲ. 결론

본 논문은 데이터 희소성이 극대화된 단일 참조 이미지 환경에서도, 다중 피사체의 정교한 정체성 보존과 자유로운 동적 변형을 동시에 달성하는 새로운 개인화 프레임워크를 제안하였다. 특히 피사체 형상 구속을 완화하기 위해 사전지식 기반 어텐션 정렬(Prior-Guided Attention Alignment)을 도입하고, 일반 클래스의 구조 정보를 앵커로 활용해 픽셀 단위 과적합을 줄이며 의미론적 유연성을 확보하였다. 또한 개념 주도 생성적 증강(Concept-Driven Generative Augmentation)과 SDE 기반 의미론적 편집(SDE-based Semantic Editing)으로 단일 이미지가 담지 못하는 상호작용·행동 패턴을 가상 공간에서 확장해 학습에 주입함으로써, 제한된 데이터의 표현력 한계를 보완하였다. 비교 실험 결과, 제안 기법은 기존 마스크 기반 접근의 구조적 경직성을 완화하고 행동(Action) 및 상호작용(Interaction) 합성에서 높은 텍스트 정합성과 정체성 유지 성능을 동시에 달성함을 보였다.

ACKNOWLEDGMENT

This work was supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2025-RS-2020-II201819)
and by the National Research Foundation of Korea (NRF) grant funded by the Kore government(MSIT)(RS2022-NR070834).

참 고 문 헌

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., "High-resolution image synthesis with latent diffusion models," Proc. IEEE/CVF CVPR, pp. 10684 - 10695, 2022.

[2] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D., "An image is worth one word: Personalizing text-to-image generation using textual inversion," arXiv preprint arXiv:2208.01618, 2022.

[3] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," Proc. IEEE/CVF CVPR, pp. 22500 - 22510, 2023.

[4] Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., and Lischinski, D., "Break-a-scene: Extracting multiple concepts from a single image," SIGGRAPH Asia 2023 Conference Papers, pp. 1 - 12, 2023.

[5] Achiam, J., Adler, S., Agarwal, S., et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[6] Kirillov, A., Mintun, E., Ravi, N., et al., "Segment Anything," Proc. IEEE/CVF ICCV, pp. 4015 - 4026, 2023.

[7] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S., "Sdedit: Guided image synthesis and editing with stochastic differential equations," arXiv preprint arXiv:2108.01073, 2021.

[8] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y., "Imagereward: Learning and evaluating human preferences for text-to-image generation," Advances in Neural Information Processing Systems, 36, 2024.