한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 3, P. 109~119

InDepthVOI: 결합 생성을 통한 기하적 특징을 인식하는 비디오 객체 삽입

한규진° 강경국 조성현*

포항공과대학교

{kyujin, kkang831, s.cho}@postech.ac.kr

InDepthVOI: Geometry-Aware Video Object Insertion via Joint Generation

Kyujin Han° Kyoungkook Kang Sunghyun Cho*

Pohang University of Science and Technology (POSTECH), South Korea

요 약

비디오 객체 삽입(Video Object Insertion, VOI)은 사용자가 지정한 객체를 기존 비디오에 자연스럽게 합성하면서, 장면의 일관성을 유지하고 주변 환경과 자연스럽게 상호작용하는 사실적인 객체 움직임을 생성하는 것을 목표로 한다. 최근 관련 연구가 활발히 진행되고 있으나, 고품질 VOI를 구현하는 것은 여전히 어려운 문제로 남아 있다. 기존 방법들은 대부분 장면의 기하 정보를 명시적으로 고려하지 않은 채 RGB 도메인(domain)에서만 객체 삽입을 수행하기 때문에, 깊이(depth) 모호성, 부정확한 오클루전(occlusion) 추론, 그리고 비현실적인 객체–장면 상호작용 문제가 발생한다. 또한 현재의 프레임워크들은 객체 움직임에 대한 제어 능력이 제한적이며, 깊이 축 방향의 움직임을 효과적으로 모델링하지 못한다. 본 논문에서는 생성 과정에 기하학적 추론을 명시적으로 통합한 깊이 인식 기반 비디오 객체 삽입 프레임워크인 InDepthVOI을 제안한다. 제안 방법은 RGB 외형 정보와 깊이 표현을 공동으로 모델링하는 듀얼브랜치(dual-branch) 디퓨전 트랜스포머(diffusion transformer) 구조를 사용한다. 또한 삽입 객체의 시공간적 이동 경로를 영상 평면과 깊이 차원에서 동시에 표현하는 구조화된 조건 신호인 깊이 궤적 맵(depth trajectory map)을 도입하여, 깊이 일관성을 유지하는 객체 움직임 제어를 가능하게 한다. InDepthVOI의 학습을 위해, 본 연구에서는 정제된 배경–합성 비디오 쌍과 구조화된 깊이 궤적 맵으로 구성된 새로운 데이터셋을 구축하였다. 다양한 실험 결과를 통해, 제안 방법은 시각적 사실성과 제어 가능성 측면에서 기존 방법들을 크게 능가하며, VOI 분야의 최고 수준의 성능을 달성하였다.

Abstract

Video object insertion (VOI) aims to seamlessly composite a user-specified object into an existing video while preserving scene consistency and generating realistic object motion that interacts naturally with the surrounding environment. Despite recent progress, achieving high-quality VOI remains challenging. Most existing approaches perform insertion purely in the RGB domain without explicitly modeling scene geometry, which often leads to depth ambiguity, incorrect occlusion reasoning, and unrealistic object–scene interactions. Moreover, current frameworks provide only limited control over object motion and cannot model movement along the depth axis. In this paper, we propose InDepthVOI, a depth-aware video object insertion framework that explicitly incorporates geometric reasoning into the generation process. Our method employs a dual-branch diffusion transformer that jointly models RGB appearance and depth representations. We further introduce a depth trajectory map, a structured conditioning signal that encodes the spatio-temporal trajectory of inserted objects across both the image plane and depth dimensions, enabling depth-consistent motion control. To train InDepthVOI, we construct a new dataset comprising carefully curated background–composite video pairs and structured depth trajectory annotations. Extensive experiments demonstrate that InDepthVOI significantly outperforms existing methods in both visual realism and controllability, establishing a new state of the art for VOI.

키워드: 비디오 객체 삽입, 객체 삽입 데이터셋, 깊이 궤적 맵, 결합 생성
Keywords: Video Object Insertion, Object Insertion Dataset, Depth Trajectory Map, Joint Generation

*corresponding author: Sunghyun Cho / Pohang University of Science and Technology (POSTECH), South Korea (s.cho@postech.ac.kr)

Received : 2026.05.27./ Review completed : 1st 2026.06.07. 2nd 2026.06.25./ Accepted : 2026.06.29.
DOI : 10.15701/kcgs.2026.32.3.109
ISSN : 1975-7883(Print)/2383-529X(Online)

1 서론

최근 생성 모델 [1, 2, 3]의 발전은 사실적인 비디오 합성을 가능하게 하며 시각 콘텐츠 생성 분야를 빠르게 변화시키고 있다. 특히 텍스트(text) 기반 비디오 생성 모델 [4, 5, 6]의 성공을 바탕으로, 최근에는 움직임 정보나 깊이(depth)와 같은 구조적 가이던스(guidance)를 포함한 다양한 조건 신호를 활용하는 사용자 친화적인 비디오 생성 프레임워크 [7]로 연구가 확장되고 있다.

비디오 객체 삽입(Video Object Insertion, VOI)은 이러한 흐름 속에서 등장한 연구 주제로, 사용자가 지정한 객체를 입력 비디오에 자연스럽게 삽입하는 것을 목표로 한다. VOI는 기존 비디오에 새로운 객체를 점진적으로 추가함으로써 사용자가 원하는 장면을 직관적으로 구성할 수 있도록 한다. 그러나 고품질의 VOI를 구현하는 것은 여전히 어려운 문제이다. 이는 입력 비디오의 배경 영역을 유지하면서도 삽입 객체의 자연스러운 움직임과 주변 장면과의 현실적인 상호작용을 동시에 생성해야 하기 때문이다.

기존 연구들은 VOI 문제를 해결하기 위해 다양한 접근법을 제안하였다. 학습 없이(training-free) 동작하는 방법들 [8, 9]은 사전 학습된 텍스트 기반 비디오 생성 모델의 디노이징 과정을 텍스트 프롬프트를 통해 제어한다. 그러나 이러한 방법들은 원하는 객체를 안정적으로 생성하지 못하거나, 원본 비디오와 크게 다른 결과를 생성하는 경우가 많다. 최근에는 학습 기반 접근법 [10, 11, 7, 12, 13, 14, 15]이 배경 비디오와 합성 비디오 간의 매핑을 학습하는 방식으로 발전하였다. 이러한 비디오 쌍(pair)을 구축하는 과정에서, *Bian et al.* [11]은 합성 비디오에서 객체 영역을 확장한 후 제거하여 배경 비디오를 생성하였고, *Zi et al.* [15]은 전경의 객체를 제거한 뒤 비디오 인페인팅(inpainting)을 수행하여 배경 영역을 복원하였다. 또한 *Jin et al.* [13]은 대규모 학습을 위해 가상 3D 장면들로부터 배경–합성 비디오 쌍을 생성하였다.

그럼에도 불구하고, 기존 방법들은 복잡한 실제 환경에서 여전히 현실적인 결과를 생성하는 데 한계를 가진다. Figure 1에서 볼 수 있듯이, 삽입된 객체가 기존 장면 구조를 관통하거나 물리적 · 기하학적 제약을 무시하는 현상이 자주 발생한다. 이러한 문제의 주요 원인은 장면 기하 구조에 대한 명시적인 모델링의 부재에 있다. 대부분의 기존 방법들은 깊이와 같은 기하학적 정보를 고려하지 않은 채 RGB 공간에서만 객체 삽입을 수행한다. 그 결과, 삽입 과정에서 장면 구조에 대한 명시적인 이해가 부족해지며, 삽입 객체와 배경 간의 깊이 모호성, 부정확한 오클루전(occlusion) 추론, 그리고 장면과의 시각적으로 부자연스러운 상호작용 문제가 발생한다. 또한 기존 VOI 프레임워크 [10, 12]은 객체 움직임에 대한 제어 능력이 제한적이다. 대부분의 방법들은 움직임의 궤적(trajectory)과 같은 단순한 단서들에 의존하여 삽입 객체의 시간적 변화를 제어한다. 이러한 신호는 단순한 움직임 패턴을 표현할 수 있으나, 객체가 이미지 평면상에서 수평적으로 이동할 뿐만 아니라 깊이 축 방향으로도 이동하는 복잡한 상황을 충분히 표현하지 못한다. 이 두 가지 한계는 기존 VOI 방법들의 현실성과 제어 가능성을 동시에 제한한다.

이러한 문제를 해결하기 위해, 본 논문에서는 기하학적 추론을 명시적으로 도입하여 깊이를 인지하는 비디오 객체 삽입 프레임워크인 InDepthVOI를 새롭게 제안한다. 제안하는 방법은 객체의 외형과 깊이 정보를 공동으로 모델링함으로써 장면과의 깊이 일관성을 유지하는 객체 삽입을 가능하게 한다. 이를 위해 외형과 기하 정보를 각각 독립적인 브랜치(branch)에서 모델링하는 듀얼 브랜치 디퓨전 트랜스포머 구조를 도입하였다. 제안한 모델은 객체의 외형을 생성하는 RGB 브랜치와 구조적인 기하(geometry) 정보를 표현하는 깊이 브랜치로 구성된다. 두 브랜치는 디노이징(denoising) 과정에서 결합 교차 어텐션(joint cross-attention) 방법 [16]을 통해 상호작용하며, 외형 특징과 기하적 특징을 정렬(alignment)한다. 또한 본 연구에서는 깊이 궤적 맵이라는 새로운 표현 방식을 제안한다. 이는 삽입 객체의 시공간적 움직임을 이미지 평면뿐만 아니라 깊이 차원까지 포함하여 표현한다. 생성 과정에서 해당 정보를 조건으로 활용함으로써, InDepthVOI는 객체 움직임에 대한 세밀한 제어를 가능하게 하며, 객체가 장면에 대해 이미지 평면 상의 이동뿐 아니라 깊이 축 방향으로도 자연스럽게 움직일 수 있도록 한다.

추가적으로, 본 연구에서는 실제 환경을 기반으로 제작한 VOI 데이터셋인 RealVOI를 제안한다. 이는 물리적으로 타당한 객체 삽입 학습을 지원하도록 설계된 데이터셋이다. 기존 공개 VOI 데이터셋들이 가상 환경 내 정적인 객체에 초점을 맞추거나 [13], 객체 제거 과정에서 그림자나 반사와 같은 시각적 흔적을 충분히 제거하지 못하는 한계 [15, 11]를 가지는 반면, RealVOI는 실제 환경 기반의 배경-합성 비디오 쌍을 제공하며 객체 및 관련 시각 효과를 깨끗하게 제거한 데이터를 포함한다. 이를 위해 공개 비디오 소스로부터 합성 비디오를 수집하고, 공개된 객체 제거 모델을 사용하여 객체와 관련 시각 효과가 제거된 배경 비디오를 획득하였다. 이후 비전 언어 모델(vision-language model; VLM)을 활용한 자동 필터링과 사람의 검증 과정을 통해 고품질의 배경 비디오를 구축하였다. 또한 각 배경-합성 비디오 쌍에 대해 삽입 객체의 시공간적 움직임을 이미지 평면 및 깊이 차원에서 표현하는 깊이 궤적 맵을 생성하였다. 이 정보는 명시적인 기하학적 방향성(supervision)을 제공하며, InDepthVOI가 깊이 일관성을 유지하는 객체 삽입 및 움직임을 학습할 수 있도록 한다.

본 논문의 기여는 다음과 같다.

- 기하학적 추론을 생성 과정에 명시적으로 도입하여 깊이 일관성을 유지하는 객체 삽입과 세밀한 움직임 제어를 가능하게 하는 새로운 비디오 객체 삽입 프레임워크 InDepthVOI를 제안한다.
- 실제 환경 기반의 새로운 VOI 데이터셋을 제안하며, 정제된 배경-합성 비디오 쌍과 구조화된 깊이 궤적 맵을 제공한다.
- 다양한 실험을 통해 InDepthVOI가 기존 방법 대비 더욱 현실적인 비디오 생성과 우수한 성능을 달성함을 입증한다.

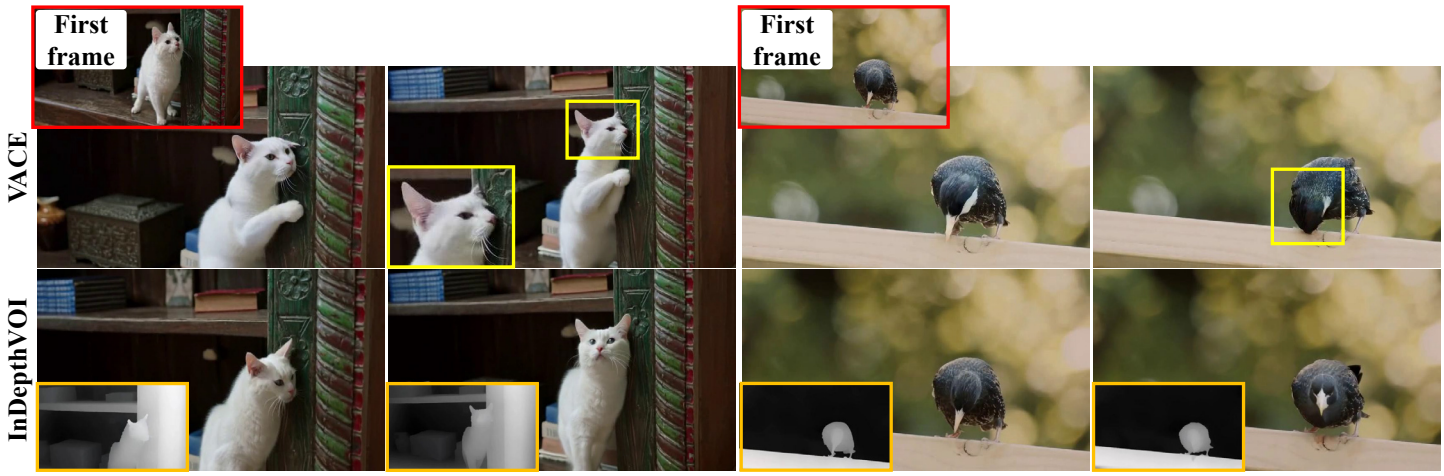

Figure 1: Qualitative comparison of different methods. VACE [7], a state-of-the-art method for video editing, does not explicitly model depth conditions, often causing depth ambiguity and unrealistic object–scene interactions, as highlighted in the yellow boxes. In contrast, InDepthVOI jointly generates object appearance and relative depth, enabling geometry-aware scene integration. Depth maps generated by our method are also shown in the orange boxes.

2 관련 연구

2.1 비디오 객체 삽입

기존 연구들은 VOI 문제를 해결하기 위해 다양한 접근법을 제안하였다. 학습 없이 동작하는 방법들 [8, 9]은 사전 학습된 디퓨전 모델의 중간 어텐션 레이어(layer)에 객체 관련 잠재 벡터를 주입함으로써 추가적인 학습 없이 객체 삽입을 수행한다. 한편, 비디오 인페인팅 기반 방법 [11]은 VOI를 마스킹된 비디오(masked video) 복원 문제로 다루며, 입력 비디오에 마스크 시퀀스(mask sequence)를 적용한 뒤 생성 모델을 통해 마스킹된 영역을 복원한다. 그러나 이러한 방법들은 삽입 객체의 움직임을 제어하기 어렵다는 한계를 가진다.

이를 해결하기 위해, 움직임 기반(motion-guided) VOI 방법들 [10, 12]은 모션 궤적 또는 단서(cue)를 조건으로 활용하여 삽입 과정에서 시간적 일관성과 제어 가능성을 향상시킨다. 최근 OmniInsert [14]는 유연한 형태의 객체 삽입을 위해 체계적인 학습 단계와 마스크 없이(mask-free) 작동하는 프레임워크를 제안하였다. 또한 InsertAnywhere [13]와 LoVoRA [17]는 정적인 객체에 대해 명시적인 객체 마스크 시퀀스를 획득하고, 4D 장면 복원 또는 광학 흐름 추정을 활용하여 객체 삽입을 가이드한다.

그러나 기존 VOI 방법들은 대부분 RGB 도메인에 국한되어 있으며, 깊이와 같은 구조적 표현에 대한 명시적인 모델링이 부족하다. 반면, 본 연구에서는 기하적 단서와 RGB 외형을 공동으로 모델링함으로써 깊이를 인지하는 비디오 객체 삽입을 가능하게 한다.

2.2 결합 디퓨전 모델

조건 기반 이미지 및 비디오 생성에 대한 관심이 증가함에 따라, 결합 디퓨전 모델에 대한 연구 또한 활발히 이루어지고 있다. 이미지 생성 분야에서는, JointNet [18]이 사전 학습된 디퓨전 모델에 밀접하게 연결된 모달리티 브랜치를 추가하여 RGB 이미지와 깊이 맵을 공동으로 모델링하였다. UniCon [16]은 RGB 이미지와 해당 조건 간의 결합 분포를 효과적으로 학습하기 위해 결합 교차 어텐션 방법을 제안하였다. 또한 JointDiT [19]은 두 모달리티 간에 불균형한 시간 단계 샘플링(timestep sampling) 전략을 도입하여 다양한 조합적 생성 작업을 보다 효과적으로 수행할 수 있도록 하였다.

비디오 생성 분야에서는 MM-Diffusion [20]과 JavisDiT [21]이 멀티모달 어텐션을 통해 비디오와 오디오 모달리티를 공동 학습함으로써 시공간적 지식 동기화(prior synchronization)를 수행하였다. IDOL [22]은 사람 중심 비디오와 깊이를 공동 생성하기 위한 통합된 듀얼모달리티(dual-modality) 디퓨전 트랜스포머를 제안하였으며, 자세(pose) 시퀀스를 조건으로 활용하여 사람의 움직임을 제어한다.

그러나 이러한 결합 비디오 디퓨전 생성 연구들에도 불구하고, 기존 방법들은 비디오 객체 삽입 문제를 고려하지 않는다. 특히 기존 연구들은 조건 신호를 활용한 생성 또는 여러 모달리티의 공동 모델링에 초점을 맞추고 있으며, 구조화된 조건 내에 새로운 객체를 삽입하는 문제는 다루지 않는다. 반면 본 연구는 RGB 비디오와 깊이 비디오에서의 객체 삽입을 공동으로 수행함으로써 이러한 한계를 해결하고자 한다.

3 방법

Figure 2은 현실적인 VOI를 위한 듀얼브랜치 디퓨전 트랜스포머 프레임워크인 InDepthVOI의 전체 파이프라인을 보여준다. 제안하는 프레임워크는 객체의 외형 장면의 기하적 정보를 명시적으로 모델링하며, 두 표현 간의 상호작용을 통해 이를 점진적으로 정렬한다.

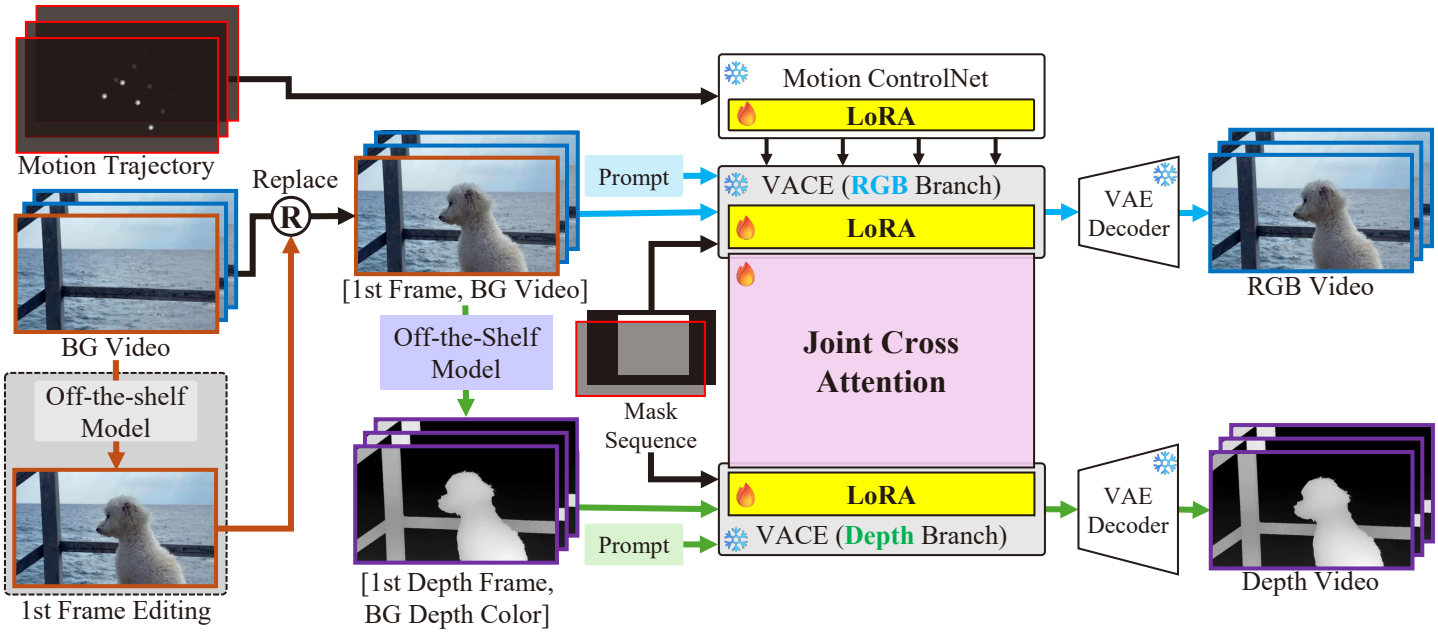

Figure 2: Overview of the propsed InDepthVOI. **(a)** Overall framework of InDepthVOI, which performs video object insertion jointly in both depth and RGB modalities.

Dataset	Composite Video	BG Video	BG Video (Visual effect removed)	0-1 Mask	Dynamic Object	Real-World Video	# of Frames
Senorita-2M [15]	✓	✓	✗	✓	✓	✓	33–64
VPData [11]	✓	✗	✗	✓	✓	✓	81<
ROSE++ [13]	✓	✓	✓	✓	✗	✗	81<
RealVOI (Ours)	✓	✓	✓	✓	✓	✓	81

Table 1: Comparison with existing open-source VOI datasets. Our dataset offers realistic background videos with visual effects removed, and dynamic object annotations, enabling realistic VOI.

RGB 브랜치는 대상 객체가 제거된 배경 비디오와 공개된 이미지 편집 모델 [23, 24]을 이용하여 객체를 삽입한 첫 번째 프레임을 입력으로 받아, 시간적으로 일관된 합성 비디오를 생성한다. 동시에 깊이 브랜치는 공개된 깊이 추정 모델 [25]을 통해 편집된 첫 번째 프레임의 깊이 맵과 배경 깊이 시퀀스를 이용하여 합성 깊이 비디오를 생성한다. 추가로, 두 브랜치에 객체의 간단한 위치 정보를 알려주는 바운딩 박스(bounding box) 형태의 마스크 시퀀스가 들어간다. 마스크 시퀀스의 첫 번째 프레임은 항상 0으로, 첫 번째 프레임에 주어진 객체 정보가 변형되는 것을 방지한다.

RGB 브랜치가 텍스처(texture), 색깔, 조명(illumination)과 같은 시각적 외형 정보를 주로 담당하는 반면, 깊이 브랜치는 객체와 배경 간의 구조적 기하 관계를 표현한다. 두 브랜치의 공동 모델링을 위해 결합 교차 어텐션 방법 [16]을 도입하였다. RGB 브랜치는 깊이 브랜치로부터 공간적 단서를 학습하며, 깊이 브랜치는 RGB 브랜치를 참조하여 구조적 일관성을 유지한다. 이러한 양방향 상호작용을 통해 InDepthVOI는 디노이징 과정에서 외형과 기하 정보를 동시에 정교화할 수 있으며, 깊이 모호성 문제를 완화하고 객체와 장면 간의 자연스러운 통합을 향상시킨다.

또한 InDepthVOI는 장면 구조와 일관된 객체 움직임 제어를 위해 깊이를 인지하는 궤적 맵을 활용한다. 이 표현 방식은 객체의 시공간 궤적을 구조적으로 인코딩함으로써, 단순한 방향 정보 기반의 제어를 넘어 장면 대비 상대적 거리 변화까지 반영하는 모션 컨트롤을 가능하게 한다. 궤적 조건은 ControlNet [26] 기반 보조 인코더(encoder)를 통해 RGB 브랜치에 주입되며, 이를 통해 모션 조건이 디퓨전 과정 전체에 걸쳐 안정적으로 전달될 수 있도록 한다. 다음 단락부터 각 구성 요소에 대한 상세한 설명이 이어진다.

3.1 깊이 궤적 맵

객체 움직임의 제어 가능성을 향상시키기 위해, 본 연구에서는 깊이 궤적 맵을 추가적인 조건으로 활용한다. 기존 움직임 기반 VOI 방법들 [10, 12]은 객체의 궤적을 (x, y) 평면상의 2D 트래킹 맵(tracking map) 형태로 표현한다. 그러나 이러한 표현 방식은 z 축 방향의 움직임을 명시적으로 모델링하지 못한다는 한계를 가진다.

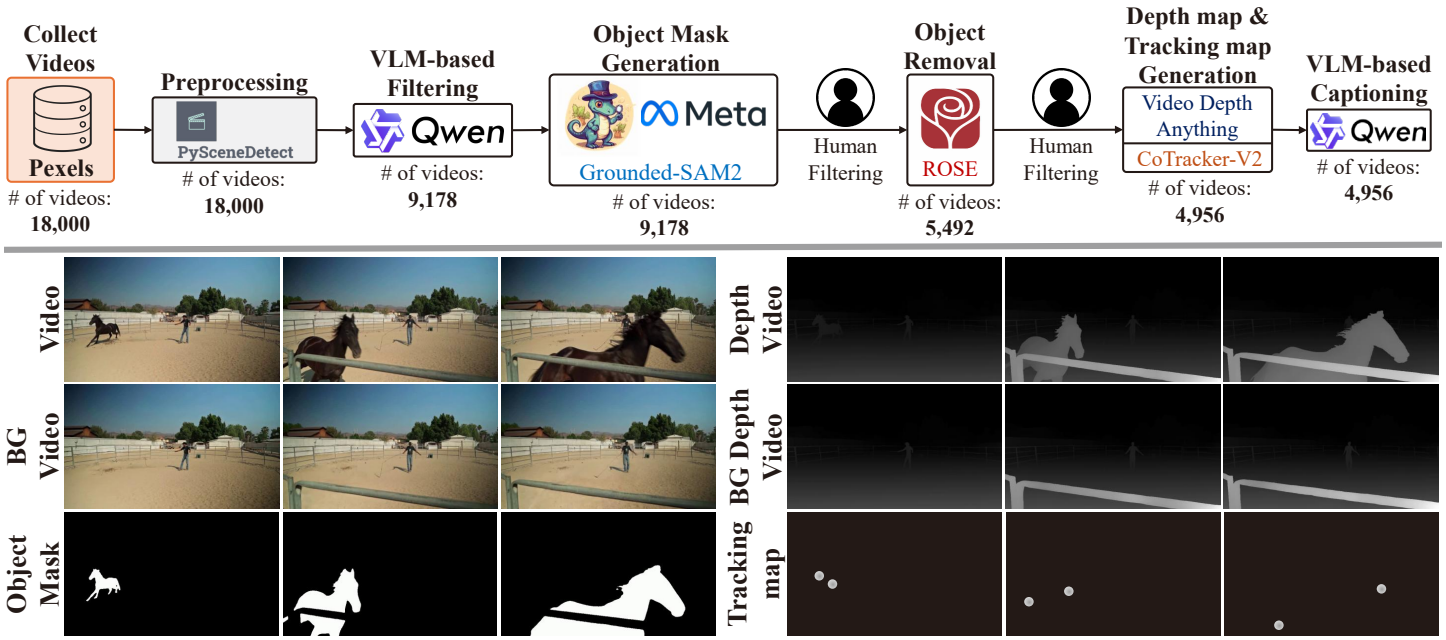

Figure 3: The detailed pipeline of RealVOI dataset. Representative examples from RealVOI dataset are shown below.

이러한 한계에 착안하여, 본 연구에서는 각 (x, y) 트래킹 좌표에 상대적 깊이 값을 함께 할당한다. 먼저 객체 영역 위의 균등 격자(uniform grid)로부터 N개의 점을 무작위로 샘플링한다. 그 다음, 공개된 사전 학습 모델 [27]을 사용하여 포인트 트래킹(point tracking)을 수행하고, 추적된 점들을 비디오와 동일한 공간 차원을 가지는 궤적 맵에 투영한다. 마지막으로 각 트래킹 포인트에 깊이 값 $V_{\text{depth}} \in [-1, 1]$을 할당하여 깊이를 인지하는 궤적 맵 $M_{\text{track}} \in \mathbb{R}^{F\times H\times W}$를 구성한다.

구체적으로, 각 트래킹 포인트의 깊이 값은 깊이 비디오로부터 추출한 객체 영역의 프레임별 평균 깊이 값으로 설정한다. M_{track}은 다음과 같이 정의된다:

$$M_{track}[f, y_f^i, x_f^i] = V_f^{depth} \tag{1}$$

여기서 f는 프레임 인덱스, (x_f^i, y_f^i)는 f번째 프레임에서의 i번째 트래킹 포인트를 의미하며, V_f^{depth}는 f번째 프레임에서 객체 영역의 평균 깊이값을 나타낸다.

이러한 표현 방식은 객체의 (x, y) 평면상 위치 정보와 함께 배경 깊이 비디오 대비 상대적인 깊이 정보까지 동시에 제공한다. 궤적 맵 M_{track}은 ControlNet [26]을 통해 잠재 벡터 표현 F^{motion}으로 인코딩되며, 이후 다음과 같이 RGB 브랜치에 주입된다:

$$F_l = \text{CrossAttn}(F_l^{joint}) + F_l^{motion} \tag{2}$$

여기서 F_l^{joint}은 l번째 결합 교차 어텐션 레이어에서 나온 특징 벡터를 의미한다. 결과적으로 M_{track}은 복잡하고 계산 비용이 큰 3D 궤적 맵을 대체할 수 있는 경량화된 표현 방식으로 동작하며, 객체 움직임에 대한 사용자 제어 성능을 향상시킨다.

3.2 RealVOI

실제 환경에서의 동적 객체 삽입을 지원하기 위해, 본 연구에서는 RealVOI 데이터셋을 제안한다. RealVOI 데이터셋은 두 가지 주요 목표를 바탕으로 구축되었다. 첫째, 각 비디오는 VOI 훈련에 필요한 모든 구성 요소를 포함해야 한다. 여기에는 객체가 존재하는 비디오, 객체가 제거된 깨끗한 배경 비디오, 객체 이진(binary) 마스크, 그리고 객체 참조 이미지(예: 객체가 포함된 첫 번째 프레임)가 포함된다. 둘째, 데이터는 실제 환경으로부터 수집되어야 하며, 배경 비디오에서 그림자나 반사와 같은 객체 관련 시각적 결함(artifact)을 충분히 제거하여 제공해야 한다. 그러나 기존 공개된 VOI 데이터셋들은 이러한 조건을 부분적으로만 만족한다. Table 1에서 볼 수 있듯이, VPData [11]는 배경 비디오를 제공하지 않으며, Senorita-2M [15]은 객체 관련 시각 효과를 완전히 제거하지 못한다. 또한 ROSE++ [13]는 가상 환경을 기반으로 구축되었으며 주로 정적 객체에 초점을 맞추고 있다.

이러한 두 가지 목표를 바탕으로, 본 연구에서는 Figure 3에 묘사된 체계적인 파이프라인을 통해 데이터셋을 구축하였다. 먼저, Pexels [28]으로부터 동적 객체를 포함하는 실제 환경 비디오를 수집하였다. 이후 PySceneDetect [29]를 사용하여 장면 전환을 검출 및 제거하였으며, 모델 학습 설정에 맞추어 비디오의 해상도와 길이를 조정하였다. 다음으로, 고품질 데이터 확보를 위해 Qwen 기반 VLM [30]을 활용하여 시각적 결함(*e.g.*, 심한 블러(blur) 및 과도한 오클루전)을 포함하는 비디오를 필터링하였다. 그 다음, 공개된 사전 학습 모델들 [31, 32, 33, 25, 27]을 사용하여 객체 제거 배경 비디오, 깊이 맵, 그리고 트래킹 포인트 맵을 생성하였으며, 추가적인 사람 검증을 통해 품질이 낮은 결과를 제거하고, 마지막으로 객체 중심의 비디오 캡션(caption)을 VLM을 통해 생성하였다. 구축된 데이터셋의 예시는 Figure 3에 묘사되어 있다.

Figure 4: Qualitative comparison with existing SOTA models.

Model	ViCLIP-T ↑	Background Consistency ↑	Subject Consistency ↑	Motion Smoothness ↑	Dynamic ↑	Imaging Quality ↑	CLIP-I ↑	DINO-I ↑
AnyV2V	23.094	0.885	0.883	0.979	0.400	0.604	0.697	0.405
ReVideo	24.478	0.905	0.908	0.989	0.360	0.634	0.725	0.512
VACE-Inp	24.545	0.932	0.940	**0.991**	0.440	0.664	0.701	0.554
VACE-VOI	24.621	**0.934**	0.942	0.989	0.480	0.675	0.698	0.563
Pika-Pro	24.103	0.911	0.937	0.989	0.400	0.646	0.726	0.562
InDepthVOI	**24.638**	**0.934**	**0.947**	**0.991**	**0.493**	**0.678**	**0.728**	**0.581**

Table 2: Quantitative comparisons with baseline methods. Here, *VACE-Inp* denotes an inpainting model using pre-trained VACE, and *VACE-VOI* denotes VACE fine-tuned on the RealVOI dataset for the VOI task.

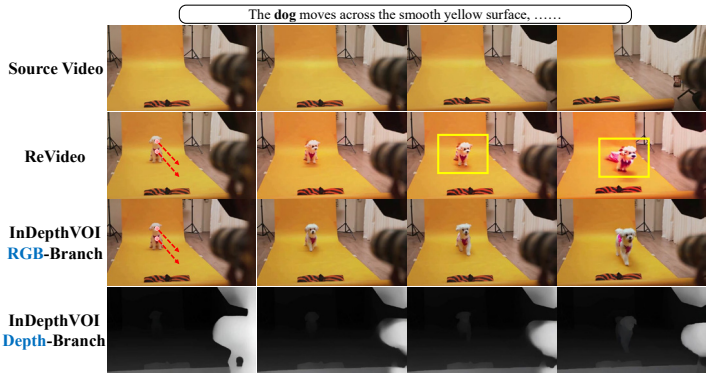

Figure 5: Qualitative ablation of motion editing using the tracking condition. In the Revideo [10] result, the object follows the trajectory but exhibits unnatural motion.

Trajectory Map Condition	ViCLIP-T ↑	Subject Consistency ↑	Dynamic ↑	Imaging Quality ↑	CLIP-I ↑	DINO-I ↑
✗	24.572	0.945	0.480	**0.682**	0.720	0.575
Point Trajectory Map (x,y)	24.581	0.943	0.480	0.674	0.726	0.571
Depth Trajectory Map (x,y,depth) (ours)	**24.638**	**0.947**	**0.493**	0.678	**0.728**	**0.581**

Table 3: Effect of trajectory representation. *Depth Trajectory Map* denotes our proposed trajectory representation incorporating depth information.

Depth Condition	Background Consistency ↑	Subject Consistency ↑	Dynamic ↑	Imaging Quality ↑	CLIP-I ↑	DINO-I ↑
✗	0.934	0.942	**0.480**	0.675	0.698	0.563
Controlnet	0.918	0.880	0.333	0.558	**0.720**	0.393
Joint Cross Attention (ours)	**0.935**	**0.945**	**0.480**	**0.682**	**0.720**	**0.575**

Table 4: Effect of joint cross attention on depth condition learning. *Joint Cross Attention* is our proposed method, which trains without tracking condition.

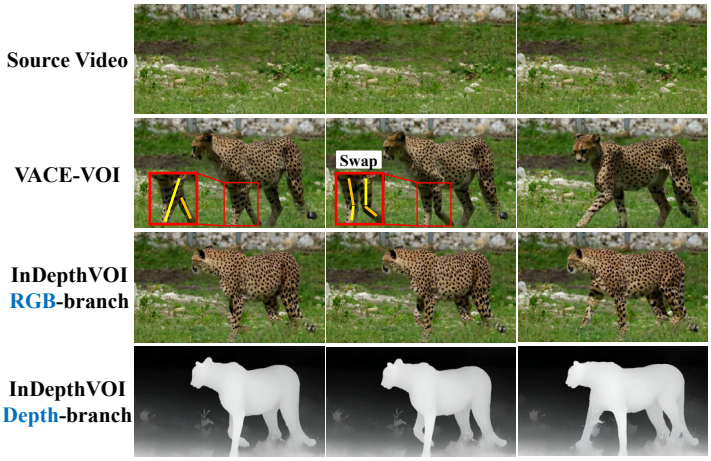

Figure 6: Qualitative ablation about the effect of depth condition. Here, the colored straight lines represent the joints of each body. In the VACE-VOI result, when an animal moves, the front and hind legs swap positions.

4 실험

4.1 구현 세부사항

훈련 및 추론. 본 연구의 실험에서 Wan2.1-VACE-1.3B [7]를 베이스라인(baseline)으로 사용하였다. 모델은 AdamW 옵티마이저(optimizer) [34]를 사용하여 학습률은 1e-4로 3 에폭(epoch) 동안 학습하였으며, 배치 사이즈는 8로 설정하였다. 결합 교차 어텐션(joint cross-attention)의 가중치 파라미터(parameter)는 베이스라인 모델의 셀프 어텐션(self-attention)의 가중치를 기반으로 초기화하였으며, ControlNet은 베이스라인의 Context Block의 가중치를 이용하여 초기화하였다. InDepthVOI에서 랭크(rank) 128의 LoRA [35]를 RGB 및 깊이 브랜치, 결합 교차 어텐션, ControlNet에 적용하였다. 깊이 궤적 맵에서 트래킹 포인트 개수 N은 1에서 10 사이로 설정하였다. 추가적으로, RGB 브랜치는 객체 중심의 프롬프트(prompt)가 들어가지만, 깊이 브랜치에서는 모달리티를 명시적으로 나타내기 위해 프롬프트 앞에 *Depth_map.*과 같은 트리거(trigger) 단어를 추가하여 활용하였다. 추론 시 classifier-free guidance [36] 스케일 값은 5.0, 샘플링 단계(sampling step)는 50으로 설정하였다. 생성된 비디오는 모두 480 × 832 해상도의 81 프레임으로 구성된다.

평가. 평가를 위해 RealVOI 데이터셋에서 다양한 객체 카테고리를 포함하는 75개의 비디오를 무작위로 샘플링하였다. 본 연구에서는 (1) 객체 일관성 평가(Subject Consistency; CLIP-I [37] and DINO-I [38]), (2) VBench [39]를 활용한 비디오 품질 평가(Video Quality ; Background/Subject Consistency, Dynamic, and Imaging Quality), (3) 텍스트-비디오 정합성(text-video alignment; ViCLIP-T [40]) 측면에서 성능을 평가하였다.

4.2 기존 연구와의 비교

본 연구에서는 AnyV2V [8], ReVideo [10], VACE-Inpainting (VACE-Inp) [7], RealVOI로 파인튜닝(fine-tuning)한 VACE-VOI, 그리고 Pika-Pro-2.5 [41]와 성능을 비교하였다. Figure 4와 Table 2에서 확인할 수 있듯이, InDepthVOI는 객체 일관성, 모션 퀄리티, 그리고 텍스트-비디오 정렬 측면에서 가장 균형 잡힌 성능을 달성하였다. 기존 방법들은 삽입 객체와 배경 간의 구조적 관계를 명시적으로 모델링하지 않아 객체-장면 정합성이 저하되는 경향이 있다. 예를 들어, VACE-VOI는 객체 외형과 배경 일관성은 비교적 잘 유지하지만, 구조적 관계를 고려하지 않아 객체와 장면 간 상호작용에서 시각적 결함과 관통(penetration) 문제가 발생한다. 반면 Pika-Pro는 첫 프레임 객체 외형을 충분히 반영하지 못하고 새로운 객체를 생성하는 경향이 있으며, 원본 비디오 대비 과도한 노출(overexposure)로 인해 배경 일관성이 저하된다.

반면 InDepthVOI는 깊이를 인지하는 결합 모델링을 통해 객체와 장면 간의 구조적 정합성을 향상시키며, 깊이 궤적 맵을 추가로 활용하여 움직임 제어 성능 또한 개선한다. 결과적으로, 객체 외형과 배경 일관성을 안정적으로 유지하면서도 보다 자연스럽고 현실적인 객체-장면 상호작용을 달성할 수 있다.

4.3 분석 및 구성 요소 검증 실험

깊이 궤적 맵. 궤적 맵 조건이 객체 움직임 편집에 미치는 영향을 분석하기 위해 추가적인 정성적 실험을 수행하였다. Figure 5에서 볼 수 있듯이, 첫 번째 프레임의 객체에 궤적 조건을 적용하면 제공된 궤적 및 텍스트 묘사에 부합하는 움직임이 생성된다. 기존 궤적 기반 방법들 [12, 10]은 2D 평면상의 정밀한 포인트 레벨(point-level) 궤적 맵에 의존하는 반면, 본 연구의 방법은 단순한 궤적만으로도 자연스럽고 일관된 객체 움직임을 생성할 수 있다. 또한 제안한 깊이 궤적 맵을 통해 객체는 공간 평면뿐 아니라 깊이 방향에서도 일관된 움직임을 보이며, 기하학적으로 자연스러운 움직임 컨트롤이 가능하다. 추가적으로 깊이 궤적 맵의 효과를 분석하기 위해 다양한 궤적 조건들을 비교하였다. Table 3에서 볼 수 있듯이, 2D 좌표와 깊이 정보를 함께 활용한 깊이 궤적 맵은 가장 높은 ViCLIP-T 성능을 달성하였다. 또한, 궤적 조건을 사용하지 않거나 2D 좌표만 활용하는 경우에 비해 Subject Consistency와 Dynamic 지표가 향상되었다. 이는 깊이 정보가 모션 제어에 추가적인 기하학적 정보를 제공하여 텍스트 정합성과 객체 및 움직임의 일관성을 높이는 데 효과적임을 보여준다.

결합 교차 어텐션. 결합 교차 어텐션의 효과를 분석하기 위해, 깊이 정보를 학습 과정에 적용하는 다양한 방법을 비교하였다. 구체적으로, (1) VACE-VOI와 (2) 배경 깊이 비디오를 ControlNet [26] 기반 조건으로 주입하는 VACE-VOI를 비교하였다. Table 4에서 볼 수 있듯이, 제안한 방법은 모든 평가 지표에서 가장 우수한 성능을 달성하였다. 이는 RGB와 깊이 비디오를 공동 생성하는 듀얼브랜치 구조와 결합 교차 어텐션이 VOI 작업에서 효과적임을 보여준다. 반면, ControlNet 기반 방법은 배경 깊이 정보만 활용하고 삽입된 객체와 장면의 기하학적인 관계를 명시적으로 모델링하지 못하기 때문에 상대적으로 낮은 성능을 보인다.

또한 깊이 정보를 활용하지 않는 경우 깊이 모호성 문제가 발생한다. Figure 6에서 볼 수 있듯이, VACE-VOI는 움직임 과정에서 상대적인 깊이 관계가 시간적으로 불안정해지는 현상을 보인다. 예를 들어, 걷는 동작 중 다리의 앞뒤 관계가 순간적으로 뒤바뀌는 문제가 발생한다. 반면, 본 연구의 방법은 삽입 객체의 깊이 정보를 내부적으로 학습함으로써 이러한 깊이 모호성을 완화하며, 보다 안정적인 깊이 관계와 자연스러운 객체 움직임을 생성한다.

데이터셋 분석. RealVOI 데이터셋은 체계적인 파이프라인을 통해 고품질 현실세계 VOI 데이터를 수집할 수 있었으나, 객체 제거의 결과가 항상 완벽한 배경을 제공하지 않으며 트래킹 오류로 인해 예측된 궤적이 실제 움직임과 완벽하게 일치하지 않는 경우가 있다. 그러나 제안된 데이터셋은 실제 환경에서 수집된 다양한 장면을 포함하고 있어 높은 실용성을 가지며, 사용자 입력 궤적의 정밀도나 품질에 보다 강건한 모델 학습을 가능하게 한다.

5 결론

본 연구에서는 생성 과정에 기하학적 추론을 명시적으로 도입한 깊이를 인지하는 비디오 객체 삽입 프레임워크 InDepthVOI를 제안하였다. 제안한 방법은 결합 교차 어텐션 기반 듀얼브랜치 디퓨전 트랜스포머를 통해 외형과 깊이를 공동 모델링함으로써, 깊이 일관성을 유지하는 객체 삽입과 장면과의 구조적 정합성을 향상시킨다. 또한 이미지 평면과 깊이 차원에서의 객체 시공간 궤도를 함께 표현하는 깊이 궤적 맵을 제안하여, 세밀하고 제어 가능한 객체 움직임 생성을 가능하게 하였다.

추가적으로, 실제 환경 비디오 기반의 고품질 배경-합성 비디오 쌍을 제공하는 RealVOI 데이터셋을 구축하였다. 제안한 데이터셋은 객체 제거, VLM 기반 자동 필터링, 그리고 사람 검증 과정을 결합하여 기존 데이터셋 대비 더욱 깨끗하고 현실적인 학습 데이터를 제공한다. 다양한 실험 결과를 통해 InDepthVOI는 높은 모션 퀄리티를 유지하면서도 일관성 및 장면-객체 상호작용 측면에서 우수한 성능을 달성함을 확인하였다. 이는 현실적이고 제어 가능한 비디오 객체 삽입을 위해 명시적인 기하적 모델링이 효과적임을 보여준다. 나아가 본 연구는 현실적인 객체-장면 상호작용이 중요한 비디오 편집 분야로의 확장 가능성을 제시한다.

한계. 제안한 방법은 듀얼브랜치 디퓨전 트랜스포머 구조로 인해 높은 메모리 사용량이 요구된다. 게다가 복잡하고 빠른 움직임을 가지는 객체에 대해서는 여전히 한계가 존재한다. 이는 RealVOI 데이터셋에 매우 동적인 비디오 데이터가 상대적으로 적어 모델이 복잡한 객체 움직임에 대해 충분히 학습하기 어렵고, 훈련에 사용된 베이스라인 모델의 제한적인 움직임 표현 능력에도 일부 기인한다. 향후 연구에서는 메모리 효율성을 개선하고, 더욱 다양한 동적인 데이터를 수집하거나 객체의 움직임 정보를 모델 내부에 효과적으로 반영할 수 있는 구조를 추가하여 복잡한 객체 움직임에 대한 성능을 향상시킬 예정이다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(IITP) (RS-2026-25511821, 개인화 미디어 서비스 추천 · 자동생성 기술 개발; RS-2024-00395401, 생성 AI 기반 특수효과 자동 생성 및 합성 기술 개발; RS-2019-II191906, 인공지능대학원지원(포항공과대학교))의 지원을 받아 수행된 연구임.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[3] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[4] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[5] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[6] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[7] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu, "Vace: All-in-one video creation and editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 17 191–17 202.

[8] M. Ku, C. Wei, W. Ren, H. Yang, and W. Chen, "Anyv2v: A tuning-free framework for any video-to-video editing tasks," *Transactions on Machine Learning Research*.

[9] G. Li, Y. Yang, C. Song, and C. Zhang, "Flowdirector: Training-free flow steering for precise text-to-video editing," *arXiv preprint arXiv:2506.05046*, 2025.

[10] C. Mou, M. Cao, X. Wang, Z. Zhang, Y. Shan, and J. Zhang, "Revideo: Remake a video with motion and content control," *Advances in Neural Information Processing Systems*, vol. 37, pp. 18 481–18 505, 2024.

[11] Y. Bian, Z. Zhang, X. Ju, M. Cao, L. Xie, Y. Shan, and Q. Xu, "Videopainter: Any-length video inpainting and editing with plug-and-play context control," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025, pp. 1–12.

[12] Y. Tu, H. Luo, X. Chen, S. Ji, X. Bai, and H. Zhao, "Videoanydoor: High-fidelity video object insertion with precise motion control," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025, pp. 1–11.

[13] H. Jin, H. Jang, J. Kim, J. Hyung, K. Kim, D. Kim, H. Choi, H. Kim, and J. Choo, "Insertanywhere: Bridging 4d scene geometry and diffusion models for realistic video object insertion," *arXiv preprint arXiv:2512.17504*, 2025.

[14] J. Chen, X. Li, X. Bai, T. Ma, P. Zhang, Z. Chen, G. Li, L. Liu, S. Zhao, B. Li, *et al.*, "Omniinsert: Mask-free video insertion of any reference via diffusion transformer models," *arXiv preprint arXiv:2509.17627*, 2025.

[15] B. Zi, P. Ruan, M. Chen, X. Qi, S. Hao, S. Zhao, Y. Huang, B. Liang, R. Xiao, and K.-F. Wong, "Se\~norita-2m: A high-quality instruction-based dataset for general video editing by video specialists," *arXiv preprint arXiv:2502.06734*, 2025.

[16] X. Li, C. Herrmann, K. C. Chan, Y. Li, D. Sun, C. Ma, and M.-H. Yang, "A simple approach to unifying diffusion-based conditional generation," *arXiv preprint arXiv:2410.11439*, 2024.

[17] Z. Xiao, L. Liu, Y. Gao, X. Zhang, H. Che, S. Mai, and Q. Tian, "Lovora: Text-guided and mask-free video object removal and addition with learnable object-aware localization," *arXiv preprint arXiv:2512.02933*, 2025.

[18] J. Zhang, S. Li, Y. Lu, T. Fang, D. McKinnon, Y. Tsin, L. Quan, and Y. Yao, "Jointnet: Extending text-to-image diffusion for dense distribution modeling," *arXiv preprint arXiv:2310.06347*, 2023.

[19] K. Byung-Ki, Q. Dai, L. Hyoseok, C. Luo, and T.-H. Oh, "Jointdit: Enhancing rgb-depth joint modeling with diffusion transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 25 261–25 271.

[20] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, "Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 219–10 228.

[21] K. Liu, W. Li, L. Chen, S. Wu, Y. Zheng, J. Ji, F. Zhou, R. Jiang, J. Luo, H. Fei, *et al.*, "Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization," *arXiv preprint arXiv:2503.23377*, 2025.

[22] Y. Zhai, K. Lin, L. Li, C.-C. Lin, J. Wang, Z. Yang, D. Doermann, J. Yuan, Z. Liu, and L. Wang, "Idol: Unified dual-modal latent diffusion for human-centric joint video-depth generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 134–152.

[23] Google. (2026) Nano banana 2 - gemini ai image generator and photo editor. [Online]. Available: https://gemini.google/overview/image-generation/

[24] S. Liu, Y. Han, P. Xing, F. Yin, R. Wang, W. Cheng, J. Liao, Y. Wang, H. Fu, C. Han, *et al.*, "Step1x-edit: A practical framework for general image editing," *arXiv preprint arXiv:2504.17761*, 2025.

[25] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang, "Video depth anything: Consistent depth estimation for super-long videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 831–22 840.

[26] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[27] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker: It is better to track together," in *European conference on computer vision*. Springer, 2024, pp. 18–35.

[28] Pexels. (2026) Pexels. [Online]. Available: https://www.pexels.com

[29] PysceneDetect. (2026) Pyscenedetect. [Online]. Available: https://www.scenedetect.com/

[30] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[31] T. Ren, Q. Jiang, S. Liu, Z. Zeng, W. Liu, H. Gao, H. Huang, Z. Ma, X. Jiang, Y. Chen, *et al.*, "Grounding dino 1.5: Advance the" edge" of open-set object detection," *arXiv preprint arXiv:2405.10300*, 2024.

[32] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[33] C. Miao, Y. Feng, J. Zeng, Z. Gao, H. Liu, Y. Yan, D. Qi, X. Chen, B. Wang, and H. Zhao, "Rose: Remove objects with side effects in videos," *arXiv preprint arXiv:2508.18633*, 2025.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[36] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[38] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[39] Z. Huang, F. Zhang, X. Xu, Y. He, J. Yu, Z. Dong, Q. Ma, N. Chanpaisit, C. Si, Y. Jiang, *et al.*, "Vbench++: Comprehensive and versatile benchmark suite for video generative models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[40] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, *et al.*, "Internvideo: General video foundation models via generative and discriminative learning," *arXiv preprint arXiv:2212.03191*, 2022.

[41] Pika-Pro. (2026) Pika-pro. [Online]. Available: https://pika.art/

〈저자소개〉

한규진

- 2024년 8월 한동대학교 ICT창업학부/수리통계 학사
- 2025년 2월 ~ 현재 포항공과대학교 인공지능대학원 석사과정
- https://orcid.org/0000-0002-9504-9768

강경국

- 2017년 2월 한양대학교 융합전자공학부 학사
- 2019년 2월 대구경북과학기술원 정보통신융합전공 석사
- 2026년 2월 포항공과대학교 컴퓨터공학과 박사
- 2026년 3월 ~ 현재 삼성전자 AI센터 근무
- https://orcid.org/0000-0002-8964-1220

조성현

- 2005년 8월 포항공과대학교 컴퓨터공학 학사
- 2012년 2월 포항공과대학교 컴퓨터공학 박사
- 2012년 3월 ~ 2014년 3월 Adobe Research 연구원
- 2014년 4월 ~ 2017년 4월 삼성전자 책임연구원
- 2017년 4월 ~ 2019년 8월 대구경북과학기술원 조교수
- 2019년 8월 ~ 2021년 8월 포항공과대학교 조교수
- 2021년 9월 ~ 현재 포항공과대학교 부교수
- https://orcid.org/0000-0001-7627-3513