Check for updates
특집논문 (Special Paper)
방송공학회논문지 제31권 제5호, 2026년 9월 (JBE Vol.31, No.5, September 2026)
https://doi.org/10.5909/JBE.2026.31.5.985
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

확산 모델 기반의 제어 가능한 텍스트-비디오 생성 기술 분류를 위한 에이전트 기반 지능형 문헌 분석 프레임워크

김 민 진[a)], 윤 소 율[a)], 박 지 예[b)], 유 제 혁[a)‡]

An Agent-Based Intelligent Literature Analysis Framework for the Taxonomy of Controllable Diffusion-Based Text-to-Video Generation

Minjin Kim[a)], Soyul Youn[a)], Jiye Park[b)], and Jehyeok Rew[a)‡]

요 약

최근 대규모 언어모델 기반 AI 에이전트는 방대한 기술 문헌을 자동으로 분석하고 다양한 목적으로 활용하는 방법론이 주목받고 있다. 본 연구는 확산 모델 기반의 제어 가능한 텍스트-비디오 생성 분야의 최근 기술적 흐름을 파악하기 위해, 문헌 선별, 정보 추출, 연구 방향 도출, 기술 분류 및 결과 통합을 수행하는 에이전트 기반 지능형 문헌 분석 프레임워크를 제안하였다. 제안 프레임워크는 단계별 분석 결과를 원문 근거와 연결하고 검증된 정보만 후속 분석에 활용함으로써 분석 과정의 추적성과 검토 가능성을 확보할 수 있다. OpenAlex에서 수집한 1,308편 중 핵심적 기술변화를 나타내는 93편을 선정하고, 대표 연구 문제와 기술 정보를 Research Axis와 기능 중심 Technology Taxonomy로 구조화하였다. 분석 결과, 명시적 움직임 및 궤적 제어, 구조적 시간 및 조건 전달 제어, 카메라·기하 및 3차원 시점 제어가 주요 연구 방향으로 나타났으며, 동일한 기술이 여러 연구 방향에서 다양한 기능으로 활용되고 있음을 확인하였다. 이를 통해 제안 프레임워크가 대규모 기술 문헌에서 분석 목적에 필요한 핵심 정보를 원문 근거에 기반해 선별하며, 이를 일관된 기준으로 구조화하여 검토 가능한 분석 결과를 제공할 수 있음을 확인하였다.

Abstract

Recent advances in large language model‐based AI agents have attracted growing attention as a methodology for automatically analyzing large volumes of technical literature and supporting diverse analytical purposes. This study proposes an agent-based intelligent literature analysis framework to identify recent technological trends in controllable diffusion-based text-to-video generation. The framework performs literature screening, information extraction, research-axis derivation, technology classification, and result integration. By linking the outputs of each analytical stage to evidence from the source text and using only verified information in subsequent stages, the framework enables traceable and reviewable analysis. From 1,308 records collected through OpenAlex, 93 papers representing key technological developments were selected, and their representative research problems and technical information were structured into Research Axes and a function-oriented Technology Taxonomy. The analysis identified explicit motion and trajectory control, structured temporal and condition-routing control, and camera, geometry, and 3D viewpoint control as major research directions. It also showed that the same technologies are utilized for different functions across multiple research directions. These results demonstrate that the proposed framework can identify relevant information from large-scale technical literature based on source evidence and systematically structure it according to consistent criteria to provide reviewable analytical results.

Keyword : Controllable Text-to-Video Generation, Diffusion Model, Agent-based Literature Analysis, Research Axis, Technology Taxonomy

Copyright Ⓒ 2026 Korean Institute of Broadcast and Media Engineers. All rights reserved.
"This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered."

Ⅰ. 서 론

영상 콘텐츠 제작에서는 기획된 내용을 시각적으로 구현하기 위해 장면 구성, 카메라 연출, 객체의 움직임 및 시각적 스타일과 같은 요소를 설계하며, 이러한 요소는 제작자의 의도에 따라 일관되게 구성될 필요가 있다[1]. 최근 생성형 AI의 발전으로 텍스트를 조건으로 비디오를 생성하는 텍스트 기반 비디오 생성(text-to-video generation, T2V)이 등장하면서, 생성형 AI의 활용 범위는 기존 제작 과정을 보조하는 수준에서 영상 생성과 편집 과정의 일부를 수행하는 방향으로 확대되고 있다[2]. 초기 T2V 모델은 고해상도 영상과 프레임 간 시간적 일관성을 확보하는 데 어려움이 있었으나[3], 확산 모델(diffusion model)의 도입 이후 T2V 기술은 생성 품질과 텍스트 조건에 대한 의미적 정합성 측면에서 발전하였다[4].

그러나 생성 품질의 향상만으로 영상 콘텐츠 제작에 필요한 연출 요소를 충분히 제어하기는 어렵다. 실제 영상 콘텐츠 제작에서는 사실적인 영상을 생성하는 것뿐만 아니라, 카메라 움직임, 인물의 동선, 장면 전환, 객체의 위치 및 시각적 스타일을 제작자의 의도에 맞게 조정할 수 있어야 한다. 자연어 프롬프트를 주된 조건으로 사용하는 기존 T2V 모델은 객체의 정확한 위치와 이동 경로, 카메라 움직임, 포즈 및 장면 구조를 세밀하게 지정하는 데 한계가 있다. 또한 여러 프레임에 걸쳐 객체의 외형과 움직임을 일관되게 유지하면서 제작자의 연출 의도를 안정적으로 반영하기 어렵다[5].

이러한 한계를 해결하기 위해 최근 확산 모델을 이용한 T2V 연구는 생성 과정의 제어 가능성을 강화하는 방향으로 발전하고 있다. 이에 따라 궤적, 포즈, 깊이, 카메라 파라미터 및 참조 이미지 등 다양한 조건을 활용하여 제작자의 연출 의도를 생성 과정에 반영하는 확산 모델을 이용한 제어 가능한 T2V 기법이 활발히 제안되고 있다. 특히 제어 대상, 조건 표현 방식, 모델 구조 및 제어 메커니즘이 다양해지면서 제어 가능성은 T2V 분야의 주요 연구 주제로 자리 잡고 있다[6]. 따라서 T2V 연구자와 영상 콘텐츠 제작자는 서로 다른 제어 기술의 기능적 차이와 상호 관계, 그리고 분야의 발전 흐름을 일관된 기준으로 파악할 필요가 있다. 기존 조사 및 분류 연구는 비디오 확산 모델 연구를 생성 과업, 입력 조건, 모델 구조 또는 제어 대상 중심으로 분류해 왔다[7]. 그러나 이러한 분류는 동일한 기술이 서로 다른 연구 문제에서 상이한 기능을 수행하거나 여러 연구 방향에서 공유되는 다대다 관계를 충분히 표현하기 어렵다.

이러한 한계를 바탕으로 본 연구는 다음과 같은 연구 질문을 설정하였다.

RQ1. 제어 가능한 확산 기반 T2V 연구는 어떠한 주요 연구 방향으로 구조화될 수 있는가?

RQ2. 각 연구 방향에서 사용되는 기술은 어떠한 기능적 역할에 따라 분류될 수 있는가?

RQ3. 연구 방향과 기술의 기능적 역할 사이에는 어떠한 다대다 관계가 존재하는가?

RQ4. 제어 가능한 확산 기반 T2V의 주요 연구 방향은 시간에 따라 어떠한 변화 양상을 보이는가?

이러한 연구 질문에 답하기 위해 본 연구는 확산 모델을 이용한 제어 가능한 T2V 분야의 연구 문제와 핵심 기술 구조를 체계적으로 분석하기 위한 에이전트 기반 지능형 문헌 분석 파이프라인을 제안한다. 제안 파이프라인은 문헌 검색 및 선별, 연구 문제와 기술 정보 추출, 연구 방향 도출, 기술 분류 및 통합 분석을 역할별 에이전트가 단계적으로 수행하도록 구성하였다. 또한 각 단계의 분석 결과를 원문 근거와 연결하여 정보의 출처와 판단 과정을 추적하고 검토할 수 있도록 하였다. 이를 통해 지속적으로 축적되는 기술 문헌을 일관된 기준으로 처리하고, 개별 논문에 분산된 연구 문제와 기술 정보를 구조화하였다.

본 연구의 기여는 다음과 같다. 첫째, 문헌 검색과 선별부터 정보 추출, 연구 방향 도출, 기술 분류 및 결과 통합까지 전 과정을 수행하는 네 개의 역할별 에이전트로 구성된 지능형 문헌 분석 파이프라인을 제시하였다. 각 단계의 산출

a) 덕성여자대학교 데이터사이언스학과(Department of Data Science, Duksung Women's University)
b) 덕성여자대학교 응용데이터사이언스학과(Department of Applied Data Science, Duksung Women's University)
‡ Corresponding Author : 유제혁(Jehyeok Rew)
E-mail: jhrew@duksung.ac.kr
Tel: +82-2-901-8757
ORCID: https://orcid.org/0000-0003-3952-7582
※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-00516023).
· Manuscript received August 4, 2026; Revised August 25, 2026; Accepted August 26, 2026.

물을 원문 근거와 연결함으로써 분석 결과의 추적성과 검토 가능성을 확보하였다. 둘째, 제어 조건이나 기술명 중심의 기존 분류를 넘어, 기술이 개별 연구에서 수행하는 기능적 역할을 기준으로 하는 기술 분류체계(Technology Taxonomy)를 정의하였다. 이를 통해 서로 다른 명칭으로 제시된 제어 기술을 공통된 기능적 기준에서 비교하고, 동일 기술의 다기능적 활용을 분석할 수 있도록 하였다. 셋째, 연구 문제에서 도출한 연구 방향(Research Axis)과 기술 분류체계(Technology Taxonomy)를 논문 단위로 연결하여 연구 문제와 기술 구조 간의 관계를 분석하였다. 이를 통해 각 연구 방향에서 활용되는 핵심 기술의 분포와 공유 구조를 파악하고, 확산 모델을 이용한 제어 가능한 T2V 분야의 기술 발전 흐름을 해석하는 방법을 제시하였다.

Ⅱ. 관련 문헌 연구

1. 생성형 AI를 활용한 비디오 생성 기술

비디오 생성은 정지 영상 생성과 달리 개별 프레임의 시각적 품질뿐 아니라 객체 움직임의 연속성, 프레임 간 시간적 일관성, 장면의 공간적 일관성을 함께 모델링해야 하는 과제이다. 생성된 객체는 시간의 흐름에 따라 일관된 외형과 정체성을 유지하면서 자연스럽게 움직여야 하며, 카메라 시점이나 장면 구조가 변하는 경우에도 프레임 간 공간 관계가 안정적으로 유지되어야 한다[8]. 따라서 비디오 생성은 공간적 표현과 시간적 동역학을 동시에 학습해야 하는 시공간 생성 문제로 볼 수 있으며, 이미지 생성보다 복합적인 표현 학습이 요구된다[2,8]. 초기 비디오 생성 연구에서는 생성적 적대 신경망(Generative Adversarial Network, GAN), 변이형 오토인코더(Variational Autoencoder, VAE) 및 자기 회귀 모델(Autoregressive Model) 등을 활용한 접근이 제안되었다[2,7]. 이러한 접근은 데이터로부터 영상의 공간적 특징과 시간적 변화를 학습하여 비디오 생성의 가능성을 보였다. 그러나 생성 길이가 증가할수록 객체의 외형과 배경 구조가 불안정해지고, 프레임 간 움직임의 연속성과 장면 일관성이 저하되는 한계가 있었다[9,10]. 또한 고해상도 또는 장시간 비디오를 생성하는 과정에서 계산량과 메모리 사용량이 크게 증가하였다[7,9,10,11].

이후 확산 모델은 데이터에 노이즈를 단계적으로 추가하는 전방 확산 과정을 설정하고, 노이즈가 포함된 데이터로부터 원본 데이터를 복원하는 역방향 잡음 제거 과정을 학습하는 생성 모델로 주목받았다. 확산 모델은 안정적인 학습과 높은 생성 품질을 보였지만, 픽셀 공간에서 반복적인 노이즈 제거 과정을 수행하는 초기 방식은 고해상도 데이터 생성에 높은 계산 비용을 요구하였다. 이미지 생성 분야에서 제안된 잠재 확산 모델(Latent Diffusion Model, LDM)은 이미지를 저차원의 잠재 표현으로 압축하고 잠재 공간에서 확산 과정을 수행함으로써 계산량을 줄이면서 생성 품질을 유지할 수 있는 구조를 제시하였다. 또한 교차 어텐션(cross-attention)을 이용하여 텍스트와 같은 외부 조건을 생성 과정에 반영함으로써 텍스트 조건부 이미지 및 비디오 생성 모델의 기반을 마련하였다[6,12]. 더불어, 이미지 생성에 적용되던 확산 모델은 공간 정보와 시간 정보를 함께 처리할 수 있도록 확장되면서 비디오 생성에 활용되기 시작하였다[2,7]. Ho 등은 이미지 확산 모델의 공간적 생성 구조에 시간축 모델링을 결합하여 비디오 데이터의 시공간 분포를 학습하는 방법을 제안하였다[13]. Singer 등의 Make-A-Video는 텍스트-이미지 데이터로부터 시각적 및 의미적 표현을 학습하고, 텍스트 주석이 없는 비디오 데이터로부터 움직임과 시간적 변화를 학습하는 방식을 제시하였다[14]. Blattmann 등은 사전 학습된 이미지 기반 LDM에 시간적 계층을 추가한 뒤 비디오 데이터로 미세 조정하는 T2V 모델을 제안하였다[15]. 이러한 연구들은 대규모 이미지 생성 모델이 학습한 공간적 표현을 재사용하면서 시간적 모델링 요소를 결합하여 비디오 생성으로 확장하는 기술적 방향을 제시하였다[2,5,7,13].

이처럼 확산 기반 접근은 비디오의 시각적 품질과 프레임 간 일관성, 텍스트와 비디오 간 의미적 정합성을 개선하는 방향으로 발전해 왔다. 그러나 자연어 프롬프트만으로는 객체의 정확한 위치와 이동 경로, 인물의 자세, 카메라 움직임 및 장면의 공간 구조와 같은 세부적인 시공간 정보를 명시적으로 지정하기 어렵다. 이에 따라 궤적, 포즈, 깊이 맵, 참조 이미지, 객체 마스크 및 카메라 파라미터 등을 추가 조건으로 활용하는 확산 모델 기반의 제어 가능한 T2V 연구가 확대되고 있다[4,5,7]. 본 연구에서는 텍스트 프

롬프트를 기본 의미 조건으로 사용하되, 별도의 조건을 추가하여 움직임, 공간 구조, 객체 배치, 카메라 시점 또는 시각적 속성을 제어하는 생성 방법을 확산 모델 기반의 제어 가능한 T2V 기술로 정의한다. 이를 바탕으로 개별 연구가 해결하려는 제어 문제와 이를 구현하기 위해 사용한 핵심 기술의 기능적 역할을 분석한다.

2. 확산 모델 기반의 제어 가능한 T2V

확산 모델 기반의 제어 가능한 T2V는 텍스트 프롬프트와 추가적인 제어 조건을 생성 과정에 반영하여 사용자가 의도한 영상을 생성하는 기술이다. 일반적인 T2V 기술이 자연어를 기반으로 영상의 전반적인 내용과 분위기를 생성하는 데 초점을 둔다면, 제어 가능한 T2V 기술은 추가 조건을 이용하여 객체의 위치와 움직임, 공간 구조, 카메라 시점, 객체의 외형 및 정체성 등을 보다 명시적으로 조절하는 것을 목표로 한다. 이러한 제어 과정은 기본적으로 확산 모델 기반 T2V의 생성 구조에서 이루어진다[5,7]. 일반적인 확산 모델 기반 T2V 모델은 텍스트 프롬프트를 조건으로 초기 노이즈에서 비디오 표현을 점진적으로 복원한다. 특히 LDM 기반 모델은 저차원의 잠재 공간에서 노이즈 제거 과정을 수행한 후 최종 잠재 표현을 비디오 프레임으로 복원하여 계산 효율성과 생성 품질을 함께 고려한다. 이 과정에서 텍스트 프롬프트 이외의 제어 정보를 입력하면 생성 결과의 공간적 및 시간적 속성을 구체적으로 지정할 수 있다[2,7,16].

구체적으로 제어 조건은 제어 대상에 따라 다양한 형태로 표현된다. 포즈와 스켈레톤은 인물의 자세와 신체 구조를 나타내며, 궤적과 모션 필드는 객체의 이동 경로와 움직임을 제어하는 데 활용된다[4]. 또한 깊이 맵, 분할 맵, 레이아웃 및 경계 상자는 장면의 공간 구조와 객체 배치를 표현한다[4,16,17]. 카메라와 관련해서는 카메라 포즈, 카메라 궤적 및 3차원 기하 정보를 사용하여 카메라의 위치, 방향 및 이동 경로를 지정할 수 있다[7]. 이와 함께 참조 이미지(Reference Image)와 정체성 특징(Identity Feature)은 객체의 외형과 정체성을 여러 프레임에서 일관되게 유지하는 데 활용된다[7,18]. 더 나아가 복수의 제어 조건을 동시에 적용하여 움직임, 공간 배치 및 시각적 속성을 함께 조절하는 다중 조건 제어도 연구되고 있다[6,18]. 이와 같은 제어 정보는 조건 인코더(condition encoder)와 교차 어텐션, 어댑터(adapter) 또는 제어 분기(control branch)를 통해 생성 모델에 반영된다. 한편, 모델을 추가로 학습하지 않고 추론 단계에서 어텐션 맵이나 잠재 특징을 조정하는 무학습 제어 방식도 제안되고 있다. 이처럼 제어 조건의 표현 형식뿐 아니라, 해당 조건을 생성 모델에 주입하고 반영하는 구조 역시 연구별로 다양하게 나타난다[6,7,18].

그러나 비디오 생성에서는 제어 조건을 정확히 반영하는 것만으로 충분하지 않다. 각 프레임이 동일한 제어 의도를 따르면서도 객체의 외형, 움직임 및 장면 구조가 시간축에서 일관되게 유지되어야 한다. 이를 위해 시간적 어텐션, 모션 모듈 및 3차원 시공간 표현이 활용된다. 이러한 기술은 개별 프레임에서 제어 정보를 반영하는 동시에, 해당 정보가 연속된 프레임으로 안정적으로 전달되도록 지원한다[2,8,18]. 특히 여러 제어 조건을 동시에 사용하는 경우에는 조건 간 상호작용도 고려해야 한다. 서로 다른 제어 조건이 동일한 공간 영역이나 시간 구간에 영향을 미치면 조건 정보가 중복되거나 상충하여 생성 품질과 제어 정확도가 저하될 수 있다. 이에 따라 조건별 영향력을 조절하거나, 제어 정보의 중요도에 따라 특징을 선택적으로 결합하고, 공간적 및 시간적 적용 범위를 분리하는 전략이 제안되고 있다[6,7,18].

3. 체계적 문헌 분석과 자동화된 문헌 선별

대규모 언어 모델(large language model, LLM)을 활용하여 문헌 선별과 정보 추출을 지원하는 자동화 연구가 진행되고 있다. 이러한 접근은 대규모 문헌을 효율적으로 처리하고 반복적인 검토 작업을 줄이는 데 활용될 수 있다[19,20,21,22]. 그러나 기존 연구와 도구는 주로 문헌 검색, 적합성 판정 또는 정보 추출과 같은 문헌 고찰의 개별 단계를 지원하는 데 초점을 두고 있으며, 연구 문제와 기술 정보의 추출부터 연구 방향 도출, 기술 분류, 결과 통합 및 시각화까지의 과정을 하나의 분석 절차로 연결하는 데에는 한계가 있다.

4. 기술 분류체계의 구성 방법론

기술 분류체계(taxonomy)는 분석 대상이 지닌 공통점과

차이점을 관점과 특성에 따라 구조화함으로써 복잡한 기술 현상을 체계적으로 이해할 수 있도록 하는 분류체계이다. Nickerson 등은 임의적인 분류체계 구성을 방지하기 위해 분류의 중심 기준이 되는 메타특성(meta-characteristic)을 설정하고, 개념적-경험적(conceptual-to-empirical) 접근과 경험적-개념적(empirical-to-conceptual) 접근을 반복하여 관점과 특성을 정련하는 방법을 제안하였다[23]. 이후 Kundisch 등은 기존 방법론의 적용 과정이 일관되지 않거나 충분히 보고되지 않는 문제와 분류체계의 사후평가가 부족한 문제를 지적하고, 분류체계의 설계, 개발, 평가 및 결과 보고를 위한 확장된 절차와 구체적인 설계 권고안을 제시하였다[24].

확산 모델 기반의 제어 가능한 T2V 분야가 빠르게 발전함에 따라 관련 연구와 기술을 체계적으로 정리하기 위한 서베이 연구가 수행되고 있다. 기존 서베이는 주로 포즈, 궤적, 깊이, 카메라 및 참조 이미지와 같은 제어 조건을 기준으로 연구를 분류하거나, 어댑터, 제어 분기 및 어텐션 기반 조건 주입 구조를 중심으로 구현 방식을 정리하였다. 이러한 분류는 제어 대상과 모델의 구현 구조를 이해하는 데 유용한 기반을 제공하였다[4,5,6,7].

그러나 제어 조건이나 모델 구조 중심의 분류만으로는 개별 연구가 해결하려는 연구 문제와 적용된 기술의 기능적 역할 간 관계를 충분히 설명하기 어렵다. 동일한 제어 조건이나 기술 요소라도 연구의 목적과 적용 맥락에 따라 서로 다른 문제를 해결하거나 상이한 기능을 수행할 수 있기 때문이다. 또한 하나의 기술이 여러 연구 방향에서 공통적으로 활용되거나, 하나의 연구 문제를 해결하는 데 여러 기술이 함께 사용될 수도 있다. 따라서 관련 연구를 체계적으로 비교하기 위해서는 기술의 명칭이나 적용 여부뿐 아니라, 해당 기술이 해결하려는 연구 문제와 생성 과정에서 수행하는 기능적 역할을 함께 분석할 필요가 있다.

5. AI 에이전트 기반 문헌 분석

LLM 기반 분석 결과에는 정보의 누락, 잘못된 추론 또는 원문과 일치하지 않는 내용이 포함될 수 있으므로, 추출된 정보를 원문 근거와 연결하여 검토하는 절차가 필요하다[19,20,22,25,26]. 이에 본 연구는 문헌 분석 에이전트(Literature Analysis Agent), 연구 방향 에이전트(Research Axis Agent), 기술 분류 에이전트(Technology Taxonomy Agent) 및 시각화 에이전트(Visualization Agent)의 네 개 역할 특화 에이전트로 구성된 지능형 문헌 분석 프레임워크를 제안한다. 각 에이전트는 문헌 검색 및 선별과 정보 추출, 연구 방향 도출, 기능 중심의 기술 분류, 결과 통합 및 시각화를 순차적으로 수행하며, 선행 에이전트의 출력은 후속 에이전트의 입력으로 전달된다. 또한 각 단계에서 추출된 정보를 해당 논문의 원문 근거와 연결하여 분석 결과의 출처와 판단 과정을 추적하고 검토할 수 있도록 하였다. 제안 프레임워크는 개별 논문의 연구 문제와 기술의 기능적 역할을 구조화하고, 두 결과를 논문별로 연결하여 확산 모델 기반의 제어 가능한 T2V 분야의 연구 방향, 기술 활용 구조 및 연도별 연구 구성의 변화를 분석하는 것을 목적으로 한다.

Ⅲ. 방법론

1. 에이전트 구성 파이프라인

본 연구에서는 확산 모델 기반의 제어 가능한 T2V 분야의 문헌을 체계적으로 분석하기 위해 에이전트 기반 통합 문헌 분석 파이프라인을 설계하였다. 파이프라인은 Literature Analysis Agent, Research Axis Agent, Technology Taxonomy Agent 및 Visualization Agent의 네 개 에이전트로 구성된다. 각 에이전트는 사전에 정의된 역할과 프롬프트에 따라 문헌 검색과 선별, 연구 문제 및 연구 목표, 기여 추출, 연구 방향 도출, 기술 분류 및 결과 시각화 과정을 자동으로 수행하며, 선행 에이전트가 생성한 결과는 후속 에이전트의 입력으로 사용된다. 이러한 순차적인 흐름에 따라 문헌 수집부터 결과 분석 및 시각화까지의 전 과정을 하나의 자동화된 분석 파이프라인으로 구성하였으며, 전체 파이프라인과 이를 구성하는 에이전트 간의 관계는 <그림 1>과 같다.

Literature Analysis Agent는 관련 문헌을 자동으로 검색 및 선별하고 원문 PDF를 분석하여 연구 문제, 연구 목표,

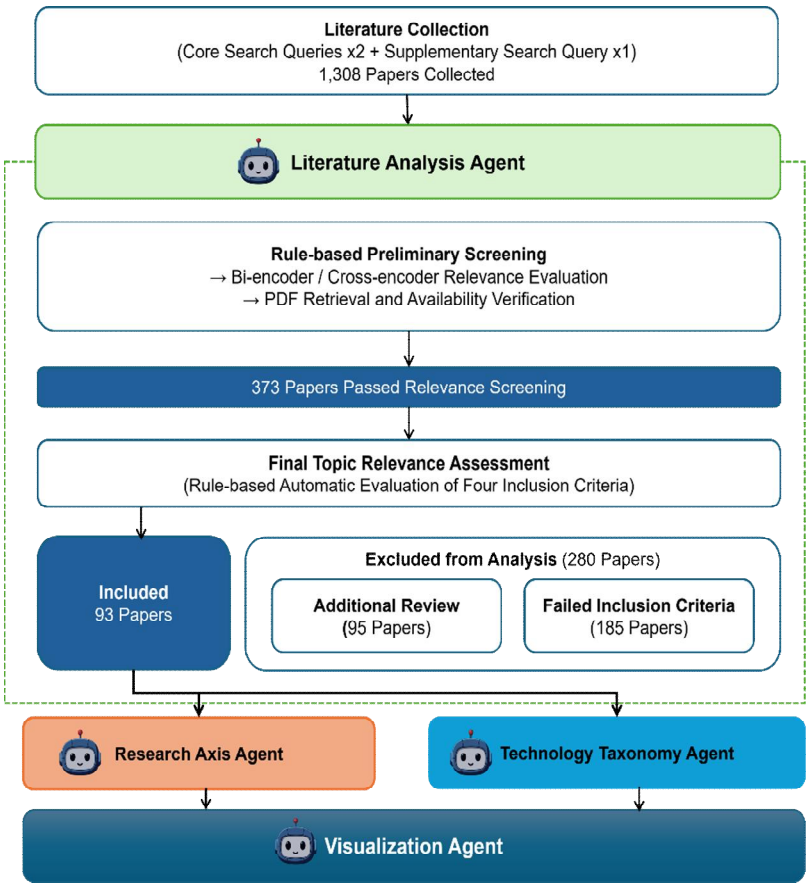

그림 1. 에이전트 기반 통합 프레임워크의 전체 분석 절차
Fig. 1. Overall workflow of the agent-based integrated literature analysis framework

연구 기여, 사용 기술 및 각 항목의 원문 근거를 구조화한다. Research Axis Agent는 Literature Analysis Agent가 추출한 연구 문제와 연구 기여의 의미적 유사성을 분석하여 주요 연구 방향을 도출한다. Technology Taxonomy Agent는 추출된 기술명, 기술 역할 및 원문 근거를 바탕으로 논문에서 각 기술이 수행하는 기능과 역할을 판별하고, 이를 사전에 정의된 기술 분류체계에 자동으로 매핑한다. Visualization Agent는 앞선 두 에이전트가 생성한 연구 방향과 기술 매핑 결과를 논문 식별자를 기준으로 통합하여 연구 방향별 기술 구성과 연도별 분포를 산출하고, 이를 표와 그래프로 시각화한다. 각 에이전트는 고유한 분석 역할을 수행하며, 선행 에이전트의 출력이 후속 에이전트의 입력으로 전달되는 순차적 의존 관계를 형성한다.

2. Literature Analysis Agent

문헌 선별 및 논문 분석은 Literature Analysis Agent가 담당한다. 이 에이전트는 사전에 정의된 검색식을 이용하여 관련 문헌을 자동으로 수집 및 선별하고, 제목과 초록을 바탕으로 연구 주제와의 관련성을 평가한다. 이후 선정된 논문의 원문 PDF에서 후속 분석에 필요한 연구 문제, 연구 목표, 연구 기여, 사용 기술, 기술 역할 및 각 항목의 원문 근거를 추출하여 구조화한다. Literature Analysis Agent에 부여한 프롬프트, 적용 모델 및 단계별 처리 절차는 <그림 2>와 같다.

<그림 2>와 같이 문헌 검색과 관련성 선별에는 사전에 정의된 검색식, 이중 인코더(Bi-encoder) 및 교차 인코더(Cross-encoder)를 순차적으로 적용하였으며, 최종 선정된

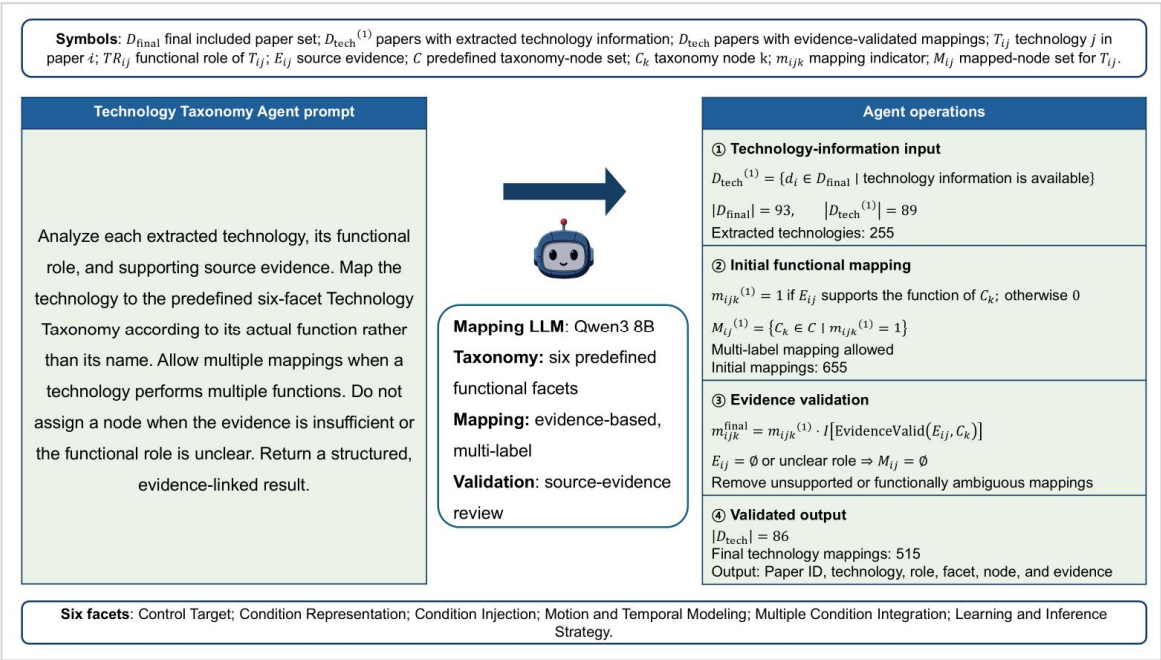

그림 2. Literature Analysis Agent의 문헌 검색과 선별 및 원문 정보 추출 절차
Fig. 2. Literature retrieval, screening, and full-text information extraction workflow of the Literature Analysis Agent

논문의 구조화된 정보 추출에는 Qwen3 8B를 사용하였다. 문헌 검색에는 핵심 검색식을 사용하였으며, 관련 문헌의 누락 가능성을 줄이기 위해 보완 검색식을 추가로 적용하였다. 핵심 검색식은 제목, 초록 및 전문을 대상으로 적용하였으며, 보완 검색식은 제목과 초록을 대상으로 적용하였다. 검색 기간은 2020년 1월 1일부터 2026년 7월 29일까지로 설정하였으며, 두 검색 결과를 통합하여 Text-to-Video 관련 후보 문헌 1,308편을 수집하였다.

규칙 기반 사전 선별을 통과한 문헌의 주제 관련성을 평가하기 위해 이중 인코더와 교차 인코더를 순차적으로 적용하였다. 기계학습 기반 문헌 선별은 대규모 후보 문헌 가운데 관련 가능성이 높은 문헌을 우선 검토하는 데 활용될 수 있다[27]. 1차 평가에는 이중 인코더 기반 문장 트랜스포머(Sentence Transformer) 임베딩 모델인 paraphrase-multilingual-MiniLM-L12-v2를 적용하였다. 논문의 제목과 초록을 결합한 텍스트와 고정 질의를 각각 임베딩 벡터로 변환하고 코사인 유사도를 관련성 점수로 사용하였다. 1차 기준을 통과한 문헌은 질의문과 제목 및 초록을 하나의 입력 쌍으로 처리하는 교차 인코더 모델인 ms-marco-MiniLM-L6-v2를 이용하여 재평가하였다[28,29]. 고정 질의문은 다음과 같다. “Controllable Diffusion-based Text-to-Video Generation: methods that guide video diffusion models using camera motion, pose, trajectory, layout, sketch, or other structural control signals to produce controllable video outputs from text prompts.”

이중 인코더 유사도 점수가 0.25 미만인 문헌은 제외하고, 0.25 이상 0.50 미만인 문헌은 경계 사례, 0.50 이상인 문헌은 높은 관련성 사례로 구분하였다. 이 가운데 유사도 점수가 0.25 이상인 문헌은 모두 교차 인코더 평가 단계로 전달하였다. 교차 인코더는 고정 질의문과 각 문헌의 제목 및 초록을 하나의 입력 쌍으로 처리하여 관련성 점수를 산출하였다[28]. 본 연구에서는 별도의 확률 정규화를 적용하지 않은 모델의 원시 출력값인 로짓(logit)을 관련성 점수로 사용하였다. 로짓은 특정 구간으로 제한되지 않으며, 본 연구에서 관찰된 점수는 -9.9612에서 7.4951까지 분포하였다. 점수가 클수록 질의문과 문헌 간 관련성이 높은 것으로 해석하였다. 교차 인코더 점수가 -1.0 미만인 문헌은 제외하고, -1.0 이상 1.0 미만인 문헌은 예비 분석 후보로 보존하

였으며, 1.0 이상인 문헌은 원문 PDF의 접근 가능성, 다운로드 가능 및 열람 가능 여부를 확인한 후 최종 주제 적합성 선별 단계로 전달하였다. 그 결과 총 373편이 해당 단계를 통과하였다. 여기서 -1.0과 1.0은 모델 출력값의 범위를 의미하는 것이 아니라 문헌 선별을 위해 설정한 임계값이다. 이 임계값은 공인된 표준값이 아니며, 예비 검토를 통해 각 점수 구간에 포함된 논문들의 실제 연구 주제 부합도를 확인한 뒤 경험적으로 설정하였다.

그러나 해당 문헌이 모두 확산 모델 기반의 제어 가능한 T2V 연구에 해당하는 것은 아니므로, 에이전트는 관련성 평가를 통과한 373편에 대하여 다음 네 가지 최종 포함 기준의 충족 여부를 판단하였다. ① 제목에서 비디오 생성 연구임이 확인되는 경우, ② 제목 또는 초록에서 확산 모델, 확산 트랜스포머(Diffusion Transformer, DiT) 또는 이를 기반으로 한 구조가 확인되는 경우, ③ 제목과 초록 또는 원문에서 카메라, 궤적, 포즈, 레이아웃, 참조 이미지, 정체성 또는 오디오와 같은 외부 제어 조건이 확인되는 경우, ④ 사전에 정의한 제외 작업이 주된 연구 목적이 아닌 경우를 판단하였다. 제목만으로 외부 제어 조건 사용 여부가 명확하지 않은 문헌은 초록 또는 원문을 추가로 분석하였다.

이 과정을 통해 Literature Analysis Agent는 93편을 후속 분석 대상으로 선정하고, 원문 PDF에서 필요한 정보를 추출 및 구조화하였다. 검증된 연구 문제와 연구 기여는 Research Axis Agent에, 사용 기술, 기술 역할 및 원문 근거는 Technology Taxonomy Agent에 각각 입력 자료로 전달하였다.

3. Research Axis Agent

Research Axis Agent는 Literature Analysis Agent가 구조화한 연구 문제와 연구 기여를 바탕으로 의미적 유사성을 분석하여 주요 연구 방향을 도출한다. 각 논문의 연구 문제와 연구 기여는 별도의 가중치 없이 하나의 텍스트로 결합하고, 결합된 텍스트를 임베딩 벡터로 변환한 후

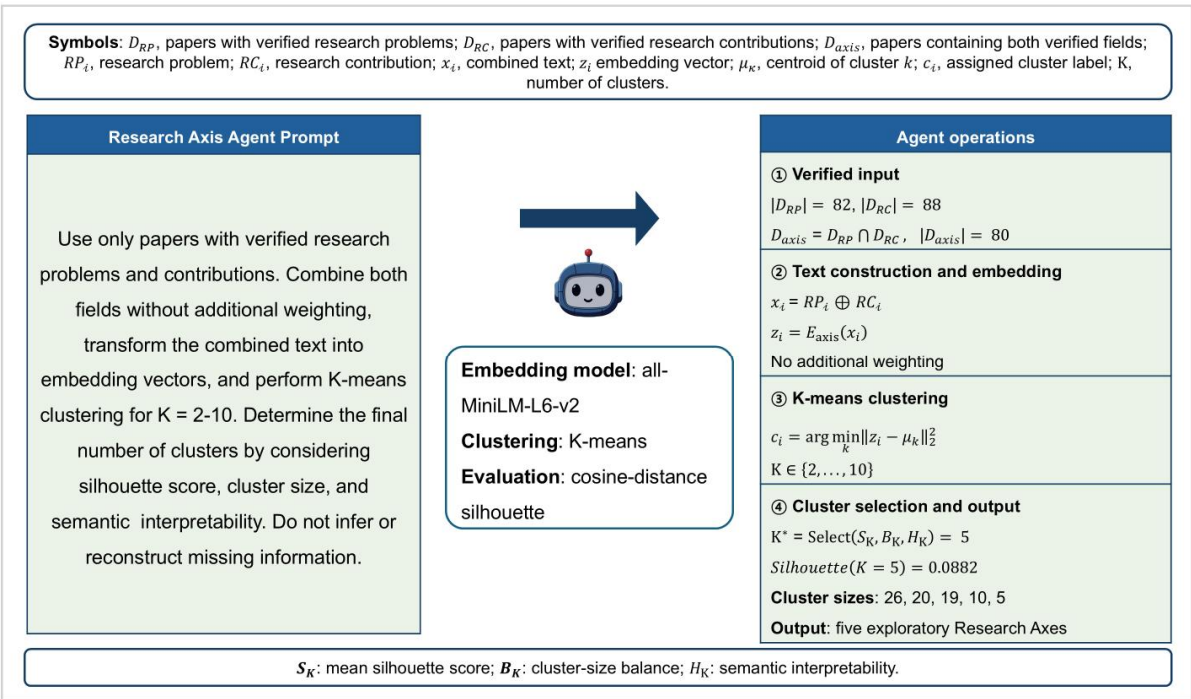

그림 3. Research Axis Agent의 검증 자료 구성과 연구 방향 도출 절차
Fig. 3. Verified-data construction and research-axis derivation workflow of the Research Axis Agent

K-means 군집화를 적용하였다[30]. 군집 수는 실루엣 점수[31], 군집별 논문 수 및 연구 내용의 의미적 해석 가능성을 종합하여 결정하였다. Research Axis Agent의 입력 자료, 프롬프트 및 연구 방향 도출 절차는 <그림 3>과 같다.

<그림 3>과 같이 Literature Analysis Agent가 선정한 93편 가운데 연구 문제는 82편, 연구 기여는 88편에서 원문 근거와의 대응 관계가 검증되었다. Research Axis Agent는 두 항목을 모두 보유한 교집합 80편만을 입력 자료로 사용하였으며, 누락된 내용은 추정하거나 보완하지 않았다. 각 논문의 연구 문제와 연구 기여를 별도의 가중치 없이 하나의 텍스트로 결합하고, 문장형 트랜스포머 기반의 all-MiniLM-L6-v2를 이용하여 임베딩 벡터로 변환하였다. 이후 K-means를 적용하여 의미적으로 유사한 논문을 군집화하였다. 군집 수는 2부터 10까지 변화시키면서 평균 실루엣 점수, 군집별 논문 수 및 연구 문제와 연구 기여의 의미적 구분 가능성을 비교하여 결정하였다. 이를 종합적으로 고려하여 최종 군집 수는 5로 결정하였으며, 80편은 각각 26편, 20편, 19편, 10편 및 5편으로 구성된 다섯 개의 Research Axis로 분류되었다. 평균 실루엣 점수는 0.0882로 군집 간 분리가 강하지 않았으므로, 도출된 Research Axis는 의미적으로 일부 중첩될 수 있는 탐색적 분류로 해석하였다. 논문별 Research Axis 분류 결과는 Visualization Agent의 입력 자료로 전달하였다.

4. Technology Taxonomy Agent

Technology Taxonomy Agent는 Literature Analysis Agent가 구조화한 기술명, 기술 역할 및 원문 근거를 바탕으로 논문별 기술을 기능 중심의 공통 기준에 따라 분류한다. 본 연구의 Technology Taxonomy는 Nickerson 등의 개념적-경험적 분류체계 개발 접근과 Kundisch 등의 설계 및 평가 원칙을 참고하여 구성하였다[23,24]. 먼저 제어 가능한 확산 기반 T2V 생성 과정에서 기술이 수행하는 기능적 역할을 메타특성(meta-characteristic)으로 설정하고, 선행연구의 분류 기준과 T2V 생성 및 제어 과정을 바탕으로 여섯 개의 관점을 개념적으로 정의하였다. 여섯 관점은 무엇을 제어하는가를 나타내는 제어 대상, 제어 정보를 어떠한 형태로 나타내는가를 설명하는 조건 표현, 제어 정보를 생성 모델에 어떻게 반영하는가를 구분하는 조건 주입 방식, 움직임과 프레임 간 시간 정보를 어떻게 처리하는가를 나타내

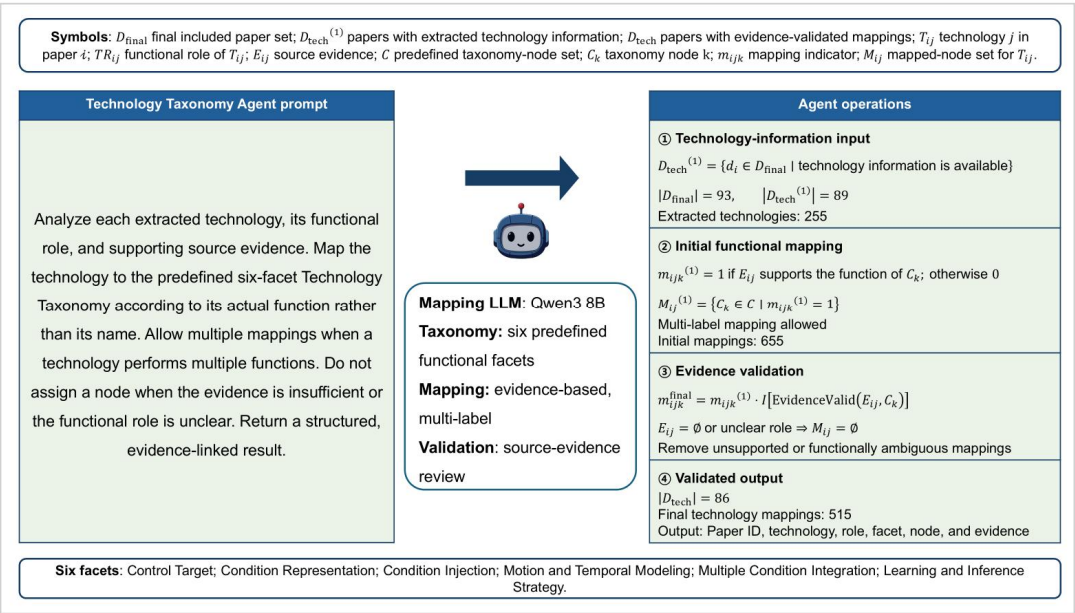

그림 4. Technology Taxonomy Agent의 기능 중심 기술 매핑 및 원문 근거 검증 절차
Fig. 4. Function-centered technology mapping and source-evidence validation workflow of the Technology Taxonomy Agent

는 움직임 및 시간 처리, 복수의 제어 조건을 어떻게 결합하고 조정하는가를 설명하는 복수 제어의 조정, 그리고 제어 기능을 학습 및 추론 과정에서 어떻게 구현하는가를 구분하는 학습 및 추론 방식으로 구성하였다. 이러한 여섯 관점은 제어 조건의 정의와 표현부터 모델 내 반영, 시간적 처리, 복수 조건 간 조정 및 학습, 추론 단계의 구현까지 제어 가능한 T2V 기술이 수행하는 기능적 역할을 분석할 수 있도록 설정하였다. 선정 문헌에서 추출한 기술명, 기술 역할 및 원문 근거는 사전에 정의한 Taxonomy 항목에 매핑하였다. 하나의 기술이 복수의 기능을 수행하는 경우에는 다중 매핑을 허용하되, 원문 근거가 부족하거나 기술의 기능적 역할이 명확하지 않은 경우에는 특정 Taxonomy 항목에 배정하지 않았다. Technology Taxonomy Agent의 입력 자료, 프롬프트, 기능 중심 매핑 및 원문 근거 검증 절차는 <그림 4>와 같다.

<그림 4>와 같이 Literature Analysis Agent가 선정한 93편 가운데 기술 정보가 확인된 89편을 Technology Taxonomy Agent의 입력 자료로 사용하였다. 에이전트는 89편에서 추출된 255개 기술의 실제 기능과 사전에 정의된 Taxonomy 항목의 부합 여부를 분석하여 655건의 1차 기술 매핑을 생성하였다. 하나의 기술이 여러 기능을 수행하는 경우에는 복수 항목에 연결하였으며, 원문에서 해당 기술의 기능을 확인할 수 없는 경우에는 특정 항목에 강제로 배정하지 않았다.

이후 에이전트는 기술 역할과 원문 근거가 Taxonomy 정의에 부합하는지를 재검토하고, 근거가 부족하거나 기능이 명확하지 않은 연결을 제외하였다. 그 결과 86편에서 515건의 최종 기술 매핑을 확정하였다. 여기서 기술 매핑 1건은 특정 논문에서 확인된 하나의 기술이 그 기능적 역할에 따라 하나의 Taxonomy 항목에 연결된 경우를 의미한다. 하나의 기술이 복수 기능을 수행하면 여러 Taxonomy 항목에 각각 매핑될 수 있으므로, 추출된 기술 수와 최종 매핑 건수는 서로 다르다. 논문별 최종 기술 매핑 결과는 Visualization Agent의 입력 자료로 전달하였다.

5. Visualization Agent

Visualization Agent는 Research Axis Agent의 연구 방향 분류 결과와 Technology Taxonomy Agent의 기술 매핑 결

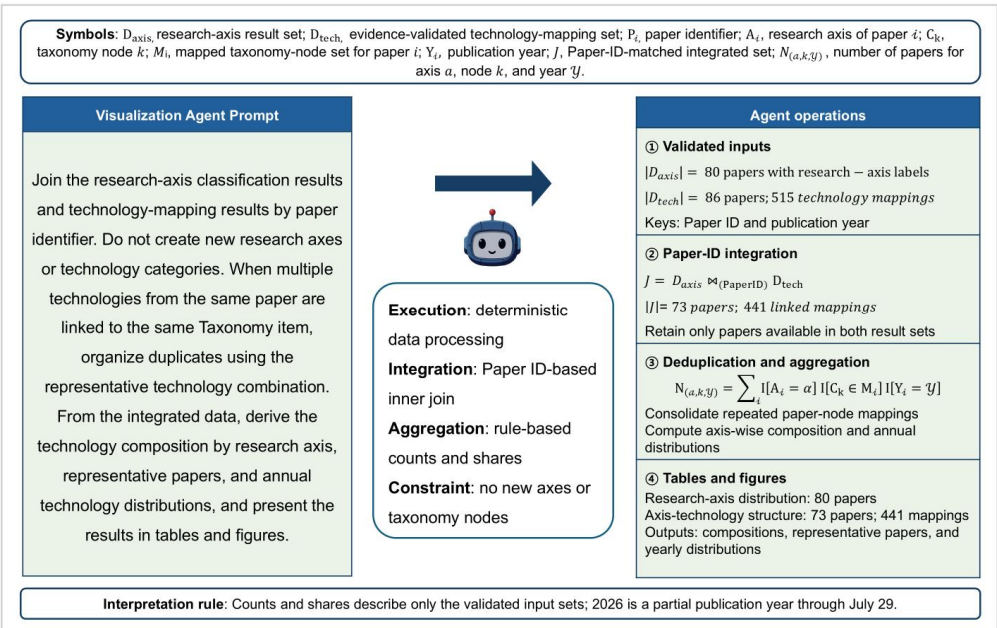

그림 5. Visualization Agent의 연구 방향과 기술 매핑 통합 및 시각화 절차
Fig. 5. Integration and visualization workflow for research-axis classifications and technology mappings in the Visualization Agent

과를 논문 식별자를 기준으로 결합한다. 이 단계에서는 새로운 Research Axis나 Technology Taxonomy 항목을 생성하지 않으며, 두 분석 결과를 모두 보유한 논문만 통합 분석 대상으로 사용한다. 동일 논문의 여러 기술이 같은 Taxonomy 항목에 연결된 경우에는 중복 매핑을 정리하고, 통합 결과를 바탕으로 연구 방향별 기술 구성, 대표 논문 및 연도별 분포를 산출하였다. Visualization Agent의 입력 자료, 결합 조건, 집계 및 시각화 절차는 <그림 5>와 같다.

<그림 5>와 같이 Research Axis가 부여된 80편과 기술 매핑이 확정된 86편의 515건 기술 매핑을 논문 식별자를 기준으로 결합하였다. 두 분석 결과를 모두 보유한 문헌의 교집합은 73편이었으며, 이에 연결된 441건의 기술 매핑을 통합 분석 대상으로 확정하였다. 이후 동일 논문과 Taxonomy 항목 간의 중복 매핑을 정리하고, 연구 방향별 기술 구성, 대표 논문 및 연도별 분포를 산출하여 표와 그래프로 시각화하였다.

Research Axis의 전체 분포와 연도별 구성비는 Research Axis가 부여된 80편을 기준으로 산출하였으며, 연구 방향별 기술 구조는 Research Axis와 Technology Taxonomy 결과가 모두 확인된 73편의 441건 기술 매핑을 기준으로 분석하였다.

Ⅳ. 연구 결과

본 장에서는 서론에서 설정한 네 가지 연구 질문에 대한 분석 결과를 제시한다. 4.1절에서는 제어 가능한 확산 기반 T2V 연구의 주요 연구 방향을 도출하여 RQ1에 답하고, 출판 연도별 연구 방향의 구성 변화를 분석하여 RQ4에 답한다. 4.2~4.4절에서는 각 연구 방향에서 사용된 기술을 기능적 역할에 따라 분석하여 RQ2에 답하고, 연구 방향별 기술 구조의 공통점과 차이를 통해 연구 방향과 기술의 기능적 역할 간 관계를 제시하여 RQ3에 답한다.

1. 확산 모델 기반의 제어 가능한 T2V 연구 방향 및 연도별 변화

Literature Analysis Agent에서 선정한 93편 가운데 연구 문제와 연구 기여가 모두 검증된 80편을 Research Axis Agent의 연구 방향 분석에 사용하였다. Research Axis Agent는 두 항목의 의미적 유사성을 바탕으로 다섯 개의 군집을 도출하고, 군집별 공통 연구 문제, 연구 기여 및 기술적 특성을 검토하여 Motion-Field and Explicit Trajectory Control(AXIS_01), Structured Temporal and Condition-

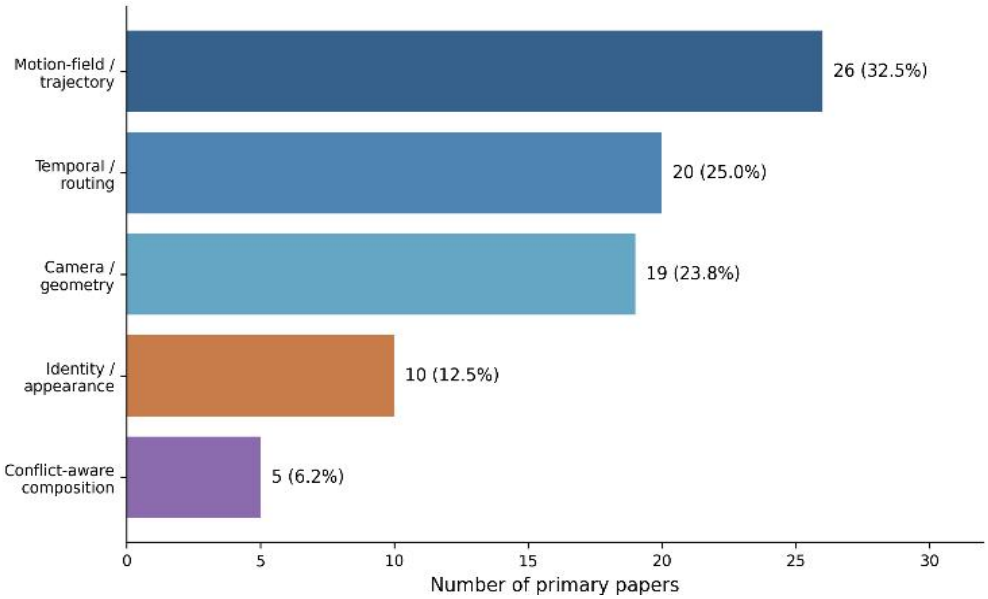

그림 6. 다섯 연구 방향별 주축 논문 분포
Fig. 6. Distribution of primary-axis papers across five research axes

Routing Control(AXIS_02), Camera, Geometry, and 3D Viewpoint Control(AXIS_03), Identity and Appearance Preservation(AXIS_04), Conflict-Aware Composite Motion Control(AXIS_05)로 명명하였다.

다섯 연구 방향별 주축 논문 분포는 <그림 6>과 같다. AXIS_01은 26편(32.5%)으로 가장 큰 비중을 차지하였으며, AXIS_02는 20편(25.0%), AXIS_03은 19편(23.8%)이었다. 세 방향은 전체의 81.3%를 차지하였다. AXIS_04는 10편(12.5%), AXIS_05는 5편(6.2%)으로 상대적으로 적었다.

다섯 연구 방향 간 관계를 설명하기 위해 각 Research Axis의 군집명, 군집별 대표 연구 문제, 연구 기여 및 기술 매핑 결과를 비교하였다. 이후 연구가 주로 제어하는 대상과 생성 과정에서 수행하는 기능의 유사성을 기준으로 다섯 연구 방향을 세 개의 상위 연구 영역으로 해석적으로 묶었다. AXIS_01·02·05는 Motion and Temporal Control, AXIS_03은 Camera and 3D Viewpoint Control, AXIS_04는 Identity and Appearance Customization으로 구분하였다. 이는 다섯 연구 방향의 관계를 설명하기 위한 구조이며 <그림 7>과 같다.

80편의 주축 연구 방향을 출판 연도별로 비교한 결과는 <그림 8>과 같다. 검색 기간 내 연구 구성의 변화가 관찰되었다. 2023년에는 Structured Temporal and Condition-Routing Control이 63.6%로 가장 큰 비중을 차지하였다. 2024년에는 Motion-Field and Explicit Trajectory Control과 Camera, Geometry, and 3D Viewpoint Control이 각각 35.7%, 2025년에는 Motion-Field and Explicit Trajectory Control이 34.8%로 가장 높은 비중을 차지하였다. Identity and Appearance Preservation의 구성비는 2024년 7.1%, 2025년 13.0%, 2026년 부분 자료에서 27.8%로 나타났다. Conflict-Aware Composite Motion Control은 2023년에는 확인되지 않았으며, 2024년 이후의 자료에서는 매년 확인되었다.

연구 방향별 사례는 군집 중심과의 코사인 유사도, 인용 수 및 기술적 대표성을 함께 검토하여 선정하였다. 인용 수는 2026년 7월 29일 기준 OpenAlex 값을 사용하였다. 사례로는 AXIS_01의 Time-to-Move: Training-Free Motion Controlled Video Generation via Dual-Clock Denoising[32], AXIS_02의 ControlVideo: Training-Free Controllable Text-to-Video Generation[33], AXIS_03의 VD3D: Taming

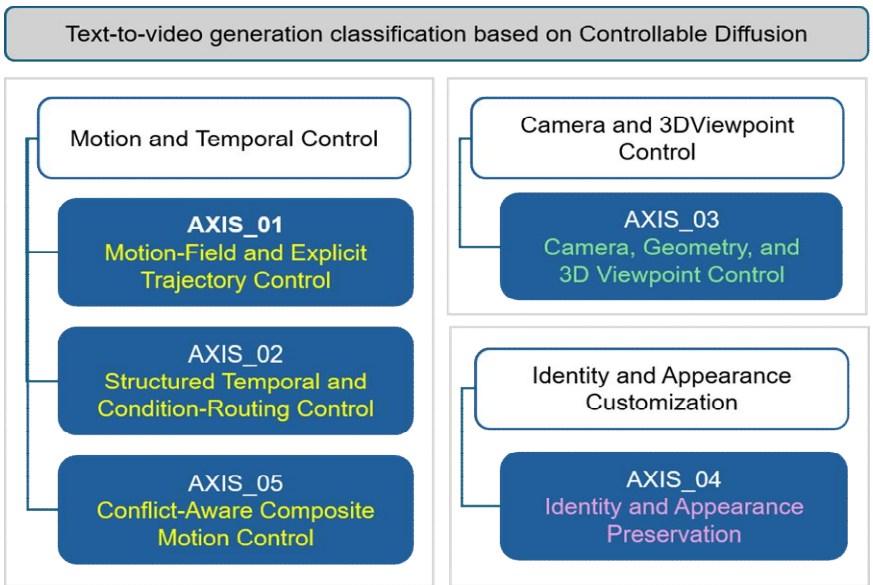

그림 7. 다섯 연구 방향의 상위 연구 영역
Fig. 7. Higher-level research domains of the five research axes

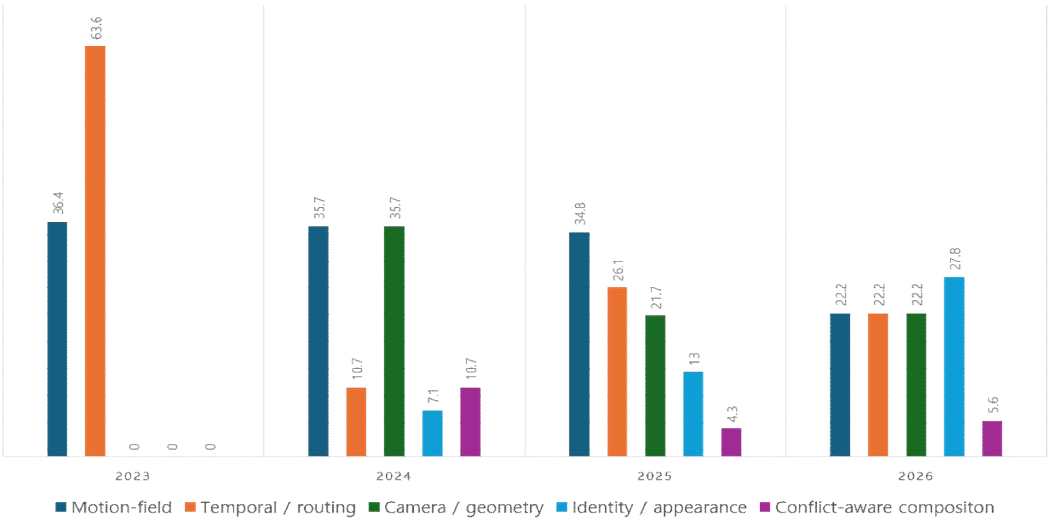

그림 8. 출판 연도별 연구 방향 구성비
Fig. 8. Annual composition of primary research axes

Large Video Diffusion Transformers for 3D Camera Control[34], AXIS_04의 Magic-Me: Identity-Specific Video Customized Diffusion[35], AXIS_05의 DualReal: Adaptive Joint Training for Lossless Identity-Motion Fusion in Video Customization[36]을 제시하였다.

한편, Technology Taxonomy Agent는 기술의 역할과 원문 근거가 검증된 86편에서 515건의 최종 기술 매핑을 확정하였다. Visualization Agent가 연구 방향이 도출된 80편과 기술 매핑이 확정된 86편을 논문 식별자를 기준으로 결합한 결과, 두 분석 결과를 모두 보유한 73편과 이에 연결된 441건의 기술 매핑이 교집합으로 확인되었다. 따라서 연구 방향의 전체 분포와 연도별 구성비는 80편을 기준으로 분석하고, 연구 방향별 기술 구조는 73편의 441건 기술 매핑을 기준으로 분석하였다. 이하에서는 세 개의 상위 연구 영역과 다섯 연구 방향의 특성을 설명하고, 통합 분석 대상 73편을 기준으로 연구 방향별 기술 구조를 제시한다.

2. Motion and Temporal Control

Motion and Temporal Control은 생성 영상에서 객체와 인물의 움직임을 사용자의 의도에 맞게 제어하고, 이러한 제어 정보를 시간축 전반에 걸쳐 안정적으로 유지하는 연구 영역이다. 본 연구에서는 이 영역을 Motion-Field and Explicit Trajectory Control(AXIS_01), Structured Temporal and Condition-Routing Control(AXIS_02), Conflict-Aware Composite Motion Control(AXIS_05)의 세 연구 방향으로 구분하였다.

세 연구 방향은 모두 움직임을 다룬다는 공통점을 가지지만 연구의 초점은 서로 다르다. AXIS_01은 움직임을 어떠한 형태로 명시적으로 표현하고 생성 모델에 전달할 것인가에 초점을 둔다. AXIS_02는 제어 조건을 영상 전체에 걸쳐 시간적으로 일관되게 유지하고 필요한 위치에 전달하는 문제를 다룬다. AXIS_05는 여러 제어 조건을 동시에 적용할 때 발생하는 충돌을 완화하고 조건 간 상호작용을 조정하는 문제를 중심으로 한다. Motion and Temporal Control 영역에 포함된 세 연구 방향의 주요 제어 조건과 처리 과정은 <그림 9>와 같다. 이 그림은 각 연구 방향에 속한 복수 논문에서 확인된 기술 요소를 입력 조건, 조건 표현, 조건 처리, 생성 및 출력 단계에 따라 종합한 개념적 처리 과정이다. 따라서 특정 논문의 단일 모델 구조나 기술의 시간적 발전 순서를 의미하지 않는다.

2.1 Motion-Field and Explicit Trajectory Control

AXIS_01은 객체나 카메라의 움직임을 사용자가 지정한 방향과 경로에 따라 제어하는 연구 방향이다. 텍스트만으로는 움직이는 대상과 이동 경로를 정밀하게 지정하기 어려우므로, 이 연구축에서는 다양한 명시적 움직임 정보를 생성 과정에 반영한다. AXIS_01로 분류된 26편 가운데 연구 방향 분류 결과와 기술 매핑 결과가 모두 확인된 논문은 22편이었다. 주요 기술 요소로는 Motion, Dense Motion Field, Control Token, Auxiliary Control Network, Input Level, Motion Field Modeling 및 Inference-time Control이 확인되었다. 이는 AXIS_01에 속한 논문들이 움직임 정보를 모델이 처리할 수 있는 형태로 표현하고 생성 과정에 반영하며, 이 가운데 추론 단계에서 제어하는 방식도 빈번하게 사용했음을 보여준다. 대표 사례인 Time-to-Move: Training-Free Motion Controlled Video Generation via Dual-Clock Denoising[32]은 입력 이미지의 일부 영역을 잘라 이동시키거나 단안 깊이 기반으로 재투영하여 만든 거친 참조 애니메이션을 움직임 단서로 사용한다. 이후 제어 영역을 나타내는 마스크와 이중 시간축 잡음 제거(Dual-Clock Denoising)를 이용하여, 사용자가 움직임을 지정한 영역에는 강한 정렬을 적용하고 나머지 영역에는 더 큰 생성 자유도를 부여한다. 이 방식은 별도의 추가 학습 없이 기존의 이미지 기반 비디오 확산 모델(image-to-video diffusion model, I2V)의 추론 과정을 조정하여 움직임과 외형을 제어하는 AXIS_01의 사례이다.

2.2 Structured Temporal and Condition-Routing Control

AXIS_02는 제어 조건을 개별 프레임에 독립적으로 적용하지 않고 영상 전체에 일관되게 전달하고 유지하는 연구 방향이다. 프레임별 생성에서 발생할 수 있는 외형 및 구조의 불일치와 깜빡임을 줄이기 위해 조건 전달과 시간 정보 처리에 초점을 둔다. AXIS_02로 분류된 20편 가운데 연구 방향 분류 결과와 기술 매핑 결과가 모두 확인된 논문은 19편이었다. 주요 기술 요소로는 Sparse Spatial, Sparse Trajectory, Attention Based, Input or Latent Initialization, Spatiotemporal Modeling 및 Temporal Consistency가 확인되었다. 다중 조건 처리에서는 Condition Routing or

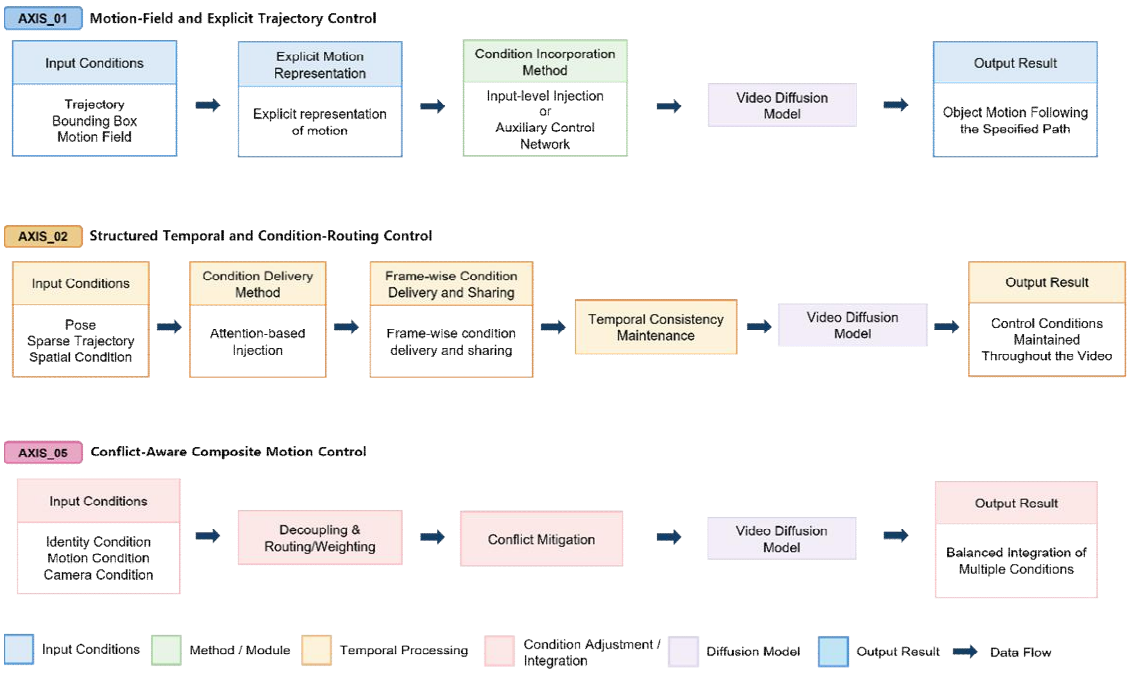

그림 9. Motion and Temporal Control 영역의 세부 연구 방향별 제어 과정
Fig. 9. Control Processes by Research Direction in the Motion and Temporal Control Domain

Weighting과 Condition Fusion이 나타났다. 이는 제어 정보를 프레임 사이에 전달하고 시간축에서 유지하는 AXIS_02의 기술적 구성을 보여준다. 대표 사례인 ControlVideo: Training-Free Controllable Text-to-Video Generation[33]은 사전 학습된 ControlNet을 별도의 추가 학습 없이 비디오 생성에 적용한다. Fully cross-frame interaction은 모든 프레임의 특징이 서로 참조되도록 하여 외형의 일관성을 높이고, interleaved-frame smoother는 연속된 세 프레임의 가운데 프레임을 보간하여 영상의 깜빡임을 줄인다. Hierarchical sampler는 먼저 주요 프레임을 생성한 뒤, 그 사이의 짧은 구간을 생성하여 장시간 비디오의 일관성을 유지한다. 이 방법은 구조적 제어 조건을 반영하면서 프레임 간 일관성과 연속성을 확보하는 AXIS_02의 접근을 보여준다[33].

2.3 Conflict-Aware Composite Motion Control

AXIS_05는 여러 제어 조건을 함께 적용할 때 발생하는 간섭과 충돌을 조정하는 연구 방향이다. 서로 다른 조건이 동일한 생성 과정에 영향을 미치는 상황에서 각 조건의 역할을 분리하고 그 영향을 조절하는 문제에 초점을 둔다. AXIS_05로 분류된 5편 가운데 연구 방향 분류 결과와 기술 매핑 결과가 모두 확인된 논문은 5편이었다. 주요 기술 요소로는 Motion, Motion Decoupling, Condition Routing or Weighting, Conflict Mitigation 및 Control Branch or Adapter가 확인되었다. 이는 복수 조건의 영향을 분리 및 조정하고 조건 간 충돌을 완화하는 AXIS_05의 기술적 구성을 보여준다. 대표 사례인 DualReal: Adaptive Joint Training for Lossless Identity-Motion Fusion in Video Customization[36]은 정체성과 움직임을 독립적으로 학습한 뒤 결합할 때 발생하는 충돌을 해결하기 위해 두 정보를 함께 학습하는 방식을 제안한다. Dual-aware Adaptation은 정체성과 움직임 학습을 번갈아 수행하면서 한쪽의 사전 정보로 다른 쪽의 학습을 유도하고, 정규화를 통해 정보 누출을 방지한다. StageBlender Controller는 노이즈 제거 단계와 DiT의 깊이에 따라 정체성과 움직임 특징의 반영 비중을 조절한다[36].

3. Camera and 3D Viewpoint Control

AXIS_03은 카메라의 위치와 방향, 이동 경로를 제어하면서 장면의 3차원 구조와 시점의 일관성을 유지하는 연구 영역이다. 자연어만으로는 정확한 카메라 경로를 지정하기 어려우므로, camera pose와 camera trajectory를 직접 사용

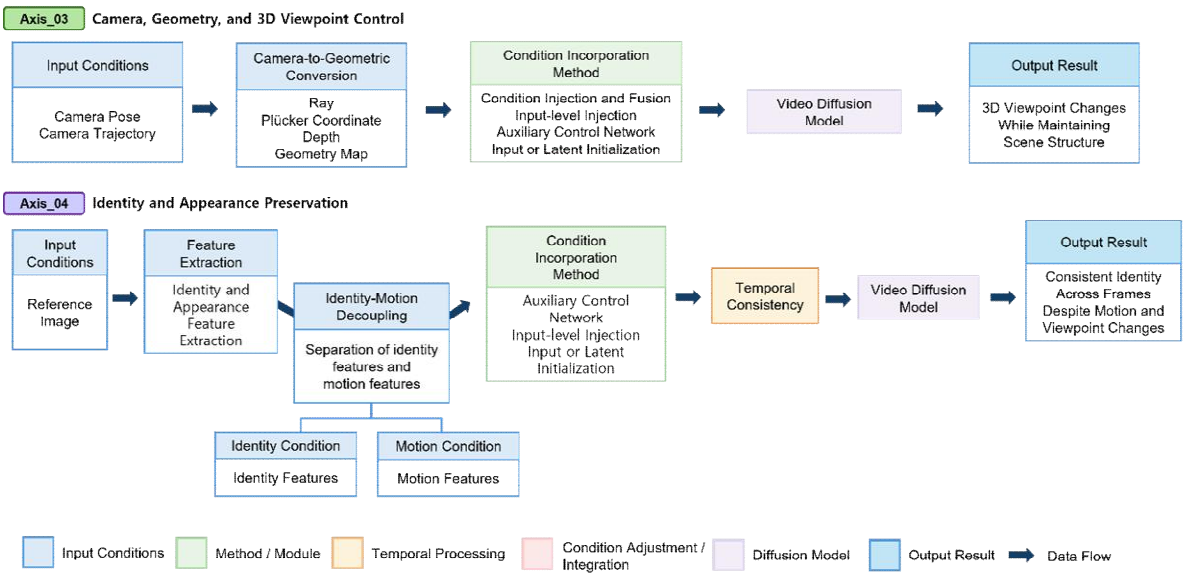

그림 10. Camera and 3D Viewpoint Control 및 Identity and Appearance Preservation 연구 방향의 제어 과정
Fig. 10. Control Processes for Camera and 3D Viewpoint Control and Identity and Appearance Preservation

하거나 ray, depth, geometry map 및 Plücker coordinate와 같은 기하학적 표현으로 변환하여 생성 모델에 전달한다. 주요 제어 과정은 <그림 10>과 같다. AXIS_03으로 분류된 19편 가운데 연구 방향 분류 결과와 기술 매핑 결과가 모두 확인된 논문은 17편이었다. 주요 제어 대상은 Camera와 Scene Geometry였으며, Camera Parameter와 Depth or Geometry Map이 주요 표현으로 확인되었다. 조건 주입에는 Input Level, Auxiliary Control Network 및 Input or Latent Initialization이 사용되었고, 카메라 정보를 기하학적 형태로 변환하는 Camera-to-Geometric Conversion과 여러 조건을 결합하는 Condition Fusion도 주요 기술 요소로 확인되었다. 원문에서는 정밀한 카메라 제어, 카메라 주석과 다중 시점 데이터의 확보, 추가 미세조정 및 생성 품질 유지가 주요 과제로 확인되었다. 대표 사례인 VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control[34]은 트랜스포머 기반 video diffusion model에 3차원 카메라 제어를 적용하는 방법을 제안하였다. 각 프레임의 카메라 내부와 외부 파라미터로부터 픽셀별 광선을 계산하고, 이를 Plücker coordinate 기반의 시공간 카메라 표현으로 변환한다. 변환된 정보는 패치 토큰으로 구성된 뒤, 교차 어텐션을 포함하는 ControlNet 유사 구조를 통해 Video Diffusion Transformer에 반영된다.

4. Identity and Appearance Preservation

Identity and Appearance Preservation은 자세, 시점, 표정, 조명 및 배경이 변하더라도 인물이나 객체의 정체성과 외형을 유지하는 연구 방향이다. 참조 이미지 기반 생성에서 대상의 특징이 손실되거나 프레임마다 외형이 달라지는 문제를 다루며, 주요 제어 과정은 <그림 10>과 같다. AXIS_04로 분류된 10편 가운데 연구 방향 분류 결과와 기술 매핑 결과가 모두 확인된 논문은 10편이었다. 주요 제어 대상은 Content and Appearance와 Subject Identity였으며, Identity or Appearance Reference가 주요 조건 표현으로 확인되었다. 조건 주입에는 Auxiliary Control Network, Input Level 및 Input or Latent Initialization이 사용되었고, 시간 처리에서는 Identity- or Appearance-Motion Decoupling, Motion Decoupling 및 Temporal Consistency가 나타났다. 원문에서는 자세, 표정, 조명과 전신 움직임에 따른 정체성 변화가 주요 문제로 확인되었으며, 이를 해결하기 위해 정체성 특징의 분리와 정렬, 다중 규모 특징 추출, 공간적 배치, 프레임 간 일관성 강화 및 참조 프레임 활용 등의 접근이 사용되었다. 대표 사례인 Magic-Me: Identity-Specific Video Customized Diffusion[35]은 소수의 참조 이미지로 특정 인물의 정체성을 유지하면서 다양한 장면과 움직임을 생성하는 Video Custom Diffusion 프레임워크를 제안한다. 확장된 Textual Inversion 기반 ID Module은 잘라낸 정체성 영역을 이용하여 배경과 구분되는 인물 특징을 학습하고, 3D Gaussian Noise Prior는 초기화 단계에서 프레임 간 상관관계를 반영하여 시간적 안정성을 높인다. 생성은 T2V VCD, Face VCD, Tiled VCD의 세 단계로 진행되며, 얼굴의 정체성 특징을 강화하고 영상의 해상도를 향상시킨다. 이는 참조 이미지의 정체성 정보를 특징 추출, 노이즈 초기화 및 후처리 단계에 걸쳐 반영함으로써 외형과 프레임 간 일관성을 유지하는 AXIS_04의 사례이다.

연구 방향별 기술 구조를 종합하면, 하나의 연구 방향에서 복수의 기술적 기능이 함께 사용되었으며, 조건 주입, 조건 결합 및 시간적 일관성 유지와 같은 동일한 기능도 여러 연구 방향에서 공통적으로 확인되었다. 이는 연구 방향과 기술의 기능적 역할이 일대일로 대응하지 않고 다대다 관계를 형성함을 보여준다. 이러한 관계의 구체적인 매핑과 그 의미는 5장에서 논의한다.

Ⅴ. 논의

1. 연구 방향의 구조화(RQ1)

본 연구에서는 Research Axis Agent를 통해 각 논문이 상대적으로 중점적으로 다루는 문제를 나타내는 다섯 개의 탐색적 연구 방향을 도출하였다. 군집 수는 평균 실루엣 점수만을 기준으로 하지 않고, 군집별 논문 분포와 연구 내용의 해석 가능성을 함께 고려하여 결정하였다. 따라서 다섯 개의 Research Axis는 통계적으로 완전히 분리된 범주가 아니라, 연구 문제와 연구 기여의 유사성을 바탕으로 논문의 중심적인 연구 방향을 구분한 탐색적 구조로 해석해야

한다. 각 Research Axis는 실제 생성 과정에서도 상호 보완적으로 연결된다. 명시적 움직임 제어는 시간적 일관성 유지와, 카메라 이동은 장면의 기하구조 및 객체 움직임과, 정체성 보존은 움직임의 자유도와 복수 조건 조정과 연결될 수 있다. 따라서 연구 방향과 각 기술의 기능을 함께 분석할 필요가 있다.

2. 기술의 기능적 역할 분류(RQ2)

RQ2와 관련하여, 본 연구는 Technology Taxonomy Agent를 통해 제어 대상, 조건 표현, 조건 주입 방식, 움직임 및 시간 처리, 복수 제어의 조정, 학습 및 추론 방식의 여섯 관점으로 구성된 기능 중심 분류체계를 정의하고, 86편에서 515건의 최종 기술 매핑을 확정하였다. 이 분류체계는 기술의 명칭이나 구현 방식이 아니라 각 기술이 생성 과정에서 수행하는 기능적 역할을 기준으로 하므로, 서로 다른 명칭이나 구현 방식을 가진 기술도 동일한 기능적 범주에서 비교할 수 있다.

3. 연구 방향과 기술의 기능적 역할 간 다대다 관계(RQ3)

RQ3과 관련하여, Research Axis는 논문이 해결하려는 연구 문제를 나타내고, Technology Taxonomy는 이를 구현하는 기술의 기능적 역할을 나타낸다. Visualization Agent가 두 결과를 논문 식별자를 기준으로 결합한 결과, 하나의 Research Axis에는 여러 기술 요소가 포함되고, 동일한 기술 요소도 여러 Research Axis에서 확인되는 다대다 관계가 <표 1>과 같이 나타났다.

조건 주입, 복수 조건 결합 및 시간적 일관성 관련 기술은 여러 Research Axis에서 공통으로 확인되었다. 이는 하나의 연구 문제를 해결하기 위해 여러 기술 기능이 함께 활용되고, 동일한 기술 기능도 서로 다른 연구 문제를 해결하는데 사용될 수 있음을 보여준다. 따라서 Research Axis Agent와 Technology Taxonomy Agent는 서로를 대체하는 분류가 아니라, 연구 목적과 기술적 해결 방식을 각각 설명하는 상호 보완적 분석 체계로 이해할 수 있다.

4. 연구 방향의 시간적 변화(RQ4)

RQ4와 관련하여, 연도별 연구 방향의 구성 변화를 분석한 결과 2023년에는 제어 조건의 시간적 전달, 2024년과 2025년에는 명시적 움직임과 카메라 및 기하 제어가 상대적으로 높은 비중을 차지하였다. 2026년 검색 종료일까지의 부분 자료에서는 정체성과 외형 유지의 구성비가 상대적으로 높게 확인되었다. 이는 분석 기간 동안 시기별로 서로 다른 제어 문제가 상대적으로 높은 비중으로 관찰되었음을 보여준다. 다만 연도별 결과는 본 연구의 검증 조건을 충족한 제한된 문헌을 기반으로 하며, 2026년 자료는 7월 29일까지 수집된 부분 자료이다. 따라서 연도 간 차이는 특정 기술의 지속적인 증가나 감소를 나타내는 추세가 아니라, 설정된 분석 조건과 검색 기간 내에서 관찰된 연구 구성의 차이로 해석해야 한다.

표 1. 다섯 연구 방향과 대표 Technology Taxonomy 요소의 매핑 관계
Table 1. Mapping relationships between the five research axes and representative technology taxonomy nodes

Research Axis	Motion	Auxiliary Control Network	Input Level	Attention Based	Input or Latent Initialization	Condition Fusion	Temporal Consistency	Condition Routing or Weighting
AXIS _01	●	●	●	●		●		
AXIS _02		●	●	●	●	●	●	●
AXIS _03		●	●	●	●	●		
AXIS _04		●	●	●	●	●	●	
AXIS _05	●	●				●		●

Ⅵ. 결 론 (Conclusion)

본 연구는 확산 모델 기반의 제어 가능한 T2V 분야의 연구 문제와 핵심 기술 구조를 분석하기 위해 Literature Analysis Agent, Research Axis Agent, Technology Taxonomy Agent 및 Visualization Agent로 구성된 에이전트 기반 통합 문헌 분석 프레임워크를 제안하였다. 각 에이전트는 사전에 정의된 역할과 프롬프트에 따라 문헌 수집 및 선별, 정보 추출, 연구 방향 도출, 기술 분류 및 결과 시각화를 순차적으로 수행하며, 선행 에이전트의 출력은 후속 에이전트의 입력으로 활용되었다. Literature Analysis Agent는 OpenAlex에서 2020년 1월부터 2026년 7월까지 수집한 1,308편을 단계적으로 선별하여 93편을 후속 분석 대상으로 선정하였다. Research Axis Agent는 연구 문제와 연구 기여가 모두 검증된 80편에 텍스트 임베딩과 K-means 군집화를 적용하여 다섯 개의 군집을 도출하고, 각 군집의 연구 문제와 기여를 연구자가 비교 및 해석하여 다섯 개의 Research Axis로 명명하였다. Technology Taxonomy Agent는 기술 정보가 확인된 문헌을 제어 대상, 조건 표현, 조건 주입 방식, 움직임 및 시간 처리, 복수 제어의 조정, 학습 및 추론 방식의 여섯 관점에서 분석하여 86편에서 515건의 최종 기술 매핑을 확정하였다. Visualization Agent는 두 분석 결과를 모두 보유한 73편과 이에 연결된 441건의 기술 매핑을 통합하여 연구 방향과 기술의 기능적 역할 간 관계를 분석하였다. 분석 결과, 명시적 움직임 및 궤적 제어, 구조적 시간 및 조건 전달, 카메라와 기하 및 3차원 시점 제어가 주요 연구 방향으로 확인되었다. 하나의 연구 문제를 해결하기 위해 여러 기술 요소가 결합되고, 동일한 기술 기능이 서로 다른 Research Axis에서 활용되는 다대다 관계가 나타났다. 이는 해당 분야를 제어 조건이나 모델 구조만으로 분류하기보다, 각 연구가 해결하려는 문제와 기술이 생성 과정에서 수행하는 기능적 역할을 함께 분석해야 함을 보여준다. 본 연구의 의의는 Research Axis와 Technology Taxonomy를 논문 식별자를 기준으로 연결하여 연구 문제와 기술 구조의 관계를 분석하였다는 데 있다. 문헌 선별, 정보 추출, 연구 방향 도출, 기술 분류, 원문 근거 검토 및 결과 통합을 네 개 에이전트의 순차적 처리 절차로 구성함으로써, 지속적으로 증가하는 문헌을 일관된 기준으로 분석하고 그 결과를 추적할 수 있는 방법론을 제시하였다. 다만 본 연구는 OpenAlex만을 문헌 수집 데이터베이스로 사용하였으므로 다른 데이터베이스에 수록된 일부 연구가 누락되었을 가능성이 있다. 2026년 자료는 7월 29일까지 수집된 부분 자료이며, 분석 단계별 검증 조건에 따라 80편, 86편 및 73편의 서로 다른 문헌 집합이 사용되었다. 문헌 선별에 사용한 이중 인코더 및 교차 인코더 임계값은 정량적 검증 절차가 아닌 예비 검토에 기반한 경험적 기준이므로, 향후 연구에서는 사람이 직접 라벨링한 표본을 통한 정량적 검증이 필요하다. 따라서 각 결과는 해당 검증 조건을 충족한 문헌의 범위에서 해석해야 한다. 향후 연구에서는 Scopus, Web of Science 및 IEEE Xplore 등 복수 데이터베이스를 활용하고, 다양한 LLM, 임베딩 모델 및 군집 알고리즘을 비교하여 분석 결과의 안정성을 검증할 필요가 있다. 이와 함께 소규모 표본에 대해서라도 도메인 전문가와 에이전트 간 분류 일치도(inter-coder reliability)를 정량적으로 검증하고 복수 연구자의 독립적 분류와 연구자 간 일치도 평가를 수행하며 에이전트별 프롬프트와 출력 스키마 및 검증 기준을 구체적으로 공개하여 재현성을 높여야 한다. 나아가 신규 문헌과 제어 기술을 지속적으로 반영하여 Research Axis와 Technology Taxonomy를 갱신하는 동적 에이전트 기반 문헌 분석 파이프라인으로 확장할 수 있다.

참 고 문 헌 (References)

[1] A. Jhala and R. M. Young, “Cinematic Visual Discourse: Representation, Generation, and Evaluation,” *IEEE Transactions on Computational Intelligence and AI in Games*, Vol.2, No.2, pp.69-81, June 2010.
doi: https://doi.org/10.1109/TCIAIG.2010.2046486

[2] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, “A Survey on Video Diffusion Models,” *ACM Computing Surveys,* Vol.57, No.2, Article No.41, pp.1-42, February 2025.
doi: https://doi.org/10.1145/3696415

[3] D. Kim, D. Joo, and J. Kim, “TiVGAN: Text to Image to Video Generation with Step-by-Step Evolutionary Generator,” *IEEE Access*, Vol.8, pp.153113-153122, August 2020.
doi: https://doi.org/10.1109/ACCESS.2020.3017881

[4] F. Xie, D. Zeng, Q. Shen, and B. Tang, “A Comprehensive Survey on Text-to-Video Generation,” *Chinese Journal of Electronics*, Vol.34, No.4, pp.1009-1036, July 2025.

doi: https://doi.org/10.23919/CJE.2024.00.151

[5] M. T. Jan, M. G. Al-Jassani, M. Nadar, E. M. Vunnava, V. Chakrapani, H. Ullah, A. Khan, S. A. Abbas, and B. Furht, "Text-to-Video Generators: A Comprehensive Survey," *Journal of Big Data*, Vol.12, No.1, Article No.253, November 2025.
doi: https://doi.org/10.1186/s40537-025-01314-3

[6] R. Jiang, G.-C. Zheng, T. Li, T.-R. Yang, J.-D. Wang, and X. Li, "A Survey of Multimodal Controllable Diffusion Models," *Journal of Computer Science and Technology*, Vol.39, No.3, pp.509-541, May 2024.
doi: https://doi.org/10.1007/s11390-024-3814-0

[7] W. Ma, X. Yang, L. Jiao, L. Li, X. Liu, F. Liu, P. Chen, Y. Yang, M. Ma, L. Sun, R. Zhang, X. Geng, Y. Guo, S. Yang, and Z. Feng, "Video Diffusion Generation: Comprehensive Review and Open Problems," *Artificial Intelligence Review*, Vol.58, Article No.338, August 2025.
doi: https://doi.org/10.1007/s10462-025-11331-6

[8] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-A-Video: Text-to-Video Generation without Text-Video Data," *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, May 2023, https://iclr.cc/virtual/2023/poster/11001 (accessed September 1, 2026).

[9] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, "Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation," *ACM Transactions on Graphics*, Vol.39, No.4, Article No.75, pp.1-13, July 2020.
doi: https://doi.org/10.1145/3386569.3392457

[10] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, Y. Shan, and T.-T. Wong, "Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance," *IEEE Transactions on Visualization and Computer Graphics*, Vol.31, No.2, pp.1526-1541, February 2025.
doi: https://doi.org/10.1109/TVCG.2024.3365804

[11] H. Zhao, J. Gu, S. Wang, T. Lu, X. Zhang, Z. Wu, H. Xu, and Y.-G. Jiang, "LSTD: Long Short-Term Temporal Diffusion for Video Generation," *IEEE Transactions on Multimedia,* Vol.28, pp.2460-2473, 2026.
doi: https://doi.org/10.1109/TMM.2026.3651052

[12] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A Survey on Generative Diffusion Models," *IEEE Transactions on Knowledge and Data Engineering*, Vol.36, No.7, pp.2814-2830, July 2024.
doi: https://doi.org/10.1109/TKDE.2024.3361474

[13] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video Diffusion Models," *Advances in Neural Information Processing Systems*, Vol.35, pp.8633-8646, December 2022, https://papers.nips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html (accessed August 3, 2026).

[14] Z. Yin, K. Chen, X. Bai, R. Jiang, J. Li, H.-D. Li, J. Liu, Y. Xiang, J. Yu, and M. Zhang, "A Survey: Spatiotemporal Consistency in Video Generation," *ACM Computing Surveys*, Vol.58, No.12, Article No.315, pp.1-41, May 2026.
doi: https://doi.org/10.1145/3802588

[15] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp.22563-22575, June 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.02161

[16] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.45, No.9, pp.10850-10869, September 2023.
doi: https://doi.org/10.1109/TPAMI.2023.3261988

[17] F. Waseem and M. Shahzad, "Video Is Worth a Thousand Images: Exploring the Latest Trends in Long Video Generation," *ACM Computing Surveys*, Vol.58, No.6, Article No.154, April 2026.
doi: https://doi.org/10.1145/3771724

[18] D. J. Zhang, D. Li, H. Le, M. Z. Shou, C. Xiong, and D. Sahoo, "MoonShot: Towards Controllable Video Generation and Editing with Motion-Aware Multimodal Conditions," *International Journal of Computer Vision*, Vol.133, No.6, pp.3629-3644, June 2025.
doi: https://doi.org/10.1007/s11263-025-02346-1

[19] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield, "Can Large Language Models Replace Humans in Systematic Reviews? Evaluating GPT-4's Efficacy in Screening and Extracting Data from Peer-Reviewed and Grey Literature in Multiple Languages," *Research Synthesis Methods*, Vol.15, No.4, pp.616-626, July 2024.
doi: https://doi.org/10.1002/jrsm.1715

[20] E. Syriani, I. David, and G. Kumar, "Screening Articles for Systematic Reviews with ChatGPT," *Journal of Computer Languages*, Vol.80, Article No.101287, August 2024.
doi: https://doi.org/10.1016/j.cola.2024.101287

[21] O. Kohandel Gargari, M. H. Mahmoudi, M. Hajisafarali, and R. Samiee, "Enhancing Title and Abstract Screening for Systematic Reviews with GPT-3.5 Turbo," *BMJ Evidence-Based Medicine*, Vol.29, No.1, pp.69-70, February 2024.
doi: https://doi.org/10.1136/bmjebm-2023-112678

[22] E. Guo, M. Gupta, J. Deng, Y.-J. Park, M. Paget, and C. Naugler, "Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study," *Journal of Medical Internet Research*, Vol.26, Article No.e48996, January 2024.
doi: https://doi.org/10.2196/48996

[23] R. C. Nickerson, U. Varshney, and J. Muntermann, "A Method for Taxonomy Development and Its Application in Information Systems," *European Journal of Information Systems*, Vol.22, No.3, pp.336-359, May 2013.
doi: https://doi.org/10.1057/ejis.2012.26

[24] D. Kundisch, J. Muntermann, A. M. Oberländer, D. Rau, M. Röglinger, T. Schoormann, and D. Szopinski, "An Update for Taxonomy Designers: Methodological Guidance from Information Systems Research," *Business & Information Systems Engineering*, Vol.64, No.4, pp.421-439, August 2022.

doi: https://doi.org/10.1007/s12599-021-00723-x

[25] R. van Dinter, B. Tekinerdogan, and C. Catal, “Automation of Systematic Literature Reviews: A Systematic Literature Review,” *Information and Software Technology*, Vol.136, Article No.106589, August 2021.
doi: https://doi.org/10.1016/j.infsof.2021.106589

[26] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting Hallucinations in Large Language Models Using Semantic Entropy,” *Nature*, Vol.630, pp.625-630, June 2024.
doi: https://doi.org/10.1038/s41586-024-07421-0

[27] R. van de Schoot, J. de Bruin, R. Schram, P. Zahedi, J. de Boer, F. Weijdema, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands, A. Harkema, J. Willemsen, Y. Ma, Q. Fang, S. Hindriks, L. Tummers, H. van den Akker, and D. L. Oberski, “An Open Source Machine Learning Framework for Efficient and Transparent Systematic Reviews,” *Nature Machine Intelligence*, Vol.3, No.2, pp.125-133, February 2021.
doi: https://doi.org/10.1038/s42256-020-00287-7

[28] I. Singh, C. Scarton, and K. Bontcheva, “Multistage BiCross Encoder for Multilingual Access to COVID-19 Health Information,” *PLOS ONE*, Vol.16, No.9, Article No.e0256874, September 2021.
doi: https://doi.org/10.1371/journal.pone.0256874

[29] W. X. Zhao, J. Liu, R. Ren, and J.-R. Wen, “Dense Text Retrieval Based on Pretrained Language Models: A Survey,” *ACM Transactions on Information Systems*, Vol.42, No.4, Article No.89, pp.1-60, February 2024.
doi: https://doi.org/10.1145/3637870

[30] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol.28, No.1, pp.100-108, 1979.
doi: https://doi.org/10.2307/2346830

[31] P. J. Rousseeuw, “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,” *Journal of Computational and Applied Mathematics*, Vol.20, pp.53-65, November 1987.
doi: https://doi.org/10.1016/0377-0427(87)90125-7

[32] A. Singer, N. Rotstein, A. Mann, R. Kimmel, and O. Litany, “Time-to-Move: Training-Free Motion-Controlled Video Generation via Dual-Clock Denoising,” *Proceedings of the International Conference on Learning Representations*, Singapore, 2026, https://proceedings.iclr.cc/paper_files/paper/2026/hash/1eaa5146756be028ad6fff1efcc8e6bd-Abstract-Conference.html (accessed September 1, 2026).

[33] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, “ControlVideo: Training-Free Controllable Text-to-Video Generation,” *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, 2024, https://proceedings.iclr.cc/paper_files/paper/2024/hash/ef74413c7bf1d915c3e45c72e19a5d32-Abstract-Conference.html (acessed September 1, 2026).

[34] S. Bahmani, I. Skorokhodov, A. Siarohin, W. Menapace, G. Qian, M. Vasilkovsky, H.-Y. Lee, C. Wang, J. Zou, A. Tagliasacchi, D. B. Lindell, and S. Tulyakov, “VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control,” *Proceedings of the International Conference on Learning Representations*, Singapore, 2025, https://proceedings.iclr.cc/paper_files/paper/2025/hash/a6ef26da2f5b7a2d5d374dd13464c781-Abstract-Conference.html (acessed September 1, 2026).

[35] Z. Ma, D. Zhou, X.-S. Wang, C.-H. Yeh, X. Li, H. Yang, Z. Dong, K. Keutzer, and J. Feng, “Magic-Me: Identity-Specific Video Customized Diffusion,” *Computer Vision –ECCV 2024 Workshops, Lecture Notes in Computer Science*, Vol.15627, pp.19-37, Springer, Cham, Switzerland, 2025.
doi: https://doi.org/10.1007/978-3-031-92808-6_2

[36] W. Wang, M. Huang, Y. Tu, and Z. Mao, “DualReal: Adaptive Joint Training for Lossless Identity-Motion Fusion in Video Customization,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Honolulu, HI, USA, pp.16565-16575, 2025.
doi: https://doi.org/10.1109/ICCV51701.2025.01538

저 자 소 개

김 민 진

- 2025년 3월 ~ 현재 : 덕성여자대학교 데이터사이언스학과 학부생
- ORCID : https://orcid.org/0009-0009-2064-6507
- 주관심분야 : 데이터 마이닝, 인공지능, 멀티모달

저 자 소 개

윤 소 율

- 2025년 3월 ~ 현재 : 덕성여자대학교 데이터사이언스학과 학부생
- ORCID : https://orcid.org/0009-0000-5398-5956
- 주관심분야 : 데이터 마이닝, 인공지능, 멀티모달

박 지 예

- 2026년 8월 : 덕성여자대학교 컴퓨터공학전공 학사
- 2026년 9월 ~ 현재 : 덕성여자대학교 응용데이터사이언스학과 석사과정
- ORCID : https://orcid.org/0009-0008-2300-9506
- 주관심분야 : 컴퓨터 비전, LLM, 멀티모달

유 제 혁

- 2020년 8월 : 고려대학교 전기전자컴퓨터공학과 박사
- 2023년 2월 : ㈜기아 책임매니저
- 2023년 3월 ~ 현재 : 덕성여자대학교 데이터사이언스학과 조교수
- ORCID : https://orcid.org/0000-0003-3952-7582
- 주관심분야 : 빅데이터, 패턴인식, 인공지능