효율적인 옴니모달 월드 모델을 위한 핵심 기술과 연구 동향

ㅁ 홍성은 / 성균관대학교

요 약

생성형 인공지능은 텍스트와 이미지를 함께 다루던 단계에서 영상, 음성, 언어, 행동 신호를 하나의 시간 흐름 안에서 처리하는 옴니모달 시스템으로 옮겨 가고 있다. 차세대 미디어 생성 역시 한 장면을 잘 합성하는 데서 그치지 않고, 환경의 상태를 추정하고 행동에 따른 미래를 예측하며 그 결과를 계속 갱신하는 월드 모델 쪽으로 나아가고 있다. 문제는 두 흐름이 결합될 때 고해상도 시공간 토큰, 장기 기억, 다중 미래 롤아웃이 한꺼번에 더해지면서 계산량과 메모리 이동이 크게 늘어난다는 점이다. 본고에서는 효율적인 옴니모달 월드 모델을 위한 기술을 문맥 압축, 모델 압축, 캐시와 기억, 적응적 연산, 시스템과 서빙의 다섯 계층으로 나누어 정리한다. 이어 본 연구진이 제안한 영차 민감도 기반 토큰 가지치기 기법 ZOO-Prune을 포함한 최근 연구를 살펴보고, 고정된 압축률 대신 선택적 지각·선택적 기억·선택적 롤아웃을 함께 설계하는 방향을 논의한다.

Ⅰ. 서 론

최근 멀티모달 인공지능은 서로 다른 입력을 덧붙여 처리하던 방식에서 점차 벗어나고 있다. 카메라 영상, 음성, 텍스트, 사용자 행위가 시간축을 따라 함께 들어오고, 모델은 그에 맞춰 언어나 음성, 시각적 결과를 실시간으로 내놓는다. Qwen2.5-Omni는 영상·음성·문자를 통합적으로 인식하면서 텍스트와 음성을 스트리밍 방식으로 생성한다[1]. 여러 모달리티를 입력으로 받는 데 더해 입력과 출력의 형태를 상황에 따라 바꿀 수 있다는 점에서, 이러한 모델을 흔히 옴니모달 AI라고 부른다.

월드 모델은 관측으로부터 환경의 잠재 상태를 구성하고, 현재 상태와 행동을 조건으로 미래 상태를 예측하는 모델이다[2]. 초기 연구가 압축된 상태 표현과 간단한 동역학 모델의 학습에 집중했다면, 최근에는 픽셀 또는 잠재 공간에서 상호작용 가능한 환경을 생성하고 범용 제어까지 수행하는 연구로 범위가 넓어졌다[3-5]. 이러한 관점에서 보면 차세대 미디어 생성은 프롬프트 한 번으로 영상을 완성하는 작업에 한정되지 않는다. 사용자의 행동과 대화, 센서 입력을 받아 장면을 계속 갱신하는 상호작용형 생성

이 점점 중요한 과제가 되고 있다.

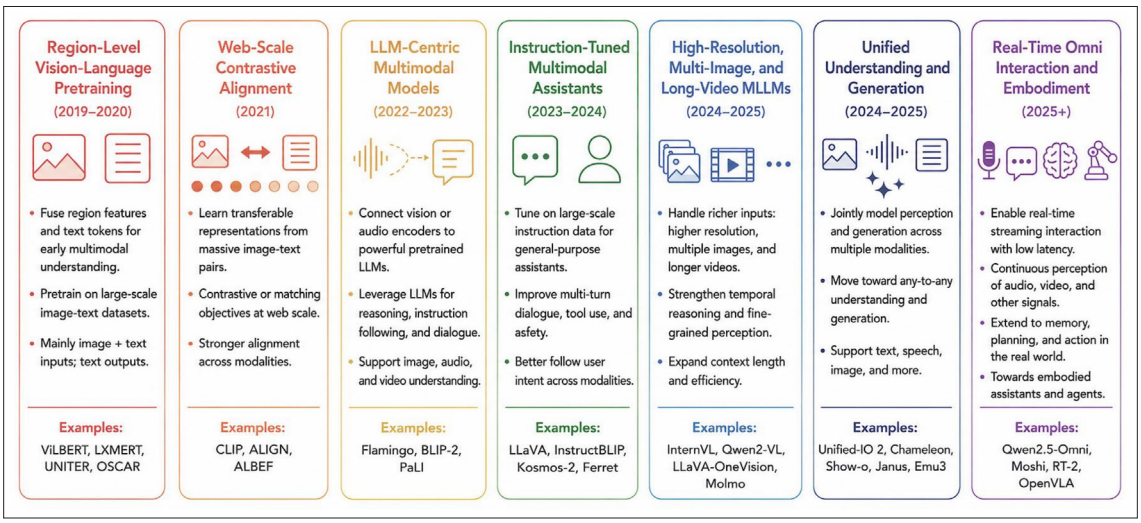
다만 옴니모달 입력과 월드 모델의 반복적 예측을 결합하면 계산 부담이 빠르게 커진다. 영상 토큰은 공간 해상도와 시간 길이에 비례해 늘어나고, 음성은 높은 샘플링 빈도 때문에 긴 토큰열을 만든다. 장기 상호작용에서는 과거 상태를 저장해 두어야 하고, 계획을 세우려면 여러 미래를 반복해서 전개해야 한다. 그래서 이 분야의 경량화는 파라미터 수를 줄이는 문제로만 보기 어렵다. 현재 장면을 얼마나 자세히 볼지, 무엇을 기억할지, 어떤 미래를 어느 깊이까지 전개할지를 상황에 맞게 배분하는 문제로 접근할 필요가 있다. 본고는 이러한 관점에서 관련 기술과 최근 연구 동향을 정리하고, 본 연구진이 진행해 온 토큰 압축 연구를 함께 소개한다.

Ⅱ. 옴니모달 AI와 월드 모델의 접점

1. 멀티모달에서 옴니모달로

멀티모달 모델은 시각 인코더와 언어 모델을 연결하는 구조를 중심으로 발전해 왔다. 초기에는 모달리티별 인코더가 특징을 추출하고, 정렬 모듈이 이를 공통 표현으로 바꾸는 방식이 주를 이루었다. 이후 Chameleon은 텍스트와 이미지를 하나의 혼합 토큰열로 모델링했고[6], Show-o는 자기회귀 생성과 확산 기반 생성을 하나의 트랜스포머에서 다루었다[7]. NExT-GPT는 입력과 출력을 모두 여러 모달리티로 확장했다[8]. 최근의 옴니모달 모델에는 여기에 실시간성이 더해진다. 입력이 모두 끝난 뒤에 답하는 것이 아니라, 영상과 음성이 계속 들어오는 동안 이해와 생성을 병행해야 하기 때문이다.

옴니모달이라고 해서 모든 모달리티를 같은 방식으로 처리해도 된다는 뜻은 아니다. 영상은 공간적 중복이 크고, 음성은 짧은 시간 동안의 변화가 중요하며, 텍스트는 토큰 수가 적어도 의미 밀도가 높다. 동일한 압축률이나 정밀도를 일괄 적용하면 중요한 모달리티의 정보가 손실되거나, 반대로 덜 중요한 신호에 계산 자원이 낭비되기 쉽다. 따라서 공통 표현을 만들더라도 각 모달리티의 시간 구조와 정보 밀도는 따로 고려해야 한다. 〈그림 1〉은 이러한 발전 흐름이 실시간 옴니모달 상호작용과 체화형 에이전트로 이어지는 과정을 정리한 것이다.



<그림 1> 멀티모달 AI에서 실시간 옴니모달 상호작용으로의 발전 흐름

2. 미디어 생성 모델에서 예측 가능한 세계로

월드 모델은 세 가지 구성 요소로 설명할 수 있다. 먼저 관측 $o_{\leq t}$를 압축해 현재의 잠재 상태 z_t를 만들고, 행동 a_t가 주어지면 다음 상태 z_{t+1}을 예측한다. 그리고 잠재 상태를 관측이나 보상, 생성 결과로 다시 복원한다. 이 구조는 미디어 생성 모델과도 잘 맞아떨어진다. 장면 속 인물, 물체, 카메라, 음향 사건을 상태로 유지하면서 사용자 입력이나 에이전트의 행동에 따라 다음 영상과 음성을 만들어 내는 과정으로 볼 수 있기 때문이다. 본고에서는 시청각·언어·행동 입력으로 상태를 갱신하고, 행동에 따른 미래를 예측·생성하는 시스템을 옴니모달 월드 모델로 본다.

Google DeepMind의 Genie는 비디오로부터 잠재 행동과 동역학을 학습해 상호작용 가능한 환경을 생성할 수 있음을 보였다[4]. V-JEPA 2는 픽셀 전체를 복원하지 않고 잠재 표현 공간에서 미래를 예측하며, 행동 조건 모델을 더해 계획 문제로 확장했다[5]. 이런 접근에서는 생성 품질 못지않게 상태의 일관성과 행동에 대한 반응성이 중요하다. 머리카락 질감이나 배경의 미세한 픽셀은 프레임마다 조금씩 달라져도 크게 문제 되지 않지만, 인물의 정체성, 물체의 위치, 사건의 선후 관계가 흐트러지면 장면의 연속성이 깨진다. 따라서 효율화의 목표도 시각적으로 덜 두드러지는 픽셀보다는 미래를 예측하는 데 덜 필요한 상태를 찾아 줄이는 데 있다.

Ⅲ. 옴니모달 월드 모델의 계산 병목

1. 시공간 토큰의 폭증

고해상도 이미지 한 장은 수백에서 수천 개의 시각 토큰으로 변환되고, 비디오는 여기에 프레임 수가 곱해진다. 토큰열이 길어지면 어텐션 연산량뿐 아니라 언어 모델의 프리필(prefill) 비용과 KV 캐시 크기도 함께 커진다. 월드 모델이 미래 프레임이나 잠재 상태를 여러 단계에 걸쳐 생성하면 같은 병목이 시간축을 따라 반복된다. 토큰 수를 줄이는 것이 가장 직접적인 해법이지만, 빠른 동작이나 작은 객체처럼 드물게 나타나는 정보가 먼저 제거될 위험이 있다.

2. 모달리티 비대칭성과 교차모달 정렬

중복의 형태는 모달리티마다 다르다. 영상에서는 인접 패치가 서로 비슷하고 연속 프레임의 배경이 반복된다. 음성에서는 발화 구간과 침묵 구간의 중요도가 크게 다르며, 입 모양이나 사건이 시작되는 시점과도 맞물려 있어야 한다. 텍스트 질의는 영상과 음성의 어느 구간이 중요한지를 정하는 기준이 된다. 각 스트림을 따로 압축하면 이러한 대응 관계가 끊어질 수 있다. 그래서 옴니모달 압축에서는 각 모달리티에서 무엇을 줄일지뿐 아니라, 모달리티 사이의 어떤 연결을 남길지도 결정해야 한다.

3. 지속 기억과 반복적 롤아웃

스트리밍 환경에서 과거 토큰을 모두 저장하는 것은 현실적으로 불가능하다. 그렇다고 최근 장면만 사용하면 오래전에 등장한 인물이나 목표를 놓치게 된다. 반대로 긴 기억을 무차별적으로 덧붙이면 현재 관측에 대한 주의가 분산되고 지연도 늘어날 수 있다. 계획 단계에서도 비슷한 문제가 생긴다. 모든 행동 후보를 같은 길이와 해상도로 전개하면 비용이 감당하기 어려울 만큼 커진다. 따라서 처음에는 여러 미래를 거칠게 탐색하고, 가능성이 높은 분기만 골라 자세히 생성하는 전략이 필요하다.

4. FLOPs 너머의 메모리 이동

실제 지연 시간과 에너지 소모는 부동소수점 연산량(FLOPs)만으로는 설명되지 않는다. 모델 가중치와 활성값, KV 캐시를 메모리 계층 사이에서 옮기는 비용이 상당

히 크기 때문이다[9]. 토큰 수를 줄였더라도 불규칙한 인덱싱이나 작은 연산이 많아지면 GPU 활용률이 오히려 떨어질 수 있다. 반면 FlashAttention과 같은 IO 인지형 커널[10]이나 PagedAttention 기반의 캐시 관리[11]는 같은 모델을 훨씬 효율적으로 실행하게 해 준다. 따라서 효율화 연구에서는 알고리즘 차원의 압축률과 함께 실제 시스템에서 측정한 종단 간 지연을 보고하는 것이 바람직하다.

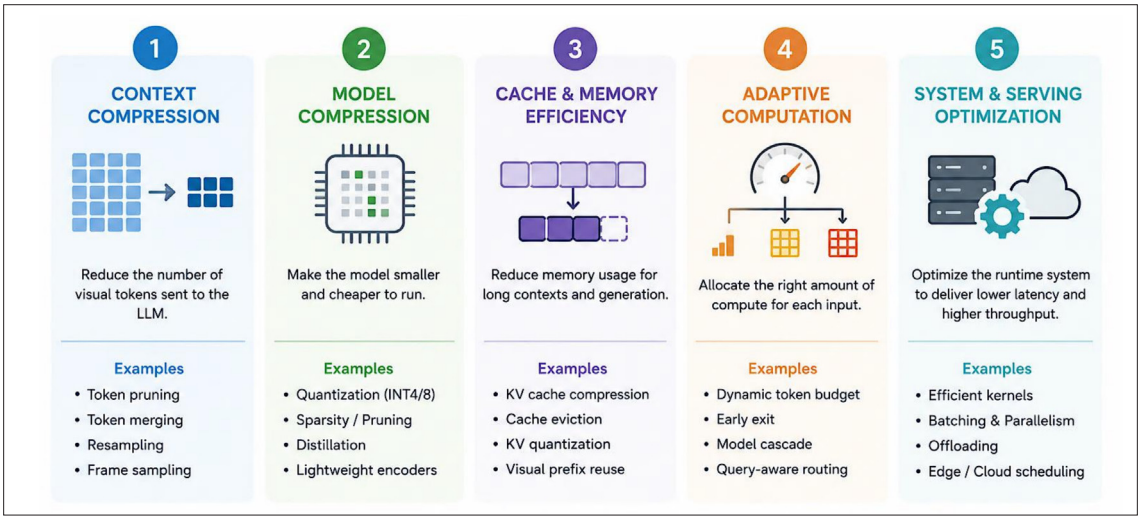
Ⅳ. 효율화를 위한 다섯 계층

효율화 기법을 하나의 기준으로 비교하기는 어렵다. 입력 문맥을 줄이는 방법, 모델 자체를 가볍게 만드는 방법, 과거 상태를 관리하는 방법, 입력에 따라 연산량을 바꾸는 방법, 하드웨어에서 실행을 최적화하는 방법은 각기 다른 병목을 겨냥하기 때문이다. 본고에서는 이를 〈그림 2〉와 같이 다섯 계층으로 나누어 정리한다. 실제 시스템에서는 한 계층만 최적화하기보다 여러 계층을 함께 설계해야 효과가 누적된다.

1. 문맥 압축

문맥 압축은 긴 입력에서 중복되는 토큰을 제거하거나 병합하는 접근이다. Token Merging(ToMe)은 유사한 토큰을 병합해 비전 트랜스포머의 연산량을 줄였고[12], FastV는 두 번째 층 이후로는 시각 토큰에 대한 어텐션 연산이 매우 비효율적이라는 관찰을 바탕으로 이후 층에서 시각 토큰을 가지치기했다[13]. VisionZip은 정보량이 많은 시각 토큰만 골라 언어 모델에 전달하여 프리필 시간을 최대 8배 단축했다고 보고했다[14]. PyramidDrop은 층이 깊어질수록 시각 정보의 중복이 커진다는 점에 착안해 토큰을 단계적으로 줄인다[15]. 전체적으로 보면 고정된 위치에서 일정 비율을 줄이던 방식이 입력과 층에 따라 축소 정도를 달리하는 방식으로 바뀌고 있다.

월드 모델에서는 토큰의 현재 중요도만으로 판단하기 어렵다. 지금은 배경처럼 보이는 물체가 다음 행동에서는 장애물이 될 수 있고, 짧게 스쳐 간 음향 사건이 이후 장면을 설명하는 원인이 되기도 한다. 따라서 토큰의 가치는 현재 출력에 대한 기여에 더해 미래 상태 전이에 대한 기여까지 반영해 평가해야 한다. 유사도 기반 병합은



<그림 2> 효율적인 VLM과 옴니모달 모델을 위한 다섯 계층

시각적 중복을 줄이는 데는 효과적이지만, 드물게 나타나면서도 인과적으로 중요한 사건을 보호하는 장치가 별도로 필요하다.

2. 모델 압축

양자화는 가중치와 활성값의 비트 수를 낮추고, 구조적 가지치기는 채널이나 블록 단위로 모델을 줄이며, 지식 증류는 큰 교사 모델의 능력을 작은 학생 모델로 옮긴다. 옴니모달 모델에서는 모달리티마다 민감도가 다르다는 점을 고려해야 한다. MBQ는 시각 토큰과 언어 토큰의 양자화 민감도 차이를 보정하여 저비트 설정에서의 정확도를 높였고, 전용 커널을 적용해 1.4배의 가속을 보고했다[16]. 월드 모델에서는 여기에 시간축의 문제가 더해진다. 인코더, 동역학 모델, 디코더는 오류가 쌓이는 양상이 서로 다르고, 반복 롤아웃에서는 작은 양자화 오차도 이후 상태 예측으로 번질 수 있다. 그래서 단일 프레임의 재구성 품질과 더불어 행동별 장기 예측 오차를 확인하고, 민감한 상태 갱신 경로에 높은 정밀도를 우선 배정하는 혼합 정밀도 설계가 적합하다.

3. 캐시와 기억

KV 캐시는 과거 토큰을 다시 계산하지 않도록 해 주지만, 문맥 길이에 비례해 메모리를 차지한다. StreamingLLM은 초기 토큰에 형성되는 어텐션 싱크(attention sink)와 최근 문맥만 유지하는 방식으로 긴 스트림을 처리했다[17]. 옴니모달 월드 모델에서는 이처럼 텍스트 중심의 캐시 관리만으로는 부족하다. 최근 장면의 고해상도 표현, 오래된 사건의 요약, 인물과 물체의 지속적인 상태를 서로 다른 시간 척도로 관리해야 한다. 이상적인 기억은 모든 프레임을 그대로 저장한 사본보다는, 미래 예측과 질의응답에 다시 활용할 수 있는 상태 표현에 가깝다. 월드 모델이라면 행동 이후에도 유지되어야 할 객체 상태와 사건을 골라 남기는 것이 관건이다. 캐시를 줄일 때도 현재 질의에 대한 정답 여부뿐 아니라 나중의 상태 예측에 쓰일 단서가 남아 있는지가 평가 기준이 된다.

4. 적응적 연산

고정된 토큰 예산을 쓰면 쉬운 장면에서는 계산이 낭비되고 어려운 장면은 지나치게 압축된다. 적응적 연산은 질의, 장면 복잡도, 불확실도, 목표 지연 시간에 따라 토큰 수와 네트워크 깊이를 조절하는 방식이다. 최근 공개된 RUTA는 어떤 토큰을 남길지와 몇 개를 남길지를 rate-utility 목적함수로 함께 학습한다[18]. 같은 아이디어를 월드 모델의 롤아웃에도 적용할 수 있다. 미래 분기의 불확실도가 높거나 안전과 관련된 장면에는 더 긴 전개를 배정하고, 변화가 적은 안정된 구간에는 짧은 전개만 수행하는 식이다. 다만 예측 불확실도가 높다는 이유만으로 무조건 연산을 늘리면 지연 한계를 넘을 수 있다. 결국 월드 모델에서는 분기별 가치와 위험을 따져 지각, 기억 검색, 롤아웃 깊이 사이에 연산 예산을 나누는 문제가 된다.

5. 시스템과 서빙

효율화 기법은 결국 실제 실행 환경에서 검증되어야 한다. 입력 스트림을 비동기적으로 인코딩하고, 서로 다른 모달리티의 토큰을 시간축에 맞춰 정렬하며, 프리필과 디코딩을 중첩해 실행해야 한다. 서버에서는 요청마다 길이가 달라 배칭 효율이 떨어지고, 엣지 장치에서는 메모리와 전력 제약이 훨씬 엄격하다. 페이지 단위의 KV 캐시 관리, 연산자 융합, 저비트 커널, 엣지-클라우드 분할과 같은 기법이 뒷받침되어야 알고리즘 수준의 압축이 실제 처리량과 응답 속도 개선으로 이어진다. 특히 월드 모델의 지각·상태 갱신·미래 생성은 서로 다른 빈도로 실행될 수 있다. 각 경로의 지연을 따로 측정하고 동기화 비용까지 포함해야 실시간 상호작용에 필요한 처리량을 판단할 수 있다.

Ⅴ. 최근 연구 동향과 대표 사례

본 절의 사례 대부분은 VLM이나 옴니모달 이해 모델에서 검증된 연구이며, 〈표 1〉은 월드 모델 관점에서 기존 사례 결과의 확장 방향을 제시한다.

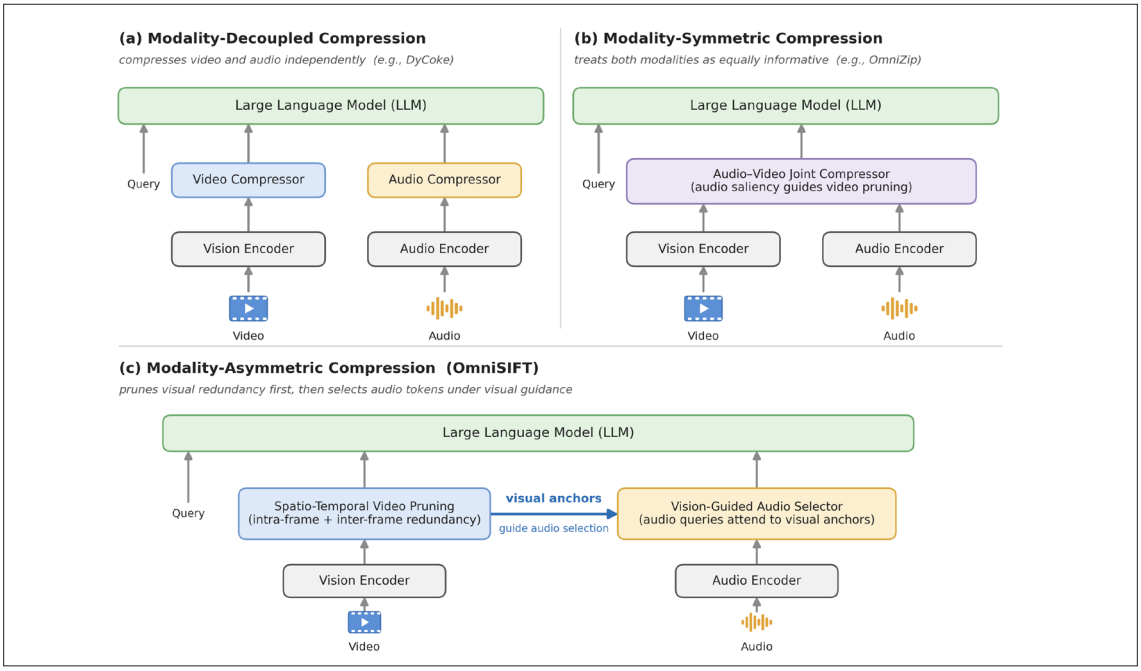
1. 이미지·비디오 토큰 압축의 진화

시각 토큰 압축은 유사 토큰의 병합, 중요도에 따른 선택, 두 방식의 결합으로 나뉜다. DivPrune은 선택 토큰의 다양성을 높여 중복을 줄였고[19], ATP-LLaVA는 입력과 LLM 계층에 따라 가지치기 비율을 조절했다[20]. 비디오용 DyCoke는 프레임 간 유사 토큰을 병합하고 디코딩 중 중요한 KV 토큰을 유지한다[21]. 이 가운데 ATP-LLaVA와 DyCoke는 고정 비율 대신 입력과 처리 단계에 따라 토큰 예산을 달리 배분한다.

2. 모달리티 비대칭 압축

영상과 음성은 중복 양상이 달라 같은 기준으로 압축하기 어렵다. OmniSIFT[22]는 〈그림 3〉처럼 두 모달리티를 독립적으로 줄이는 분리형, 음성 중요도로 영상 선택을 유도하는 대칭형, 영상 앵커를 먼저 만든 뒤 음성을 선택하는 비대칭형을 구분한다. 분리형은 모달리티 간 의미적 연결을 놓칠 수 있다. 대칭형은 두 모달리티를 똑같이 중요하게 취급하는 데다, 어텐션 점수에 의존해 FlashAttention 같은 효율적 연산자와 병용하기 어렵다는 한계가 지적된다.

OmniSIFT의 비대칭형은 영상의 공간·시간 중복을 먼저 줄인 뒤, 음성 토큰이 남은 시각 앵커를 참조하도록 한다. 화면 속 화자나 사건에 따라 음성의 중요도가 달라진다는 관찰을 반영한 설계다. OmniMem은 스트리밍 시청각 모델에서 영상·음성의 기억 예산을 분리하고 교란 민감



<그림 3> 옴니모달 LLM의 토큰 압축 패러다임 비교(OmniSIFT[22]의 Figure 2를 바탕으로 재구성)

도를 이용해 KV 상태를 보존했다[23]. 압축 대상이 입력에서 지속 기억으로 확장되고 있다.

3. 영차 민감도 기반 ZOO-Prune

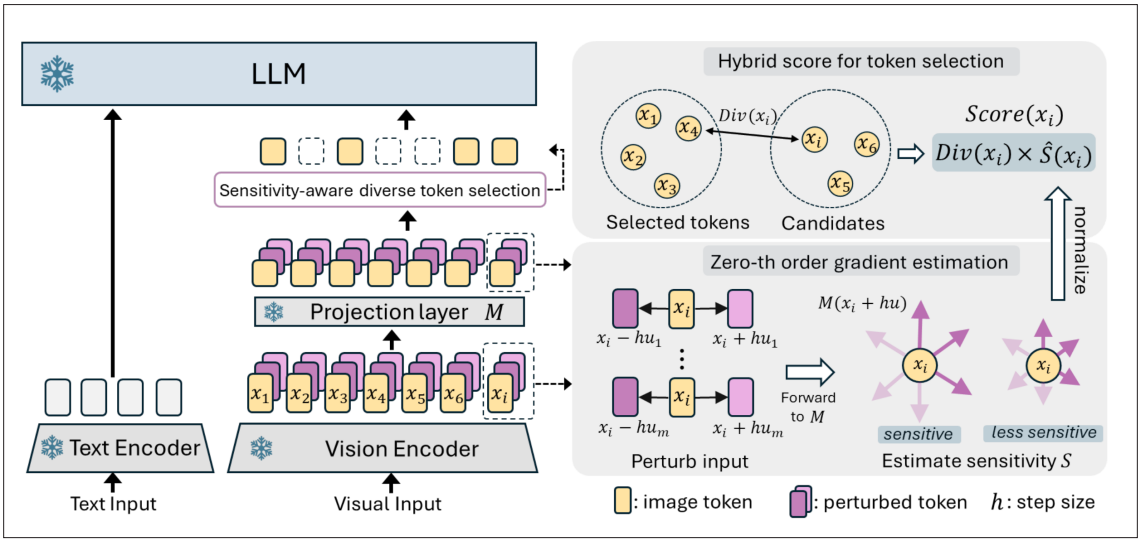
본 연구진이 제안한 ZOO-Prune은 시각 토큰에 작은 교란을 가하고 출력 특징의 변화를 측정해 중요도를 추정한다[24]. 역전파 없이 순전파만 쓰는 영차(zeroth-order) 추정이며, 비전 인코더 대신 가벼운 프로젝션 층에서 민감도를 계산한다. 두 위치에서 얻은 토큰 순위가 일정한 양의 상관을 보인다는 분석을 근거로 삼고, 중요한 토큰 사이의 다양성도 고려한다. 구조는 〈그림 4〉에 제시했다.

ZOO-Prune은 별도 학습 없이 LLaVA-NeXT-7B의 시각 토큰 94.4%를 제거하면서 원래 성능의 95.4%를 유지했고, 종단 간 추론을 최대 2.30배 가속했다. 다만 이는 VLM에서 측정한 결과다. 월드 모델로 옮기려면 한 번의 응답 변화뿐 아니라 여러 단계 뒤의 상태와 행동 결과에 대한 민감도를 검증해야 한다.

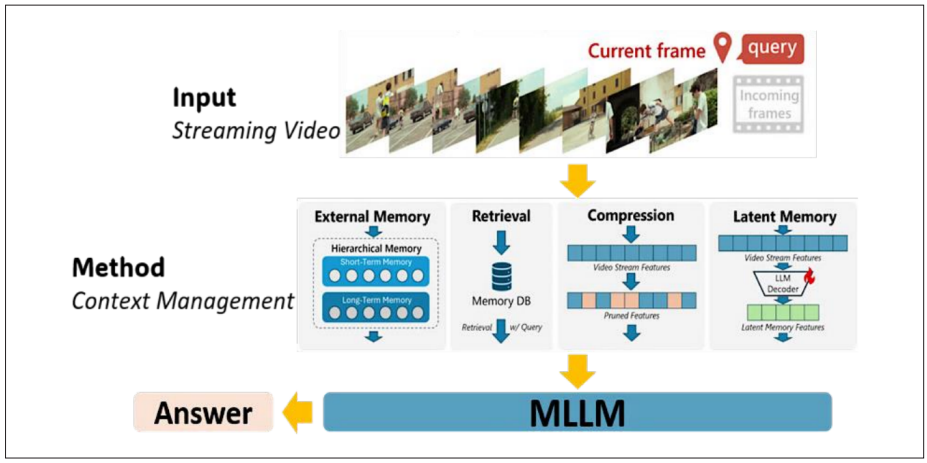
4. 장기 비디오와 스트리밍 기억

장기 비디오 이해에서는 압축과 기억을 떼어 놓고 설계하기 어렵다. LongVU는 질의와 프레임 간 의존성을 이용해 시공간 중복을 줄였다[25]. SimpleStream은 최근 네 프레임만 쓰는 기준선으로 OVO-Bench 67.7%, StreamingBench 80.59%를 보고했다[26]. 같은 연구에서 긴 문맥은 회상에는 도움이 될 수 있지만 현재 장면의 지각은 해칠 수 있었다. 따라서 최근 장면의 지각과 장기 회상은 나누어 평가하는 것이 바람직하다.

생성형 월드 모델에서도 긴 기억이 실제로 장면의 연속성을 높이는지 확인해야 한다. 최근 관측을 처리하는 경로와 과거 사건의 요약 경로를 나누고, 행동 목표에 맞춰 기억을 검색하는 설계가 가능하다. WorldKV는 카메라·행동 대응 관계로 과거 KV 캐시를 다시 불러오고 각 청크의 중복 토큰을 줄여, 추가 학습 없이 전체 KV 캐시를 유지할 때와 비슷하거나 더 나은 장면 일관성을 약 2배의 처리량으로 얻었다[27]. 스트리밍 비디오 모델의 문맥 관리 방식은 〈그림 5〉에, 이 장에서 살펴본 주요 연구와 월드 모델



<그림 4> ZOO-Prune의 영차 민감도 추정 및 다양성 기반 토큰 선택 구조[24]



<그림 5> 스트리밍 비디오 이해의 대표적 문맥 관리 방식(SimpleStream[26]에서 발췌·편집)

<표 1> 최근 대표 효율화 연구와 월드 모델 관점의 해석

연구	핵심 아이디어	월드 모델 관점의 시사점
VisionZip [CVPR 2025]	대표 토큰 선택과 집약	긴 관측의 프리필 단축
ATP-LLaVA [CVPR 2025]	층·입력별 적응형 가지치기	상태별 동적 관측 예산
DyCoke [CVPR 2025]	시간 병합과 디코딩 중 동적 제거	단계별 시공간 예산
MBQ [CVPR 2025]	모달리티 민감도 기반 양자화	구성 요소별 정밀도 배분
LongVU [ICML 2025]	질의·프레임 의존성 기반 시공간 압축	장기 관측의 계층적 요약
ZOO-Prune [CVPR 2026]	영차 민감도·다양성 기반 학습 없는 선택	장기 롤아웃 민감도로 확장
OmniSIFT [ICML 2026]	영상 우선 압축과 시각 유도 음성 선택	교차모달 정렬 보존
OmniMem [arXiv 2026]	모달리티·교란 인식 메모리 압축	시청각 기억 예산 분리
SimpleStream [arXiv 2026]	최근 프레임 기준선	지각·기억 분리 평가
RUTA [arXiv 2026]	rate-utility 기반 동적 토큰 수	상황별 계산 예산 배분

관점의 시사점은 〈표 1〉에 정리했다.

VI. 월드 모델 기반 미디어 생성과 향후 연구 과제

1. 선택적 지각과 조건부 연산

선택적 지각은 목표와 행동에 필요한 관측만 세밀하게 처리한다. 디지털 휴먼은 입 모양과 발화를 높은 해상도로 보되 정적인 배경은 요약할 수 있다. 앞으로는 장면 복잡도, 예측 불확실도, 행동 위험도, 지연·전력 한계를 종합해 입력별 토큰과 연산 예산을 정해야 한다. 평균 비용이 낮더라도 어려운 입력에서 최대 지연과 메모리 사용량이 급증하면 실시간 서비스에는 쓰기 어렵다.

2. 선택적 기억과 장기 상태 보존

선택적 기억은 모든 프레임 대신 객체의 정체성·위치, 사건의 원인과 결과처럼 미래 예측에 필요한 상태를 남긴

다. 각 기억에는 신뢰도와 유효 기간을 두고, 새 관측과 충돌하는 상태는 갱신하거나 폐기한다. 화면 밖 객체와 오래된 사건이 이후 생성에 영향을 줄 수 있으므로, 압축 후에도 객체 지속성과 사건의 선후 관계가 유지되는지 확인할 필요가 있다. 희소 사건 회상과 반사실적 질문은 이런 손실을 드러내는 평가가 될 수 있다.

3. 선택적 롤아웃과 교차모달 정렬

선택적 롤아웃은 여러 행동의 결과를 저비용 잠재 상태에서 먼저 탐색하고, 유망하거나 위험한 분기만 영상·음성으로 자세히 복원한다. 상태 전환이 임박하거나 예측이 불확실한 구간에 연산을 더 쓸 수 있다. 이때 입 모양과 발화, 충돌 장면과 소리처럼 모달리티 사이의 시간적 대응도 유지되어야 한다. 토큰을 각각 잘 줄여도 이 연결이 어긋나면 행동에 대한 생성 결과가 일관되지 않을 수 있다.

4. 적용 시나리오와 엣지 협력

디지털 휴먼은 최근 표정·억양을 자세히 처리하면서 이전 대화는 사건 기억으로 요약하고, 선택한 응답만 고품질 얼굴 영상과 음성으로 복원할 수 있다. 인터랙티브 영상은 화면 밖 객체와 서사를 기억하되 사용자의 선택에 따라 필요한 미래 분기만 전개한다. 로봇과 원격 현장에서는 장치가 위험 사건을 빠르게 감지하고 클라우드가 장기 계획을 맡는 분할도 가능하다. 이 경우 네트워크 지연과 전송량까지 예산에 포함하고, 중요한 사건의 원본이나 고해상도 특징을 선택적으로 전달해야 한다.

5. 생성 품질과 효율의 공동 평가

생성 결과는 이미지 품질뿐 아니라 인물·객체의 시간적 일관성, 행동에 따른 인과적 타당성, 최근 장면의 지각과 장기 사건의 회상을 나누어 평가해야 한다. 효율 지표에는 첫 응답과 프레임당 지연, 최대 메모리, 에너지, 통신량, 추가 학습 비용을 포함할 필요가 있다. 같은 하드웨어와 서비스 수준 목표에서 품질-비용 곡선과 행동 분기별 비용을 비교해야 월드 모델의 실제 상호작용 성능을 판단할 수 있다.

Ⅶ. 결 론

옴니모달 AI와 월드 모델은 차세대 미디어 생성의 범위를 이해에서 상호작용으로 넓혀 가고 있다. 하지만 고해상도 관측, 장기 기억, 반복적인 미래 예측을 그대로 결합하면 계산 비용을 감당하기 어렵다. 본고에서는 문맥 압축, 모델 압축, 캐시와 기억, 적응적 연산, 시스템과 서빙의 다섯 계층으로 관련 연구를 정리하였다. 이미지·비디오 토큰 축소에서 출발한 효율화 연구는 모달리티 비대칭 압축과 동적 예산 배분으로 옮겨 가고 있으며, 본 연구진이 제안한 ZOO-Prune은 교란에 대한 민감도를 순전파만으로 추정하여 과제에 중요한 토큰을 효율적으로 고를 수 있음을 보였다. 앞으로의 경량화는 모델 전체를 일괄적으로 줄이는 방식만으로는 한계가 있을 것이다. 현재 장면을 선택적으로 보고, 미래 예측에 필요한 상태만 기억하며, 불확실하거나 중요한 미래만 깊게 전개하는 방향으로 세 요소를 함께 최적화해야 한다. 이러한 설계가 뒷받침될 때 옴니모달 월드 모델은 품질, 반응성, 비용 사이의 균형을 맞추면서 실시간 인터랙티브 미디어, 디지털 휴먼, 로봇 등으로 활용 범위를 넓혀 갈 수 있을 것이다.

참 고 문 헌

[1] J. Xu et al., “Qwen2.5-Omni Technical Report,” arXiv:2503.20215, 2025.
[2] D. Ha and J. Schmidhuber, “World Models,” arXiv:1803.10122, 2018.
[3] D. Hafner et al., “Mastering Diverse Control Tasks through World Models,” Nature, vol. 640, no. 8059, pp. 647-653, 2025.
[4] J. Bruce et al., “Genie: Generative Interactive Environments,” Proc. ICML, pp. 4603-4623, 2024.
[5] M. Assran et al., “V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning,” arXiv:2506.09985, 2025.
[6] Chameleon Team, “Chameleon: Mixed-Modal Early-Fusion Foundation Models,” arXiv:2405.09818, 2024.
[7] J. Xie et al., “Show-o: One Single Transformer to Unify Multimodal Understanding and Generation,” Proc. ICLR, 2025.
[8] S. Wu et al., “NExT-GPT: Any-to-Any Multimodal LLM,” Proc. ICML, pp. 53366-53397, 2024.
[9] M. Horowitz, “1.1 Computing’s Energy Problem (and What We Can Do about It),” Proc. ISSCC, pp. 10-14, 2014.
[10] T. Dao et al., “FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness,” Proc. NeurIPS, 2022.
[11] W. Kwon et al., “Efficient Memory Management for Large Language Model Serving with PagedAttention,” Proc. SOSP, 2023.
[12] D. Bolya et al., “Token Merging: Your ViT but Faster,” Proc. ICLR, 2023.
[13] L. Chen et al., “An Image Is Worth 1/2 Tokens after Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models,” Proc. ECCV, pp. 19-35, 2024.
[14] S. Yang et al., “VisionZip: Longer Is Better but Not Necessary in Vision Language Models,” Proc. CVPR, pp. 19792-19802, 2025.
[15] L. Xing et al., “Conical Visual Concentration for Efficient Large Vision-Language Models,” Proc. CVPR, pp. 14593-14603, 2025.
[16] S. Li et al., “MBQ: Modality-Balanced Quantization for Large Vision-Language Models,” Proc. CVPR, pp. 4167-4177, 2025.
[17] G. Xiao et al., “Efficient Streaming Language Models with Attention Sinks,” Proc. ICLR, 2024.
[18] J. Zou et al., “RUTA: Principled Visual Token Allocation via Rate-Utility Optimization,” arXiv:2608.04132, 2026.
[19] S. R. Alvar et al., “DivPrune: Diversity-Based Visual Token Pruning for Large Multimodal Models,” Proc. CVPR, pp. 9392-9401, 2025.
[20] X. Ye et al., “ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models,” Proc. CVPR, pp. 24972-24982, 2025.
[21] K. Tao et al., “DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models,” Proc. CVPR, pp. 18992-19001, 2025.
[22] Y. Ding et al., “OmniSIFT: Modality-Asymmetric Token Compression for Efficient Omni-Modal Large Language Models,” Proc. ICML, PMLR 306, 2026.
[23] G. Sun, Y. Li, Y. Yang, and C. Zhang, “OmniMem: Perturbation-Aware Memory Compression for Streaming Audio-Visual LLMs,” arXiv:2606.07577, 2026.
[24] Y. Kim, Y. Zhang, H. Liu, A. Jung, S. Lee, and S. Hong, “ZOO-Prune: Training-Free Token Pruning via Zeroth-Order Gradient Estimation in Vision-Language Models,” Proc. CVPR, pp. 39572-39582, 2026.
[25] X. Shen et al., “LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding,” Proc. ICML, pp. 54582-54599, 2025.
[26] Y. Shen, S. Tian, J. Yang, and Z. Liu, “A Simple Baseline for Streaming Video Understanding,” arXiv:2604.02317, 2026.
[27] J. Yi et al., “WorldKV: Efficient World Memory with World Retrieval and Compression,” arXiv:2605.22718, 2026.

저 자 소 개



홍 성 은

- 2010년 : 한양대학교 컴퓨터공학 학사
- 2012년 : KAIST 전산학부 석사
- 2018년 : KAIST 전산학부 박사
- 2018년 ~ 2020년 : SK텔레콤 T-Brain 연구원
- 2020년 ~ 2023년 : 인하대학교 조교수
- 2023년 ~ 현재 : 성균관대학교 인공지능융합학과 부교수
- 주관심분야 : 멀티모달 학습, 효율적 머신러닝, 도메인 적응, 로봇 학습