실시간 확산 확률 모델 기반 생성형 시맨틱 비디오 통신 연구 동향

권순원, 안세영, 권용석, 김유신, 조성현*

한양대학교, *한양대학교 ERICA

kwonsonwon@hanyang.ac.kr, tpdud1014@hanyang.ac.kr, totoey200@hanyang.ac.kr,
hpwgg045@hanyang.ac.kr, *chopro@hanyang.ac.kr

A Survey on Generative Semantic Video Communications based on Real-Time Diffusion Probabilistic Model

Kwon Sun Won, Ahn Se Young, Kwon Yong Seok, Kim Yu Shin, Cho Sung Hyun*

Hanyang Univ., *Hanyang ERICA Univ.

요 약

시맨틱 비디오 통신은 비디오를 그대로 전송하는 대신 장면 의미를 유지하는 잠재 표현을 전달하고, 수신단에서 이를 조건으로 프레임을 합성 및 복원함으로써 초저비트율 환경에서도 체감 품질을 유지하려는 새로운 통신 방식이다. 최근에는 일반 장면에서 더 높은 복원 품질을 얻기 위해 확산 모델 기반 생성형 모델이 시맨틱 비디오 통신에 본격적으로 적용되고 있다. 그러나 확산 모델은 다단계 denoising 반복으로 계산량이 커 실시간 전송 시나리오에서 적용이 제한된다. 본 논문은 이를 완화하고 실시간성을 확보하려는 연구들을 두 가지 흐름으로 정리한다. 첫째, 이전 프레임과 모션 등 직전 정보를 활용해 초기 노이즈를 유리하게 만들고 역확산 구간을 단축하는 추론 가속 접근이다. 둘째, 참조 프레임을 앵커로 고정한 뒤 이후 프레임은 변화분 중심으로 복원하여 생성 범위를 줄이는 앵커 기반 생성 최소화 접근이다. 또한 큰 모션과 장면 전환처럼 프레임 간 상관이 낮아지는 구간에서는 초기 노이즈와 앵커의 신뢰도가 저하될 수 있음을 한계로 주목하고, 이를 보완하기 위한 후속 연구 방향을 논의한다.

Ⅰ. 서 론

시맨틱 비디오 통신은 원 신호의 완전 복원보다 장면의 의미를 유지하는 잠재 표현을 추출해 전달하고 수신단에서 조건부 생성으로 프레임을 합성 및 복원함으로써, 초저비트율 환경 및 열악한 채널 조건에서도 유의미한 시각 정보와 체감 품질을 유지하려는 기술이다 [1]. 초기 시맨틱 비디오 통신은 화상회의와 같이 대상과 변형 양상이 제한된 도메인에 주로 적용되었으며, 키포인트나 랜드마크와 같은 명시적 시맨틱을 전송한 뒤 이를 기반으로 프레임을 생성하는 방식이다 [1, 4]. 그러나 실제 비디오 트래픽의 다수는 특정 객체에 국한되지 않는 콘텐츠로 구성되며, 장면 구조와 객체 간 관계가 복잡하게 얽혀있는 양상을 보인다. 이때 명시적 시맨틱만으로는 배경 구조, 객체 경계, 상호작용과 같은 비디오 생성에 필요한 핵심 단서를 보존하기 어려우므로, 시맨틱 비디오 통신은 도로, 도시, 실외 환경 등 객체와 구조가 복잡한 일반 장면으로 확장될 필요가 커졌다.

일반 장면 확장에서 핵심으로 부각된 접근이 생성형 시맨틱 통신(generative semantic communication, GSC)이다 [1]. GSC는 프레임에서 시맨틱 맵을 추출해 전송하고, 수신단에서 생성 모델을 구동해 프레임을 재구성한다 [2]. 기존에는 빠른 추론 속도를 바탕으로 생성적 적대 신경망 기반 조건부 합성이 주로 활용되어 왔다 [2, 5]. 그러나 일반 장면에서는 세밀한 구조 보존과 안정적인 합성이 동시에 요구되며, 단일 조건이나 제한된 변형을 가정하는 합성은 시각적 왜곡과 시간적 불안정을 야기하기 쉽다 [4]. 이 때문에 최근 GSC에는 확산모델 기반 합성이 본격적으로 도입되고 있다 [1, 2, 4]. 확산모델은 단계적 복원 과정을 통해 복잡한 객체·텍스처·구조를 더 정교하게 재현하는 경향이 있고, 이전 프레임 참조와 프레임, 잔차, 모션과 같은 다중 조건을 결합한 조건부 생성에서도 높은 표현력과 설계 유연성을 제공한다 [2, 3, 4]. 그 결과 일반 장면에서 더 높은 품질과 더 자연스러운 프레임 생성을 기대할 수 있다.

그러나 확산모델을 사용하는 GSC에서 지속적으로 지적받는 문제는 반복되는 denoising 과정으로 인한 높은 계산량이며, 이는 곧 저지연·실시간 전송 시나리오에서의 적용 가능성을 제한한다 [1, 2, 4]. 특히 시맨틱 비디오 통신은 수신단 합성 과정이 전송 주기와 결합되므로, 생성 지연은 곧 시스템 지연으로 누적된다. 따라서 확산 기반 GSC의 성능을 논할 때에는 품질만이 아니라, 제한된 시간 예산 내에서 프레임을 복원할 수 있는가를 고려해야한다.

본 논문은 확산 모델의 초기 노이즈를 정교화하여 확산 추론을 가속화하는 접근과 참조 프레임을 앵커로 사용해 생성 자체를 최소화하는 접근, 두 가지로 정리하여 확산 기반 GSC의 실시간성을 고려한 최신 연구를 소개한다.

Ⅱ. 본 론

Ⅱ-1. 초기 노이즈 정교화 기반 확산 추론 가속

확산모델 기반 GSC에서 계산 병목은 denoising 반복에 의해 발생한다. 이를 완화하기 위한 한 흐름은 처음부터 순수 잡음에서 시작해 긴 역확산을 수행하는 대신, 직전 정보로 초기 노이즈를 만들거나 추론 경로를 짧게 유도해 필요한 denoising 스텝 수를 줄이는 방식이다. 이 절에서는 초기 노이즈 정교화 관점에서 확산 추론을 가속화하는 연구를 정리한다.

[2]에서는 denoising 스텝을 줄이는 ControlVideo-SemCom (CVSC)를 소개한다. CVSC는 수신단이 주기적으로 수신 및 복원의 기준 프레임을 두고, 프레임 간 모션을 누적해 기준 프레임의 각 위치가 현재 시점에서 어디로 이동해야 하는지를 추정한 뒤 기준 프레임을 현재 시점 좌표로 정렬해 현재 프레임에 가까운 초기 노이즈를 만든다. 이후 이 초기 노이즈를

표 1. 확산모델을 사용하는 GSC 실시간성 연구 비교

실시간성 연구 기법	[2]	[3]	[4]	[5]
확산 추론 가속	O	O	X	X
생성 최소화	X	X	O	O
큰 모션/장면 변화 감지	X	X	X	X
불확실성 추정	X	X	X	X
품질 저하 안전 장치	O	O	X	X

특정 노이즈 수준까지 변형한 상태에서 denoising을 시작하도록 구성해, 매 프레임을 순수 잡음에서 새로 생성하지 않고 짧은 역확산 구간에서 잔여 왜곡과 세부 성분만 복원하도록 만든다. 결과적으로 초기 노이즈 정교화해 역확산 구간을 단축하고 계산 부담을 줄인다.

[3]에서는 프레임 간 모션 일관성을 이용해 확산 추론을 가속하는 Denoising Reuse를 제안한다. 한 기준 프레임에 대해서만 확산 백본을 끝까지 수행해 각 denoising 단계의 중간 잠재 표현을 미리 저장해 두고, 나머지 프레임은 모션 특징으로 기준 프레임의 해당 단계 노이즈와 잠재를 정렬해 이후 단계만 denoising 하도록 만든다. 전환 스텝은 모션 크기에 따라 적절한 지점을 선택하며, 모션 네트워크와 전환 스텝 선택기를 두어 초기 구간의 계산을 건너뛰면서도 후반 구간에서 품질이 유지된다.

Ⅱ-2. 앵커 프레임과 변화분 복원으로 생성 최소화

확산모델 기반 GSC에서 실시간성을 고려한 또 다른 접근법은 매 프레임을 새로 생성하기보다 참조 프레임을 앵커로 고정하고, 이후 프레임은 앵커 대비 변화분만 복원하도록 생성 범위를 축소하는 방식이다. 이 절에서는 앵커 기반으로 생성과 복원을 최소화하는 연구를 정리한다.

[4]에서는 Wireless Video Semantic Communication with Decoupled Diffusion Multi frame Compensation (WVSC-D)를 제시한다. WVSC-D는 비디오를 구간으로 나누고 구간의 첫 시맨틱 I 프레임을 참조 앵커로 전송한 뒤, 구간 내 나머지 프레임은 I 프레임 대비 변화분을 잔차로 표현해 전송한다. 수신단은 앵커로 구간 내에서 공유되는 기본 장면 구조를 확보하고, 복원 과정에서는 참조 성분과 잔차 성분을 분리해 앵커의 구조는 유지하면서 프레임별 고유한 변화가 복원되도록 구성한다.

[5]에서는 Content Frame Motion Latent Diffusion Model (CMD)를 제시한다. CMD는 비디오를 콘텐츠 프레임과 모션 잠재 표현으로 분해해 생성 대상을 줄인다. 오토인코더가 입력 비디오를 콘텐츠 성분과 모션 성분으로 인코딩하며, 콘텐츠 프레임은 프레임별 중요도를 학습해 여러 프레임을 가중 결합한 형태로 구성되어 비디오 전체의 기준이 되는 앵커로 사용된다. 생성 단계에서는 사전학습된 이미지 확산모델로 콘텐츠 프레임을 먼저 생성해 장면의 기본 외관을 고정하고, 별도의 경량 확산모델로 모션 잠재 표현만 추가로 생성한다. 이후 디코더는 콘텐츠 프레임이 제공하는 공간적 기반 특징과 모션 잠재 표현이 제공하는 시간적 변화 특징을 결합하여 전체 프레임열을 복원한다.

Ⅱ-3. 프레임 상관 저하 구간에서의 신뢰도 판단과 앵커 갱신 한계

표 1에서 보듯, 실시간성을 목표로 한 확산 기반 GSC 연구들은 이전 정보로 초기 노이즈를 유리하게 만들거나 앵커 프레임을 기준으로 변화분만 복원하는 구조를 택한다. 그러나 대부분 인접 프레임 간 높은 중복을 전제로 추론 시간 단축에 집중하며, 큰 모션이나 장면 전환처럼 프레임 상관이 급격히 낮아지는 구간과 그에 따른 불확실성은 충분히 고려하지 않는다. 하지만 실시간 시나리오에서 감지와 불확실성 추정은 부가 기능이 아니라, 고정된 시간 예산 안에서 실패를 제어하기 위한 필수 요소다. 큰 모션과 장면 전환은 정렬 기반 초기화와 앵커 기반 변화분 복원 가정을 동시에 무너뜨리며, 신뢰도 판단 없이 동일한 스텝 수와 보정 강도를 적용하면 품질 저하가 급격히 발생하거나 오류가 누적되어 안정적인 서비스 품질을 보장하기 어렵다.

요약하면, 실시간성을 위한 연구들은 정상 구간에서는 매우 효율적이다. 그러나 큰 모션과 장면 전환처럼 프레임 간 상관이 낮아지는 구간에서는 초기 노이즈와 앵커의 신뢰도를 판단하기 어렵고, 특정 상황에 맞게 기준을 재설정하는 메커니즘을 다루지 않았다. 따라서 후속 연구에서는 전환 감지, 모션 불확실성 추정, 적응적 스텝 수 조절, 앵커 갱신 정책 같은 시스템 수준의 보완이 핵심 과제로 이어져야 한다.

Ⅲ. 결 론

본 논문에서는 확산모델 기반 GSC에서 실시간성을 확보하기 위한 관련 연구를 소개하였다. 해당 접근은 초기 노이즈 정교화를 통한 추론 가속과 앵커 기반 생성 범위 축소로 구분할 수 있다. 향후에는 큰 모션과 장면 전환 구간에서도 안정적으로 동작하도록 전환 감지와 적응적 제어를 포함한 시스템 수준의 보완이 필요할 것이다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임 (No. KRIT-CT-22-021, 우주공간 신호정보 특화연구실)

참 고 문 헌

[1] C. Eteke, A. Griessel, W. Kellerer, and E. Steinbach, “Real-time semantic video communication of general scenes,” in Proc. IEEE International Conference on Image Processing (ICIP), Athens, Greece, Oct. 2024, pp. 1916 - 1920.

[2] C. Eteke, A. Griessel, W. Kellerer, and E. Steinbach, “Real-time semantic video communication with temporally consistent and controllable diffusion models,” in Proc. IEEE International Conference on Image Processing (ICIP), Abu Dhabi, UAE, Oct. 2025, pp. 361 - 365.

[3] C. Wang et al., "Denoising Reuse: Exploiting Inter-Frame Motion Consistency for Efficient Video Generation," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 9, pp. 8436-8451, Sept. 2025.

[4] B. Xie, Y. Wu, Y. Shi, B. Feng, W. Zhang, J. Park, and T. Q. S. Quek, “Wireless video semantic communication with decoupled diffusion multi-frame compensation,” IEEE Transactions on Communications, early access, 2025.

[5] S. Yu, W. Nie, D.-A. Huang, B. Li, J. Shin, and A. Anandkumar, “Efficient video diffusion models via content-frame motion-latent decomposition,” in Proc. International Conference on Learning Representations (ICLR), Vienna, Austria, 2024.