11주차

Instruduction

생성형 AI 콘텐츠 관련 대부분의 기존 연구는 하나의 단일 모달리티 내에서 콘텐츠를 생성하는 데 집중

→ 생성된 영상은 음성이 동반되지 않거나 생성된 음성이 시각적 효과와 동기화되지 않는 문제가 발생

⇒ 본 연구에서 visual-audio 생성 문제 탐구

- 가능한 해결책 = 시각과 음향 콘텐츠를 두 단계로 나누어 생성하기

  예를 들어, 사용자가 입력 문장을 기반으로 먼저 텍스트-투-비디오(Text-to-Video, T2V) 모델을 사용해 영상을 생성하고 그다음 비디오-투-오디오(Video-to-Audio, V2A) 모델로 정렬된 음성을 만들게 함.

  - 기존의 V2A나 A2V 생성 방식은 특정 도메인에 한정되거나 생성 품질이 낮은 문제 있음
  - **Joint Visual-Audio generation, Joint-VA** 에 대한 연구는 매우 제한적이고 성능이 낮음

본 연구에서는 **Open-domain** 시각-음향 생성에 대한 새로운 생성 패러다임을 제안

1. 잘 훈련된 **단일 모달 텍스트 조건부 생성 모델(single-modality text-conditioned generation models)** 들이 이미 각각의 모달리티(시각 또는 음향)에 대해 우수한 성능을 보인다는 점.

   → 이들을 활용하면 각 모달을 별도로 학습하는 데 드는 비용을 줄일 수 있다.

2. 사전 학습된 ImageBind 모델은 서로 다른 데이터 모달 간에 효과적인 연결을 구축할 수 있는 능력을 가지고 있으며, 동일한 의미 공간(semantic space) 내에서 이들을 정렬시킬 수 있다

⇒ ImageBind를 다리로 사용하여 서로 다른 모달리티를 효과적으로 연결하고 통합하는 방법을 탐구해보자

- **ImageBind를 확산 기반 생성 모델의 잠재 공간내 정렬기(aligner)** 로 사용하는 방법을 제안

- 한 모달리티를 생성하는 동안 해당 모달의 노이즈가 포함된 latent과 다른 모달의 조건을 결합해 생성 과정을 유도하는 **guidance signal**를 만든다.
- 이 가이드는 점진적으로 확산 복원(denoising) 과정에 주입되어, 생성된 콘텐츠가 ImageBind 의미 공간에서 입력 조건과 더 정렬되도록 함.
- Joint-VA 에서는 이 가이드를 양방향 으로 작용시켜 두 모달의 생성 과정을 동시에 영향을 미치도록 설계.

→ text-guided joint video-audio generation 에 대한 최초의 연구라고 할 수있음

Method

Preliminaries

Latent Diffusion Model

Classifier Guidance

무조건(unconditional) 확산 모델에 조건 정보를 추가해 특정 클래스의 샘플을 생성하도록 하는 방법

- 클래스 라벨 y 가 주어졌을 때 확산 모델 p_(z_t|z_t+1) :

$$p_{\theta,\phi}(z_t|z_{t+1},y) = \mathcal{Z}p_\theta(z_t|z_{t+1})p_\phi(y|z_t,t)$$

형태로 근사

- ϕ : 시간 인식 분류기(time-aware classifier)

  라벨 확률의 기울기를 이용해 다음과 같이 조정된 노이즈를 얻음

$$\hat{\epsilon}(z_t) = \epsilon_\theta(z_t) - \sqrt{1-\bar{\alpha}_t}\nabla_{z_t}\log p_\phi(y|z_t)$$

Linking Multiple Modalities

본 연구의 목표는 서로 다른 모달리티로 생성된 샘플들을 joint semantic space 에서 가깝게 만드는 것

→ **ImageBind**를 정렬기(aligner)로 사용

- 이미지, 텍스트, 비디오, 오디오, 깊이, 열감지(thermal) 등 여러 모달리티를 하나의 의미 공간에 결합하는 모델

$$L_{M_1,M_2} = -\log \frac{\exp(q_i^T k_i/\tau)}{\exp(q_i^T k_i/\tau) + \sum_{j\neq i} \exp(q_i^T k_j/\tau)}$$

Diffusion Latent Aligner

Problem Formulation

M1 = generative modality, M2 = conditional modality 이라 할 때

잠재 확산 모델 θ 가 M1 데이터를 생성한다고 할 때 목표는 **M2 의 조건 정보 x^M2 를 이용해 M1 의 중간 생성 결과가 조건에 맞게 정렬되도록 하는 것**

→ 정렬기 A 가 주진 시점 t 의 잠재 변수 z_t 를 입력 받아 조건을 반영한 수정 잠재 변수 z 햇_t 생성

$$\hat{z}_t^{M_1} = \mathcal{A}(z_t^{M_1}, x^{M_2})$$

→ joint VA 는 두 모달 모두로부터 정보를 받아 결합

$$(\hat{z}_t^{M_1}, \hat{z}_t^{M_2}) = \mathcal{A}(z_t^{M_1}, z_t^{M_2})$$

Multimodal Guidance

정렬기를 훈련 없이 적용하기 위해 ImageBind의 표현 학습 능력을 활용하여 노이즈 제거 과정을 유도

- 각 시점 t 에 서 예측된 노이즈 e햇 을 이용하면 노이즈 없는 데이터 z_0 을 복원할 수 있음

$$\tilde{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} z_t - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}} \hat{\epsilon}$$

- 이후 z_0 과 x^M2 를 imagebind 에 입력해 두 임베딩 간 거리 F 를 계산하고 이 거리를 손실 항으로 사용하여 역전파

$$L(\tilde{z}_0, x^{M_2}) = 1 - \mathcal{F}(E^{M_1}(\tilde{z}_0), E^{M_2}(x^{M_2}))$$

$$\hat{z}_t = z_t - \lambda_1 \nabla_{z_t} L(\mathcal{D}(\tilde{z}_0), x^{M_2})$$

Dual/Triangle Loss

음성 데이터는 종종 의미 정보가 부족하므로 이를 보완하기 위해, 세 번째 모달리티(예: 텍스트)를 도입해 의미 정렬을 강화

- A2V (Audio→Video):

$$L_{a2v} = \mathcal{F}(e_v, e_a) + \mathcal{F}(e_v, e_p)$$

- V2A (Video→Audio):

$$L_{v2a} = \mathcal{F}(e_a, e_v) + \mathcal{F}(e_a, e_p)$$

- Joint-VA (양방향):

$$L_{joint-va} = \mathcal{F}(e_v, e_p) + \mathcal{F}(e_v, e_a) + \mathcal{F}(e_a, e_p)$$

→ 세 모달 간 관계를 삼각 구조로 정렬시켜 불완전한 의미 정보를 보완

Guided Prompt Tuning

시각→음향 변환에서는 좋은 정렬 결과를 보였으나 음향→시각 생성에서는 일관성이 낮음

→ 프롬프트 임베딩 y 를 최적화하여 의미 손실의 그래디언트를 반영하도록 한다:

$$\hat{y} = y - \lambda_2 \nabla_y L$$

→ 모든 시점에서 일관된 의미적 가이드를 유지하며 품질 향상 효과

Experiments

Experimental Setup

- 데이터셋: **VGGSound, Landscape**
  - Image-to-Audio 실험을 위해 각 영상에서 키 프레임을 추출
  - VGGSound에서 **3,000개의 영상-오디오 쌍**을 무작위로 샘플링
  - 3,000쌍은 Video-to-Audio, 3,000쌍은 Audio-to-Video, 또 다른 3,000쌍은 Image-to-Audio 생성에 사용
  - Landscape 데이터셋에서 **200개의 영상-오디오 쌍**을 무작위로 선택 하여 Joint Video-Audio 생성 평가에 사용
- 구현 세부사항
  - Video-to-Audio 및 Image-to-Audio 생성: **AudioLDM**
  - Audio-to-Video 생성: AnimateDiff
  - Joint Video-Audio 생성: 둘 다

Baselines

- Video-to-Audio : **SpecVQGAN** 사용
- Image-to-Audio: Im2Wav 사용
- Audio-to-Video: TempoToken 사용
- joint Video-Audio Generation: MM-Diffusion을 기준선으로 사용.

Ours-vanilla 버전

- Video-to-Audio 태스크에서는 키 프레임 추출하고 Image Caption 모델을 이용해 자막을 생성한 후 AudioLDM으로 오디오를 생성
- Audio-to-Video에서는 오디오 자막 모델을 통해 텍스트를 추출하고 AnimateDiff로 영상을 생성
- Joint 생성 태스크에서는 프롬프트 텍스트를 AudioLDM과 AnimateDiff에 동시에 입력하여 수행

Visual-to-Audio Generation

- Image-to-Audio의 경우 의미적정렬이 중요
- Video-to-Audio의 경우 의미적 정렬에 더해 **시간적 정렬(temporal alignment)** 도 필요

평가 지표

- **KL Divergence (KL↓)**
- **Inception Score (ISc↑)**
- **Frechet Distance (FD↓)**
- **Frechet Audio Distance (FAD↓)**

Task	Method	Metric			
V2A		KL↓	ISc↑	FD↓	FAD↓
	SpecVQGAN [26]	3.290	5.108	37.269	7.736
	Ours-vanilla	3.203	5.625	40.457	**6.850**
	Ours	**2.619**	**5.831**	**32.920**	7.316
I2A		KL↓	ISc↑	FD↓	FAD↓
	Im2Wav [37]	2.612	**7.055**	**19.627**	7.576
	Ours-vanilla	3.115	4.986	33.049	7.364
	Ours	**2.691**	6.149	20.958	**6.869**
A2V		FVD↓	KVD↓	AV-align↑	-
	TempoToken [46]	1866.285	389.096	0.423	-
	Ours-vanilla	417.398	36.262	0.518	-
	Ours	**402.385**	**34.764**	**0.522**	-
Joint VA Generation		FVD↓	KVD↓	FAD↓	
	Landscape: MM [36]	**1141.009**	**135.368**	7.752	-
	Landscape: MM [36] + Ours	1174.856	135.422	**6.463**	-
		AV-align_{bind}↑	VT-align_{bind} ↑	AT-align_{bind} ↑	AV-align ↑
	Open-domain: MM[36]	N/A	N/A	N/A	N/A
	Open-domain: Ours-vanilla	0.074	0.322	0.081	0.226
	Open-domain: Ours	**0.096**	**0.324**	**0.138**	**0.283**

Table 1. Quantitative comparison with baselines on four tasks.

→ 훈련 없이도 대규모 오디오-비디오 페어 학습을 요구하는 기준선보다 우수한 성능

Audio-to-Video Generation

생성된 영상이 입력 오디오와 의미적, 시간적으로 정렬되어야 함

- 표 1의 결과에서, 제안된 훈련 없는 방식이 기존의 훈련 기반 모델보다 의미 정렬 및 영상 품질에서 모두 우수함을 확인

Figure 5. Compared with baseline on the audio-to-video task. Given the input audio, the generated videos by TempoToken are not aligned with the input audio and the generation with poor visual quality. Our method can produce visually much better and semantically aligned content with the input condition.

→ TempoToken이 시각 품질과 정렬에서 불안정한 반면, 우리의 방법은 공유된 오디오-비주얼 표현 공간을 이용해 균형 잡힌 품질과 정렬 성능을 달성

Joint Video and Audio Generation

입력 텍스트로부터 고품질의 영상과 음향을 함께 생성, 텍스트-오디오 및 텍스트-비디오 정렬을 동시에 유지해야 함

지표

- **FVD** (영상 품질)
- **FAD** (음질)
- **AV-align** (오디오-비디오 정렬)
- **TA-align** (텍스트-오디오 정렬)
- **TV-align** (텍스트-비디오 정렬)

우리의 latent aligner 는 기존 MM-Diffusion 구조에 직접 결합될 수 있으며, 그 결과 오디오 품질이 향상되면서도 영상 품질은 유지되었음.

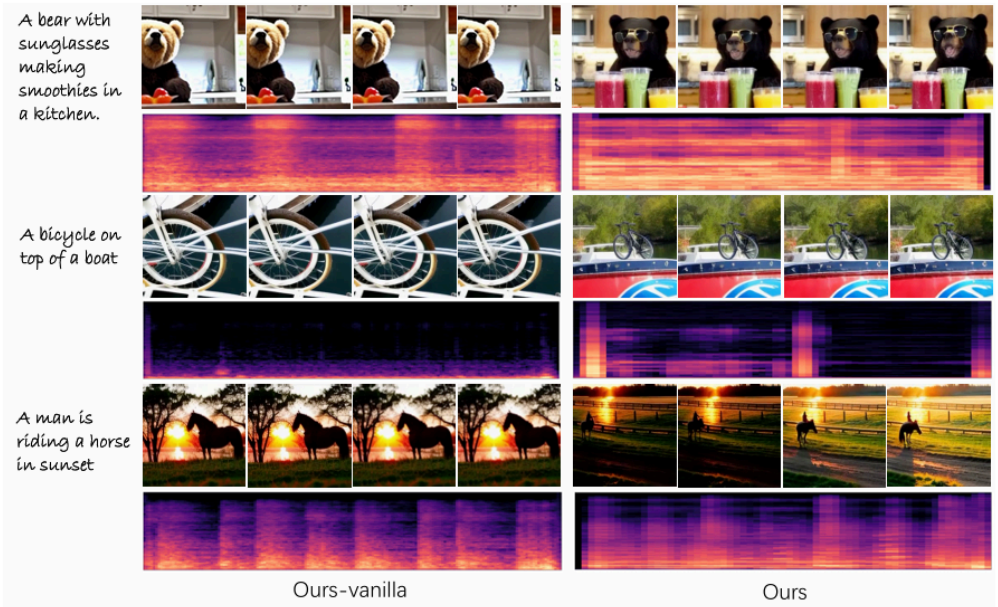

Figure 4. Compared with baseline on the joint video-and-audio generation task. Our method can produce better text-aligned visual content than the vanilla model. Besides, our generated audio is also of better quality and better alignment with the generated videos.

→ Text-to-Audio, Text-to-Video, Audio-Video 정렬 모두 개선

Limitations

성능이 기반 생성 모델(AudioLDM, AnimateDiff) 의 표현력에 의존