



VGGnet

Video Graphic Generation network

팀 구성

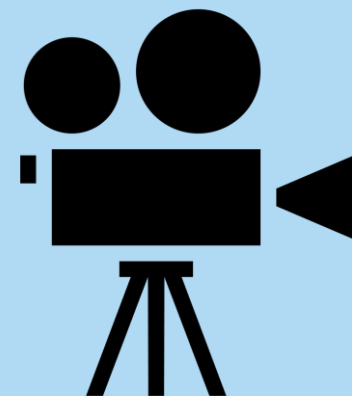
9기 박찬혁

11기 최가운

12기 박승호

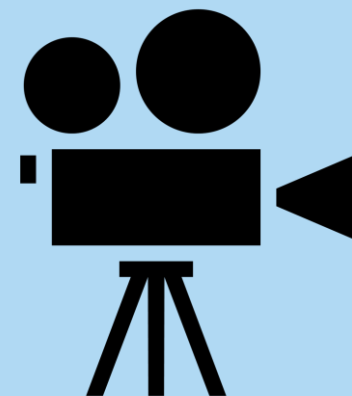
12기 유선재

12기 제갈건



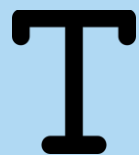
짧은 비디오 생성

- Quality
- Multimodality
- Time



예상 결과물

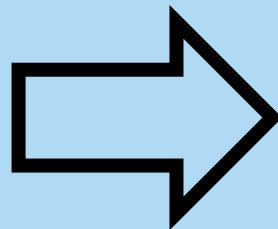
Text



Audio

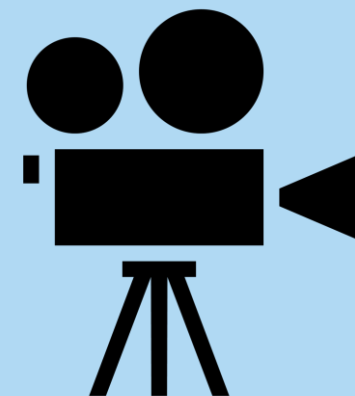


Image



Video

- Show-1
- ImageBind
- Binding Network
- 실험결과



Show-1 (Baseline)

Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation

Text-to-Video 생성 분야에서 SoTA



Toad practicing karate



A burning lamborghini on the road.



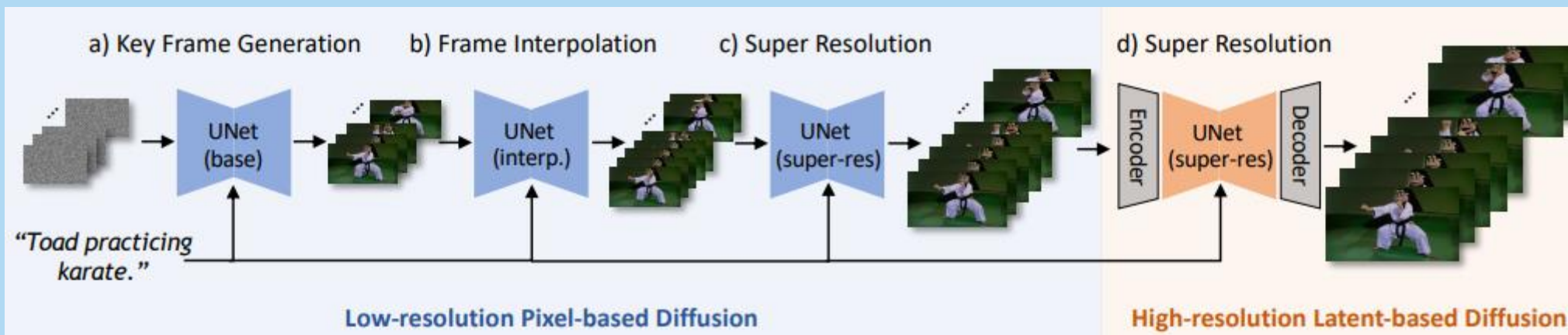
Giant octopus invades new york city.

Show-1 (Baseline)

- Pixel based diffusion + Latent based diffusion

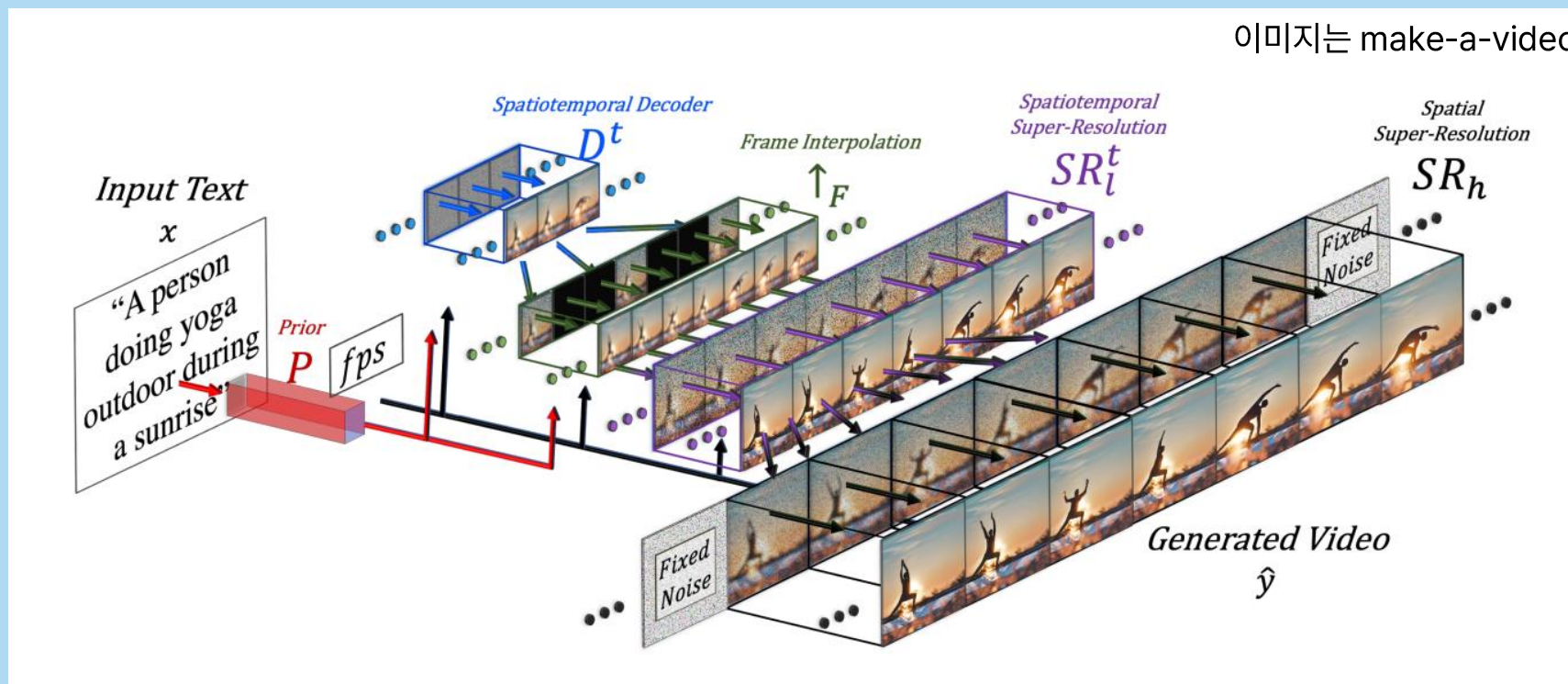
Pixel based: 시간, cost **high**, text-video alignment **good**

Latent based: 시간, cost **low**, text-video alignment **bad**



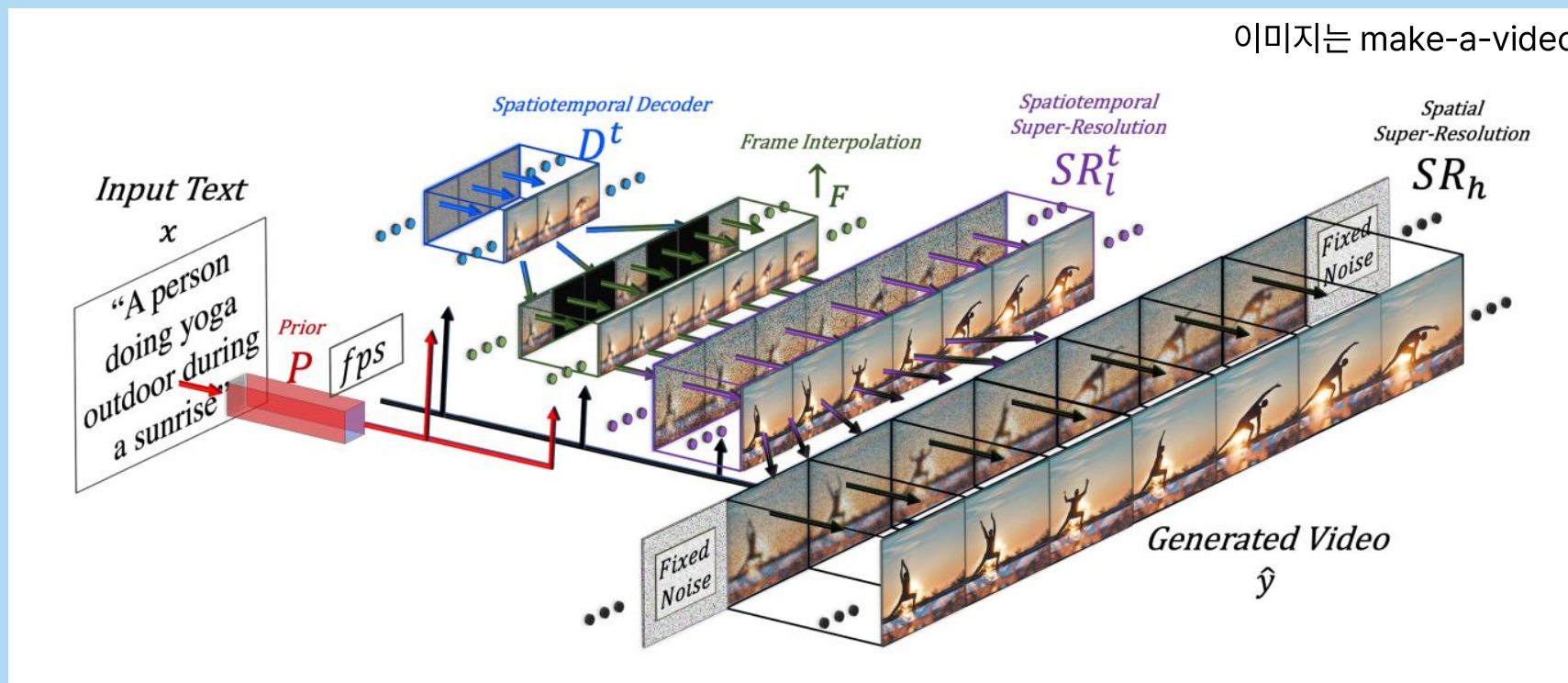
성능 향상 + 기존 모델 inference 72G -> 15G

Show-1 (Baseline)



1. 텍스트를 기반으로 Keyframes 생성 (8장, fps=2)
2. Keyframe 사이를 interpolation (fps=7.5)

Show-1 (Baseline)



3. 1차 super-resolution ($64 \times 40 \rightarrow 256 \times 160$)

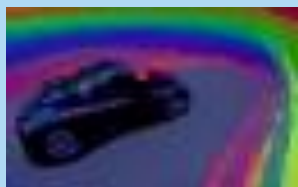
4. 2차 super-resolution ($256 \times 160 \rightarrow 576 \times 320$)

Show-1 (Baseline)

"A burning lamborghini driving on rainbow."



Base (8*64*48)



Interpolation (29*64*48)



SR1 (29*256*160)



SR2 (29*576*320)

Show-1 (Baseline)

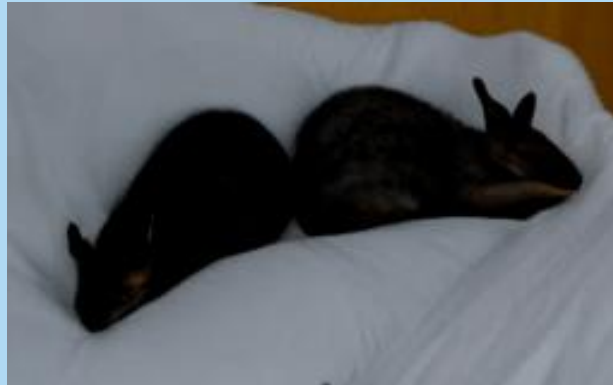
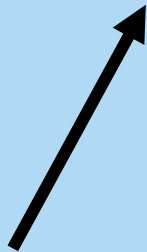
"sleeping shrews in small bed."



Base (8*64*48)



Interpolation (29*64*48)



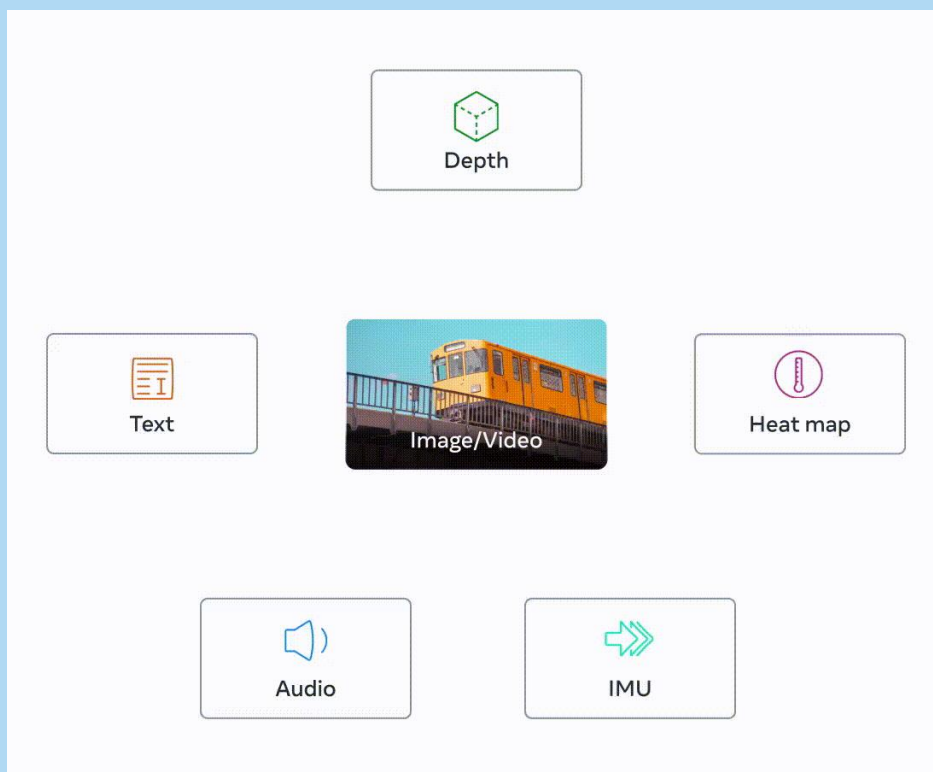
SR1 (29*256*160)



SR2 (29*576*320)

ImageBind (Baseline)

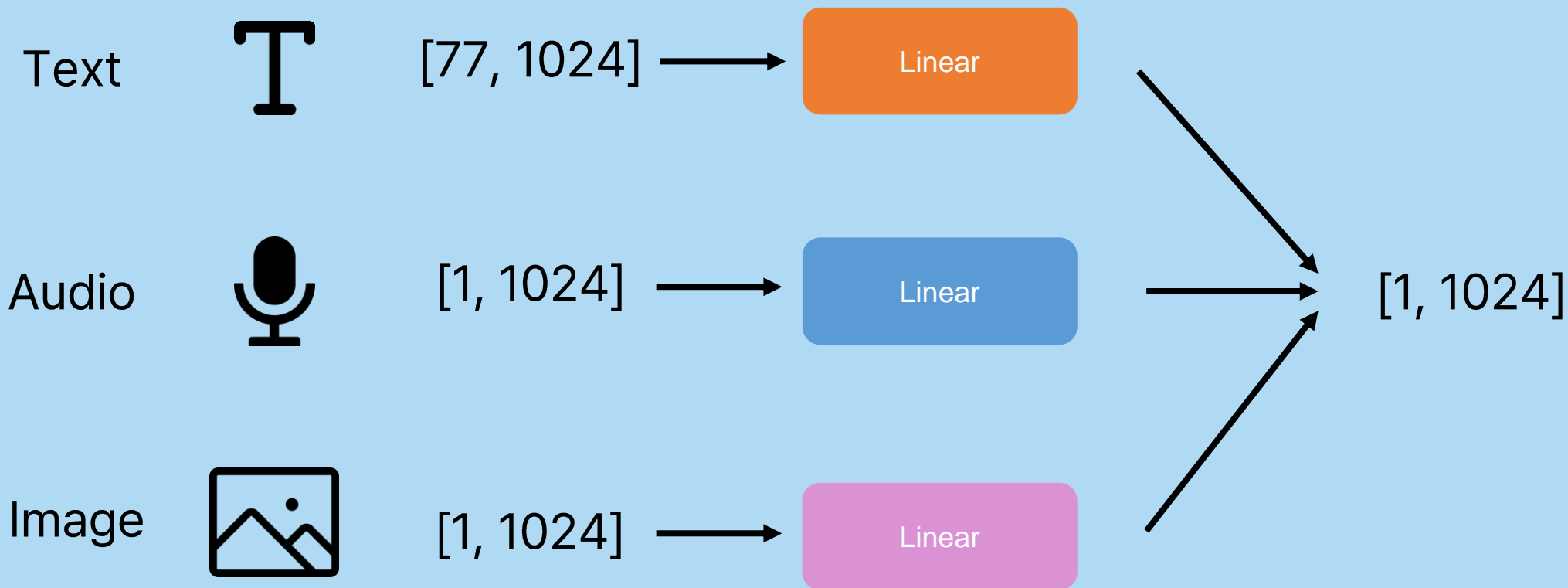
6개 modality를 하나의 embedding space에 표현하자!



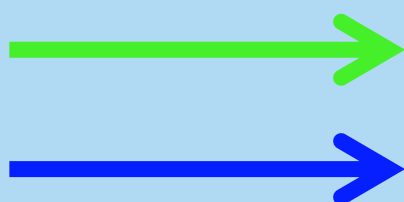
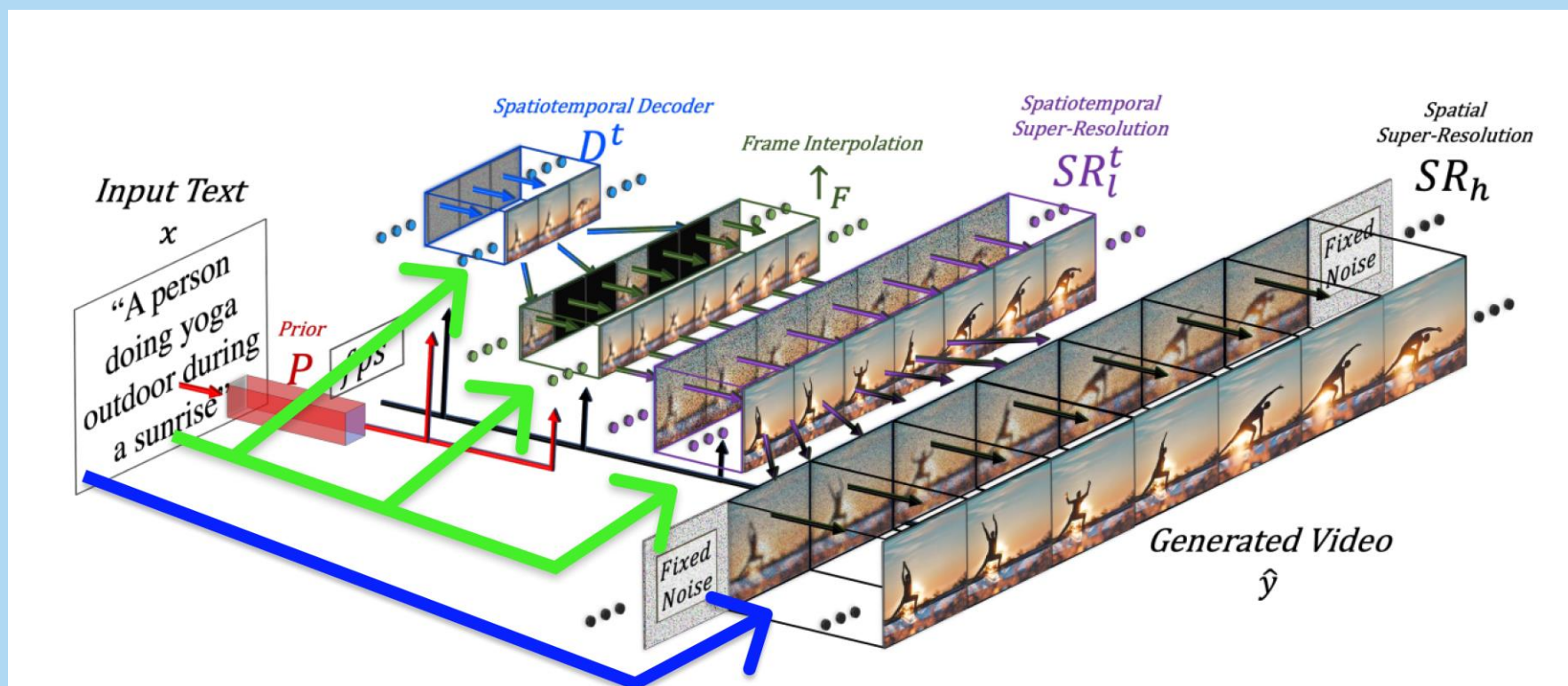
- Text
- Audio
- Image
- HeatMap
- Depth
- IMU

ImageBind (Baseline)

6개 modality를 하나의 embedding space에 표현하자!



Our Task

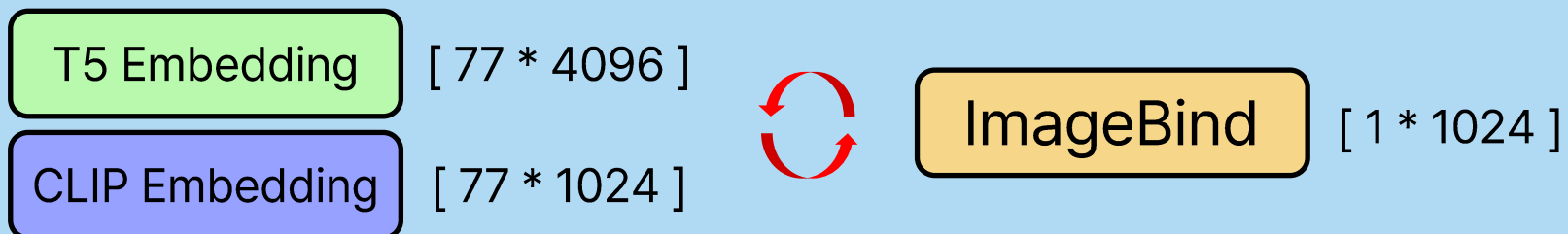


T5 Embedding [77 * 4096]
CLIP Embedding [77 * 1024]



ImageBind [1 * 1024]

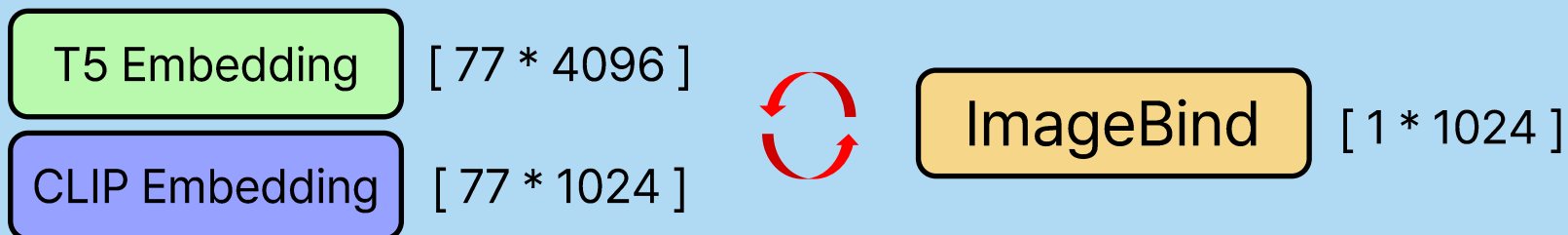
Our Task



Option 1

임베딩 모델을 변경해서 각 레이어를 다시 학습

Our Task

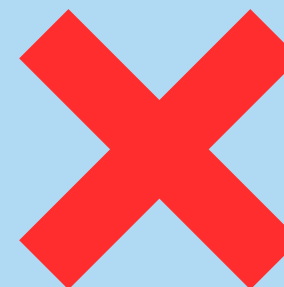


Option 1

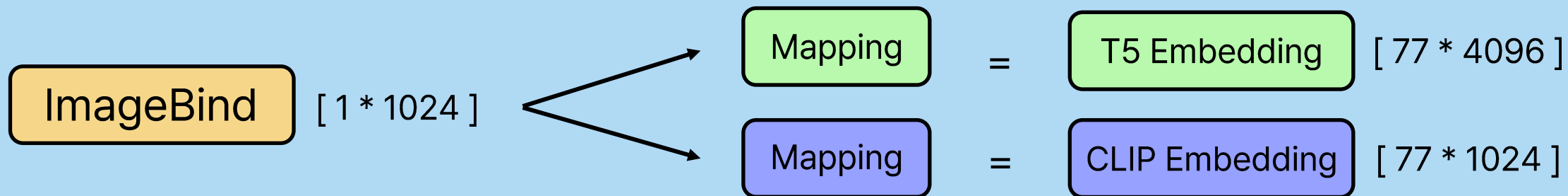
임베딩 모델을 변경해서 각 레이어를 다시 학습

Diffusion Model 4개 다시 학습

첫번째 모델 A100 48개로 72시간 학습...



Our Task



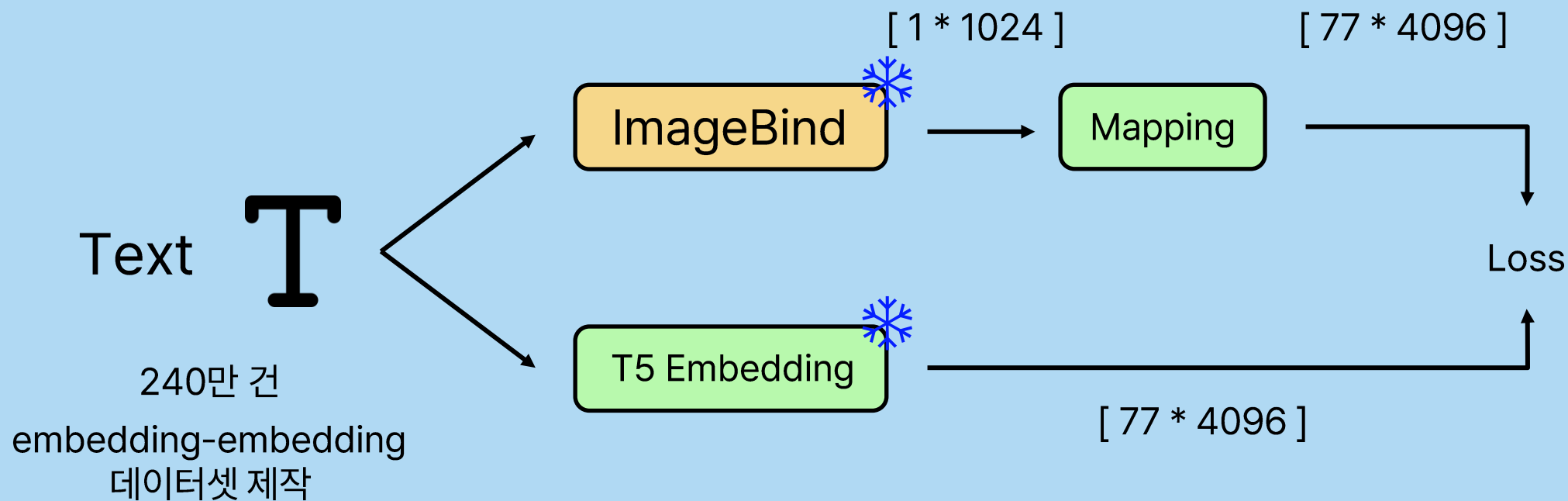
Option 2

ImageBind의 embedding space를

T5, CLIP Embedding의 embedding space로 mapping

Mapping Network

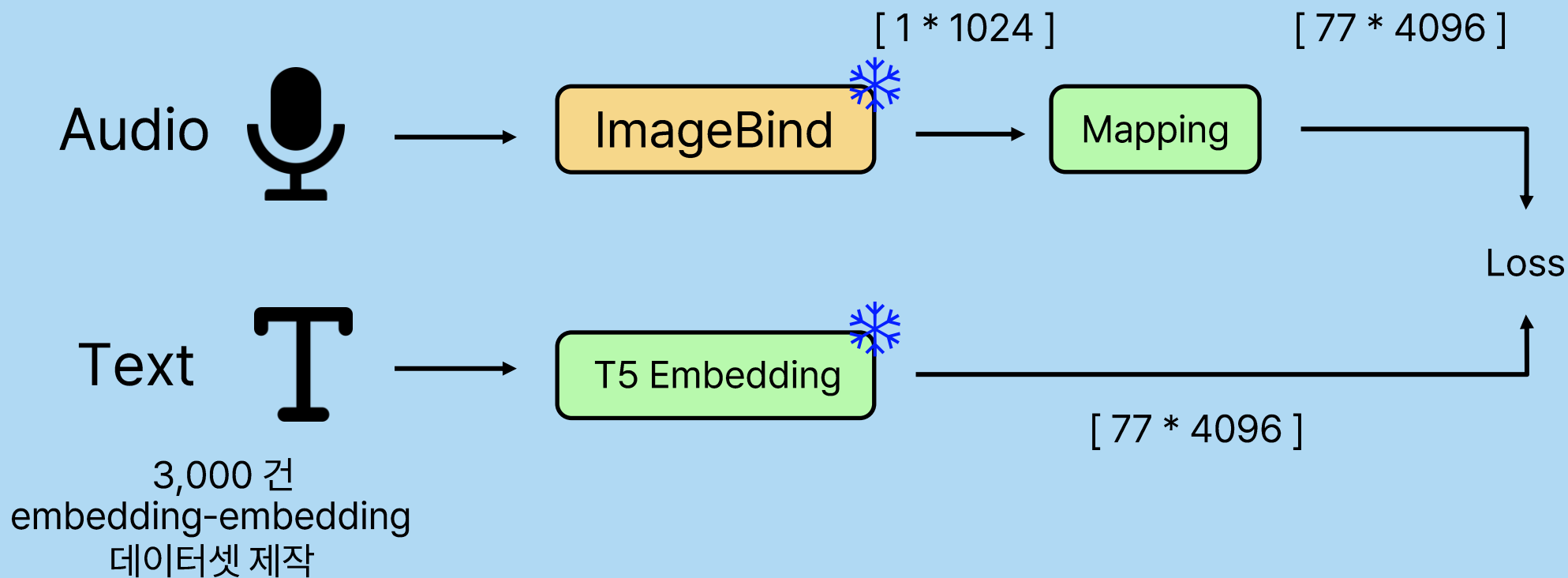
Data



ImageBind + Mapping Model의 출력값이 T5 임베딩 출력값과 같아지도록 학습

Mapping Network

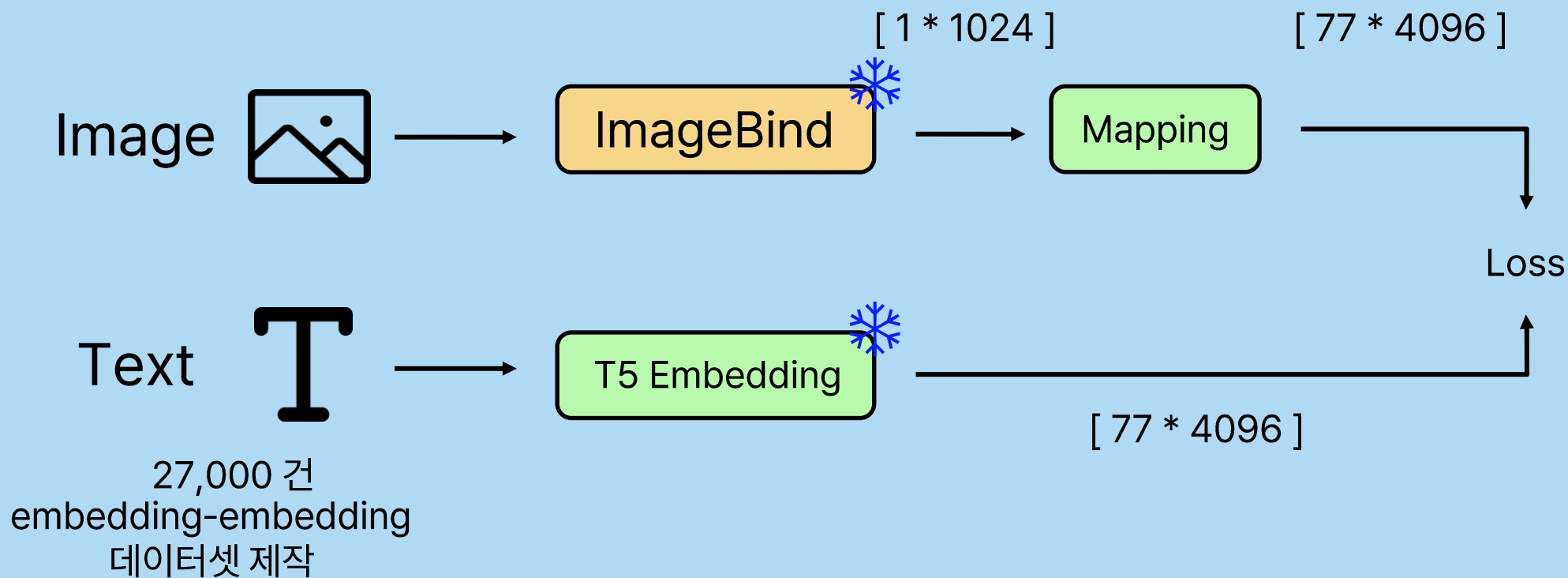
Data



ImageBind + Mapping Model의 출력값이 T5 임베딩 출력값과 같아지도록 학습

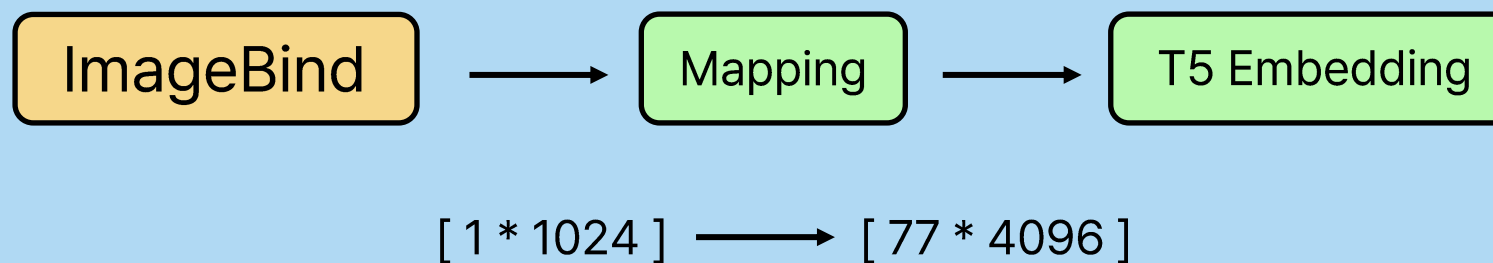
Mapping Network

Data



ImageBind + Mapping Model의 출력값이 T5 임베딩 출력값과 같아지도록 학습

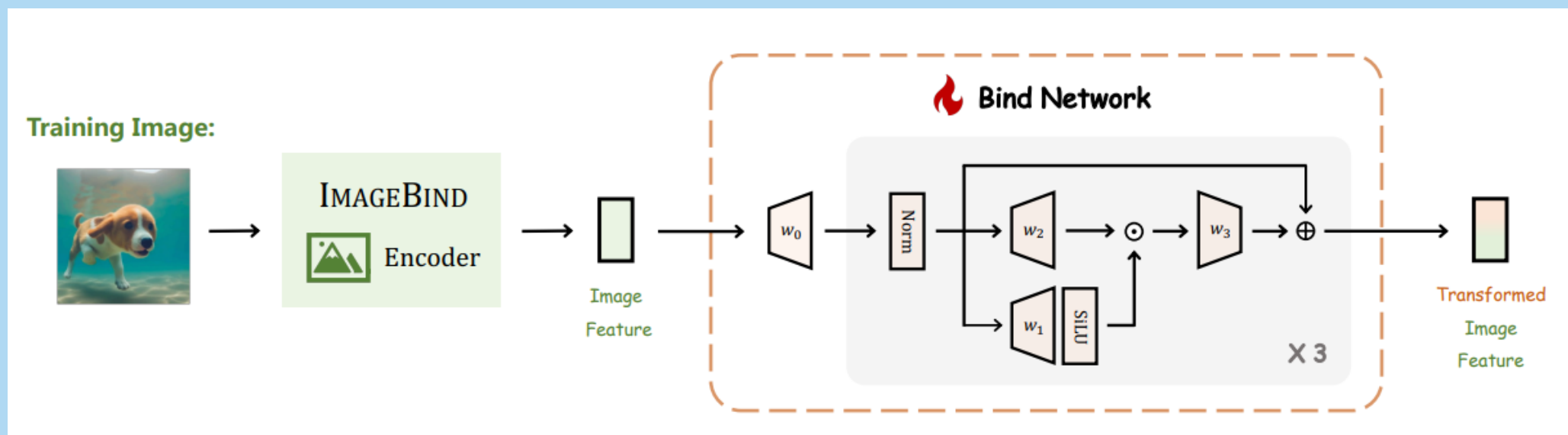
Mapping Network



- Linear
- 1DConv
- Transformer
- Residual
- ...

Mapping Network

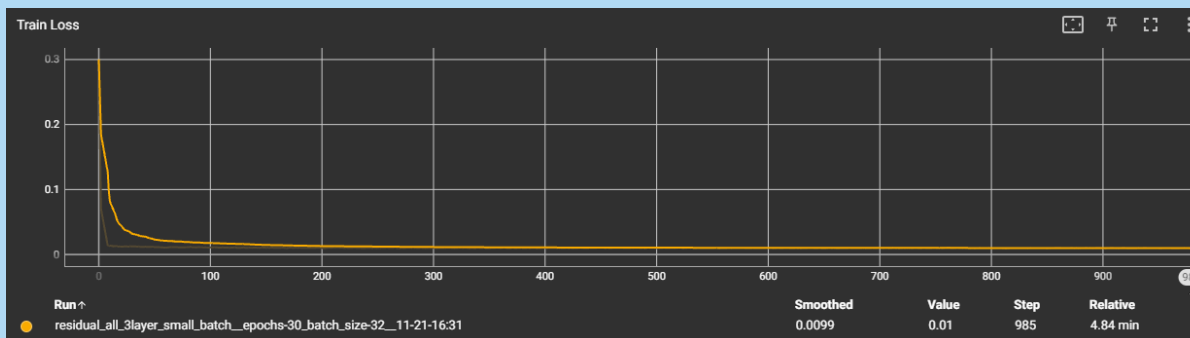
ImageBind-LLM: Multi-modality Instruction Tuning



목적이나 사용하는 방법은 본 task와는 많이 다르지만
실험 결과 단순 Linear model보다 학습 속도, cost면에서 낮다고 판단

Mapping Network

Input이 달라도 출력이 거의 같은 현상



강아지 짖는 소리



새소리



자동차 경적 소리

Mapping Network

모델 구조 변경하고 배치를 줄여도 비슷한 현상 발생

새소리



사이렌 소리



장작 타는 소리



트럭 소리



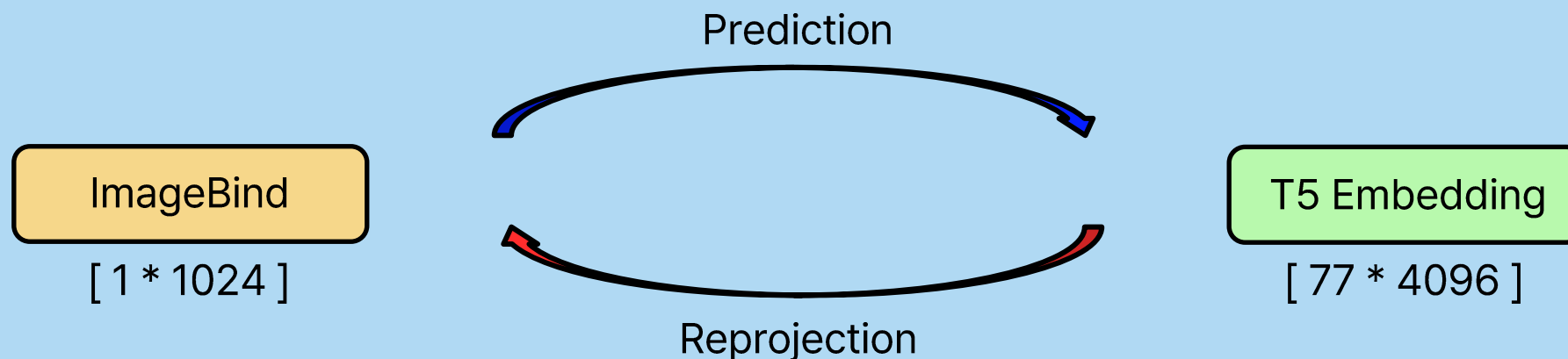
Mapping Network

모든 입력에 대해 Loss 값을 적당히 작게 만드는 특정 값으로 수렴...?

Mapping Network

Mutual Information

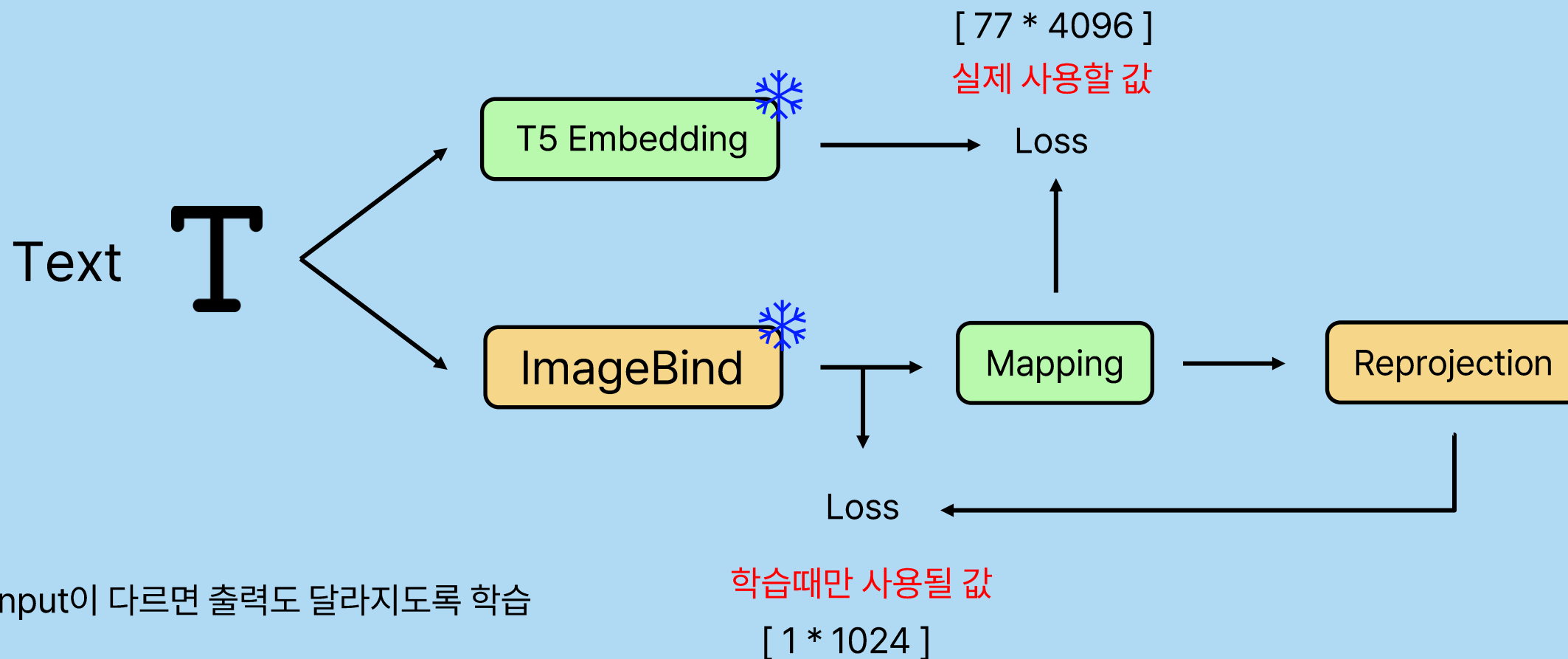
문제점: T5의 값이 ImageBind의 값과 관계없이 항상 비슷하게 나온다.



T5의 값을 서로 다르게 만들어야 확실하게 서로 다른
Imagebind 출력 값을 다시 만들 수 있다.

Mapping Network

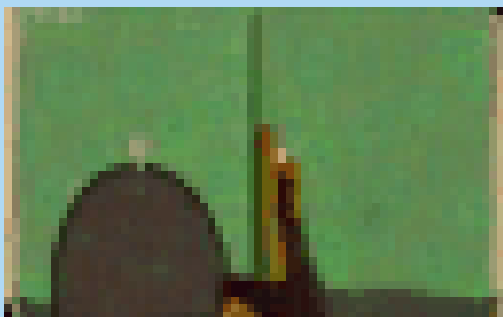
Mutual Information



Mapping Network

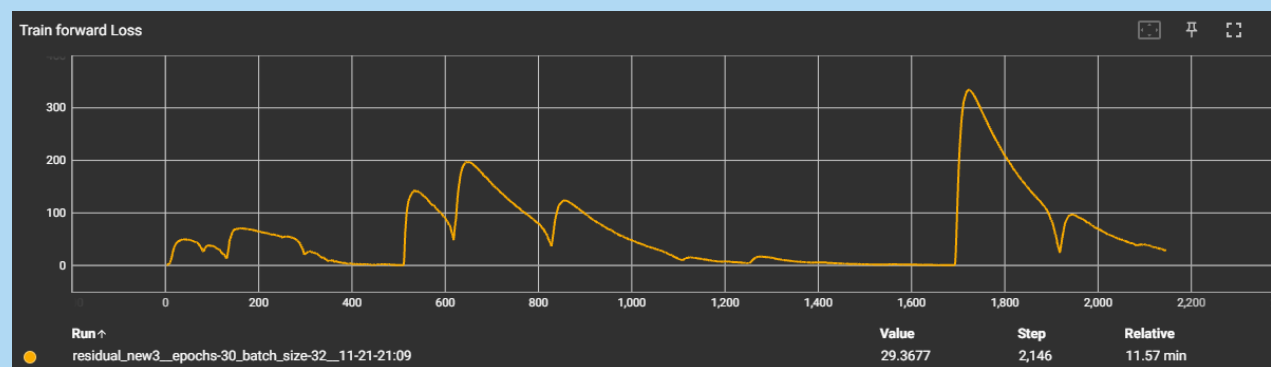
Mutual Information

학습 초기부터 다른 입력에 대해선 확실히 다른 출력을 보임

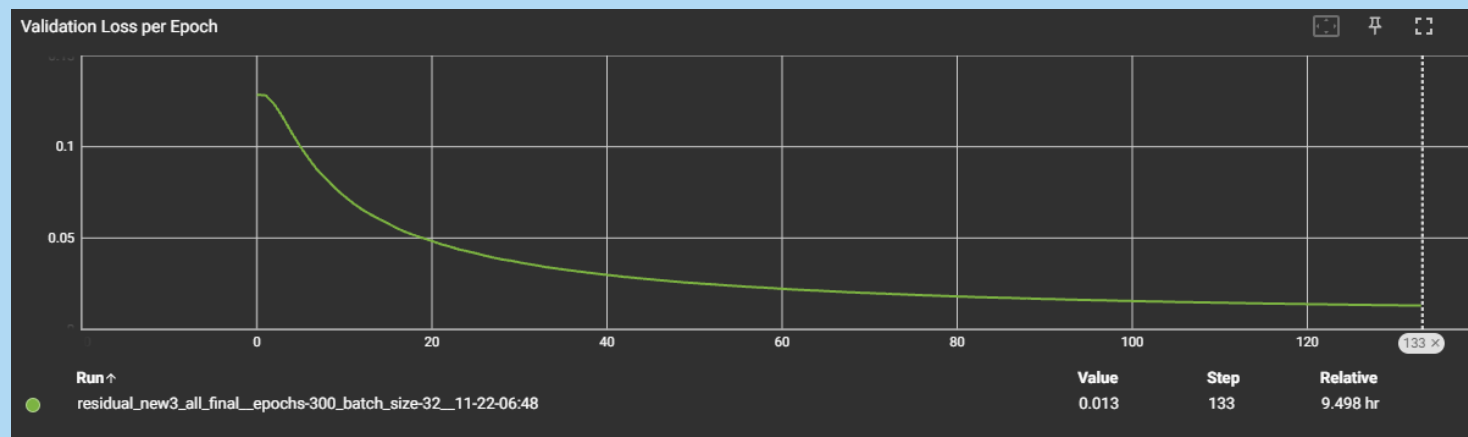


Mapping Network

Mutual Information

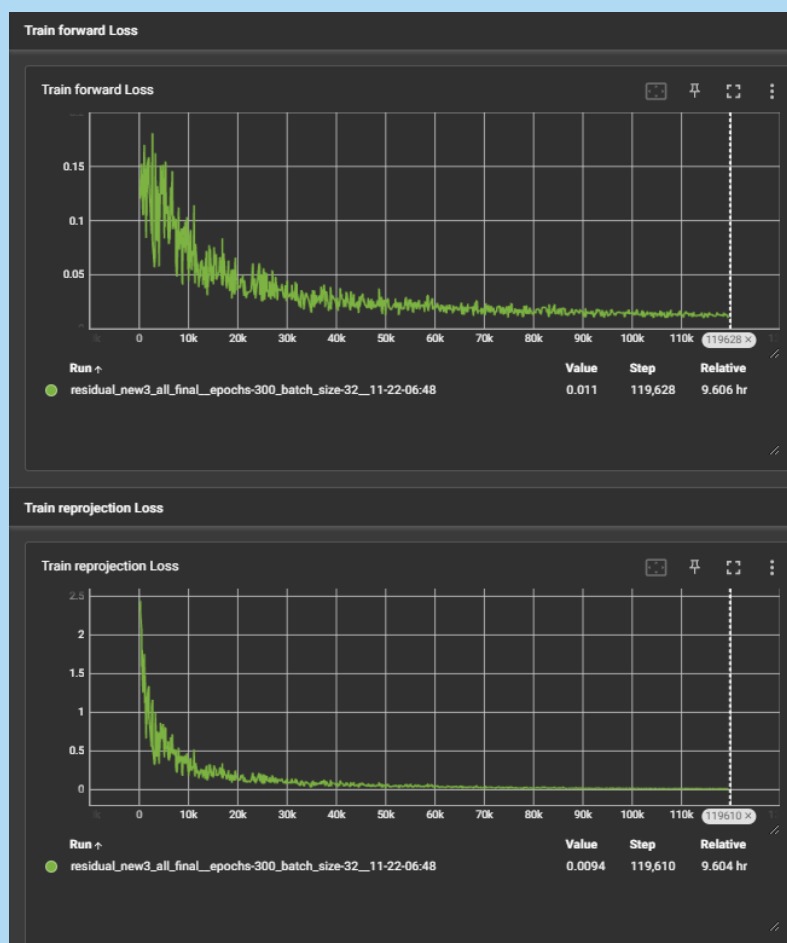


많은 시도 후..



Mapping Network

Mutual Information



Text, Audio, Image 다 포함한 데이터

약 3만 건으로 130 epoch 학습

Mapping Network

Mutual Information

최종 학습 결과 - Audio



Mapping Network

Mutual Information

최종 학습 결과 - Image



Mapping Network

Mutual Information

최종 학습 결과 - Text



Beautiful lake aerial view_base



Happy family using laptop on bed at home

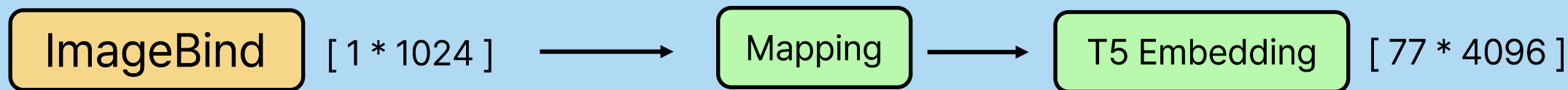


Beautiful young woman runs up_base



4k. time lapse view cityscape at bangkok city thailand_base

1. 큰 차원의 괴리









Token-wise Embedding이 고려될 수 없음

차원의 크기 차이가 너무 심함 (308배)


Discussion

2. ImageBind 자체의 문제


1) Cross-Modal Retrieval

Audio	Images & Videos	Depth	Text
 Crackle of a Fire			"A fire crackles while a pan of food is frying on the fire." "Fire is crackling then wind starts blowing." "Firewood crackles then music..."
 Baby Cooing			"A baby is crying while a toddler is laughing." "A baby is laughing while an adult is laughing." "A baby laughs and something..."

2) Embedding-Space Arithmetic

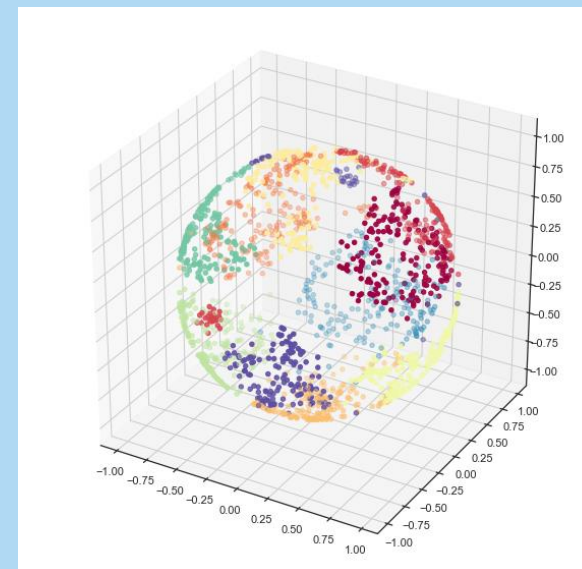
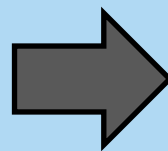
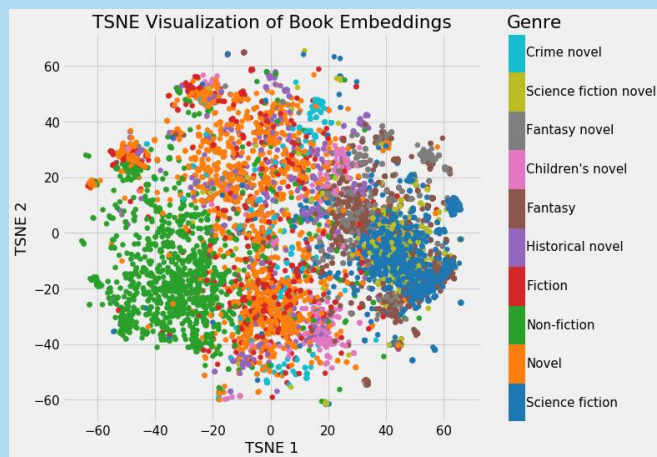


3) Audio to Image Generation



Retrieval 할 수 있을 정도로만
텍스트와 일대일 매칭시킬 수 있을 정도의 정확성은 부족하다.

3. Embedding Space mapping의 어려움



모델을 통해 완벽하게 매핑시키기 위해선 임베딩 모델을 만들 때 사용했던 거의 모든 데이터가 필요

임베딩의 차원이 조금 더 낮은 Video Generation 모델이 있었다면

최대한 다양한 분포의 데이터를 구할 수 있었다면



VGGnet

Video Graphic Generation network