

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

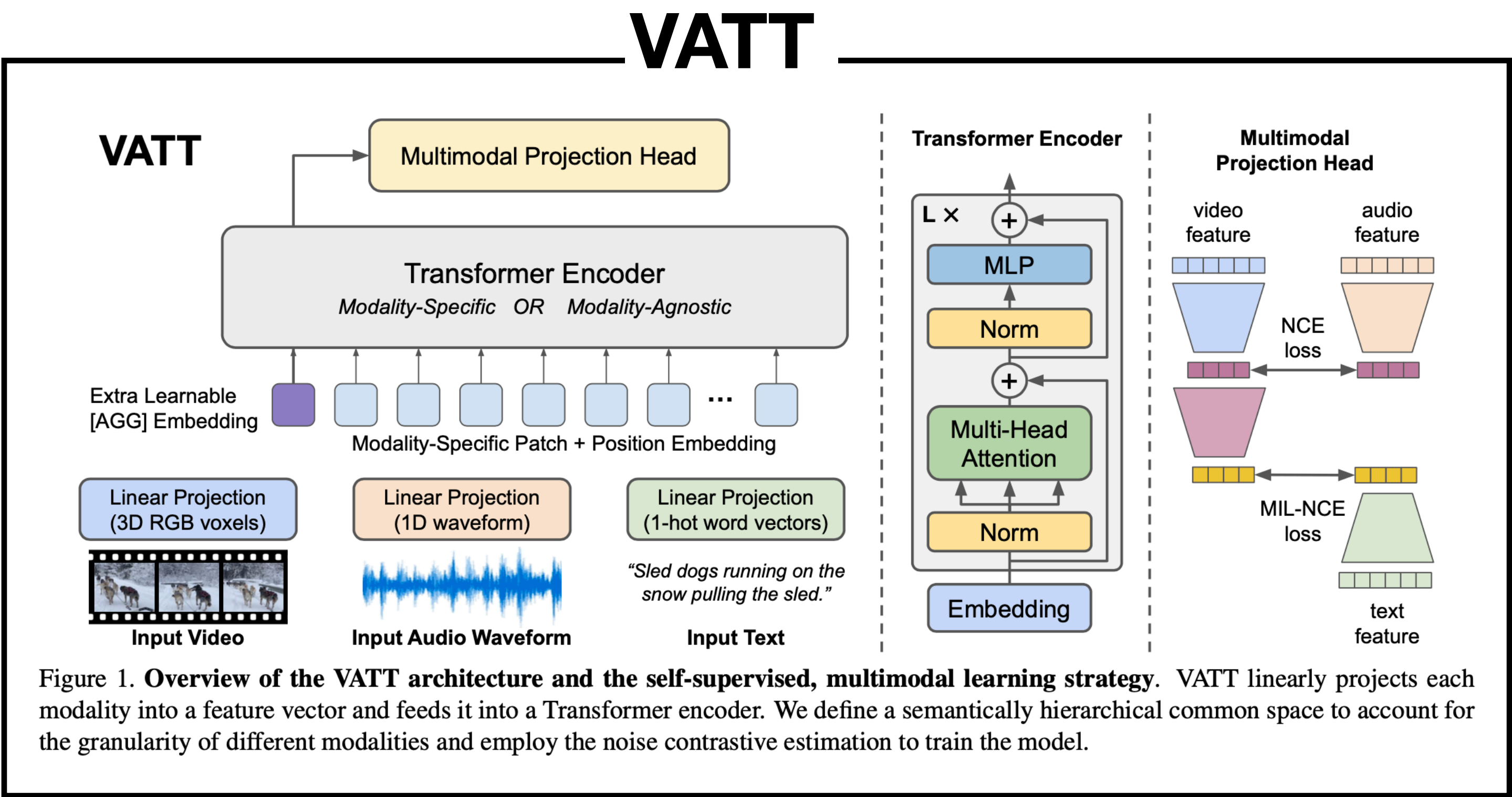
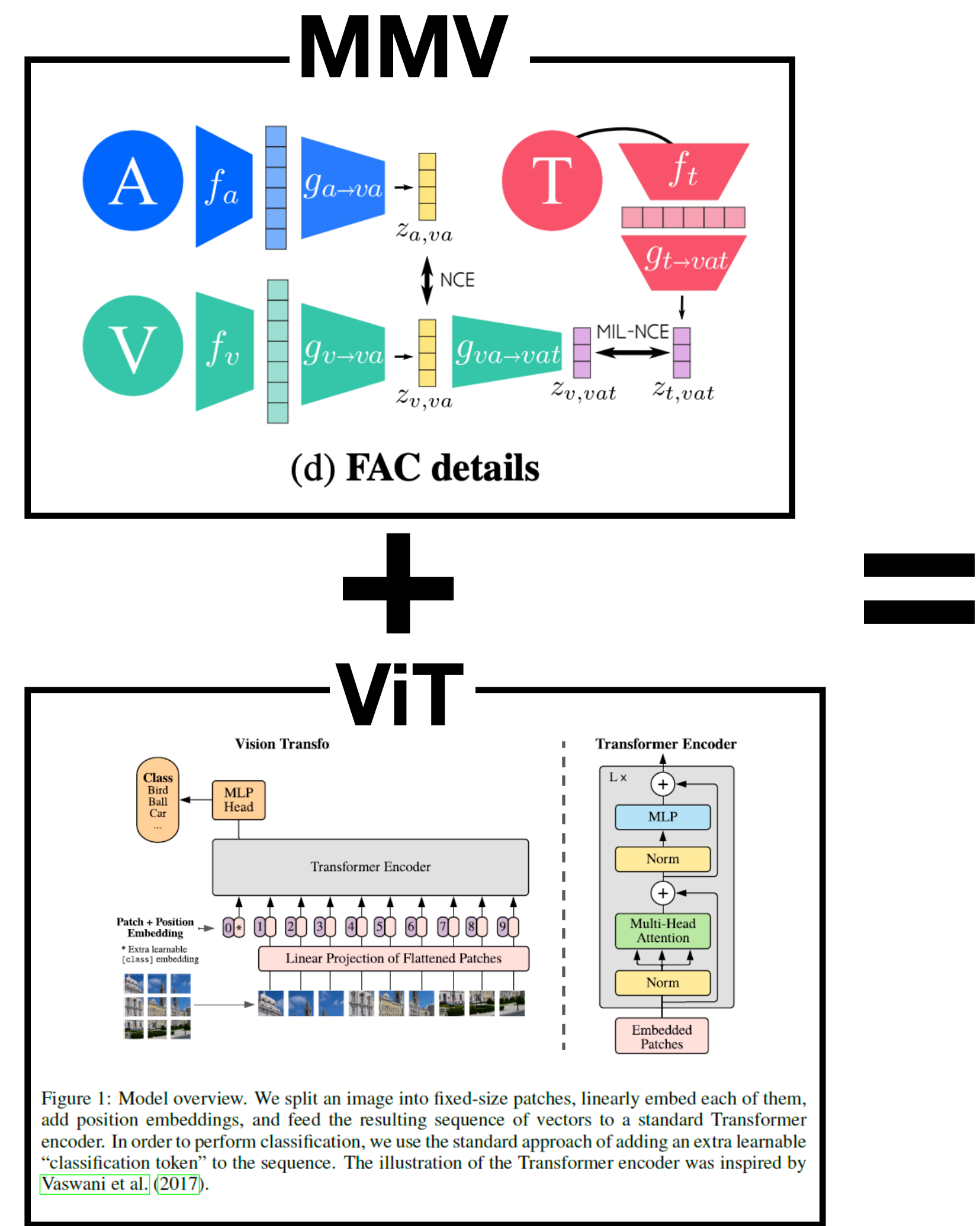
H. Akbari et. al.

Google, Columbia Univ., Cornell Univ.

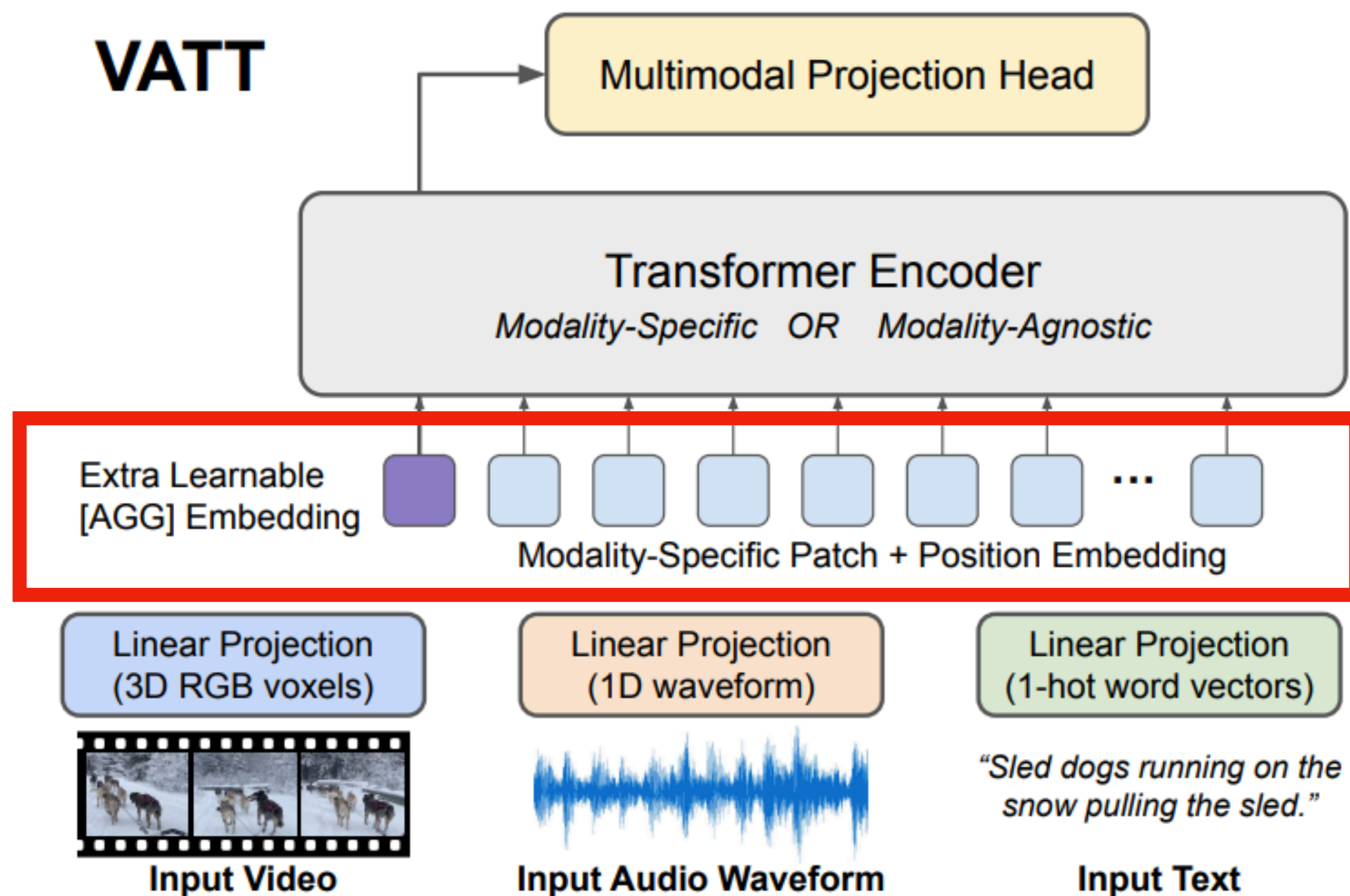
Summary

- Multi modality : Video, Audio, and Text
- Transformer
- Self-supervised learning
- Downstream tasks :
 - video action recognition,
 - audio event classification,
 - zero-shot video retrieval,
 - image classification

Overview

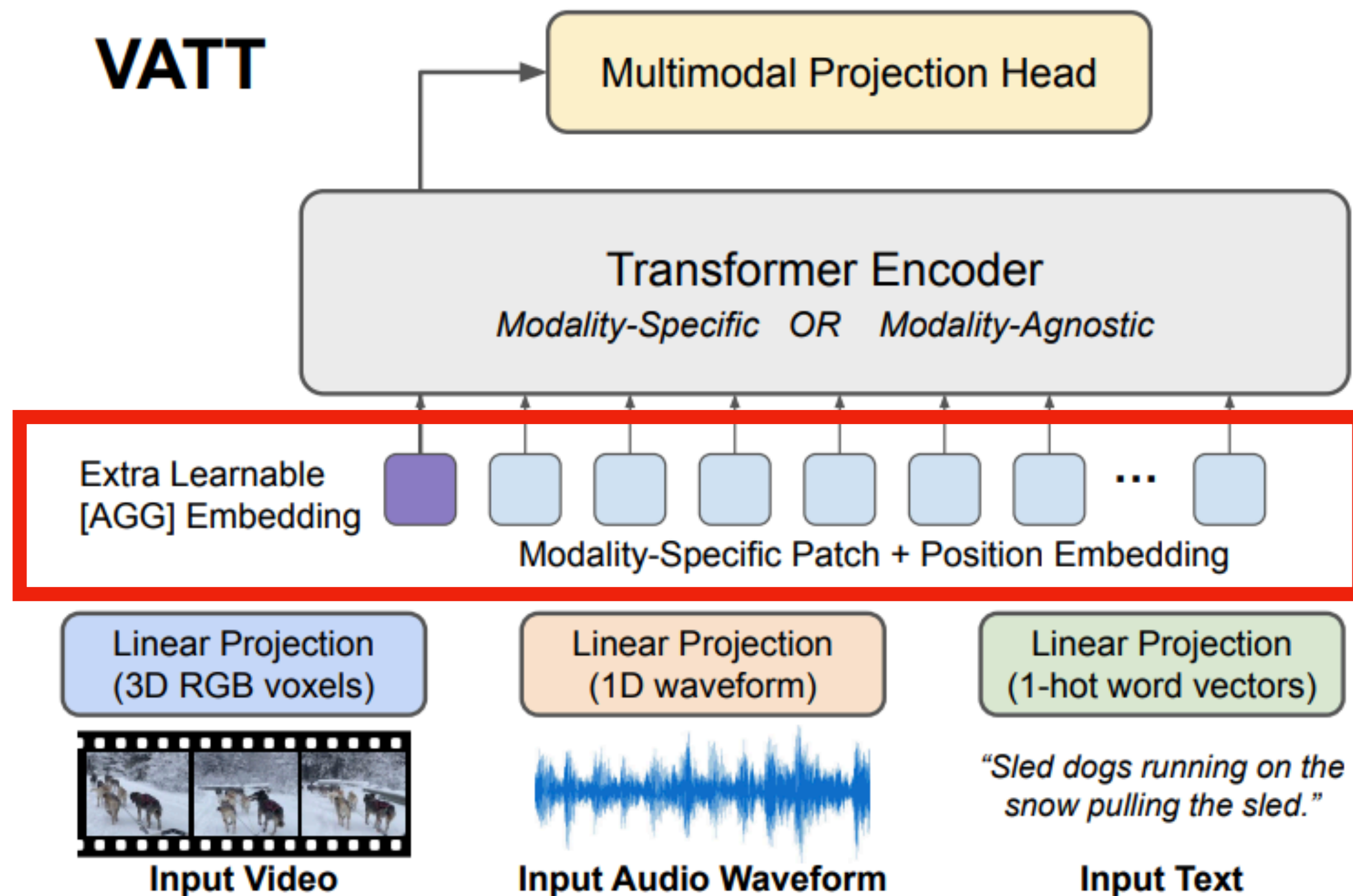


Tokenization



- **Video** : $(T, H, W, 3)$
 - $(t, h, w, 3)$ Patch
 - flatten, linear mapping \rightarrow d차원
- **Audio** : $(T, 1)$
 - $(t, 1)$
 - linear mapping \rightarrow d차원
- **Text** : v차원의 one-hot vector
 - linear mapping \rightarrow d차원
- AGG : aggregation token == CLS

Positional Encoding (Learnable)



- **Video :**
 - Horizontal, Vertical, Temporal 축에 대해 각각 positional encoding한 결과를 합함
- **Audio :**
 - positional encoding
- **Text :**
 - relative positional encoding 사용

DropToken

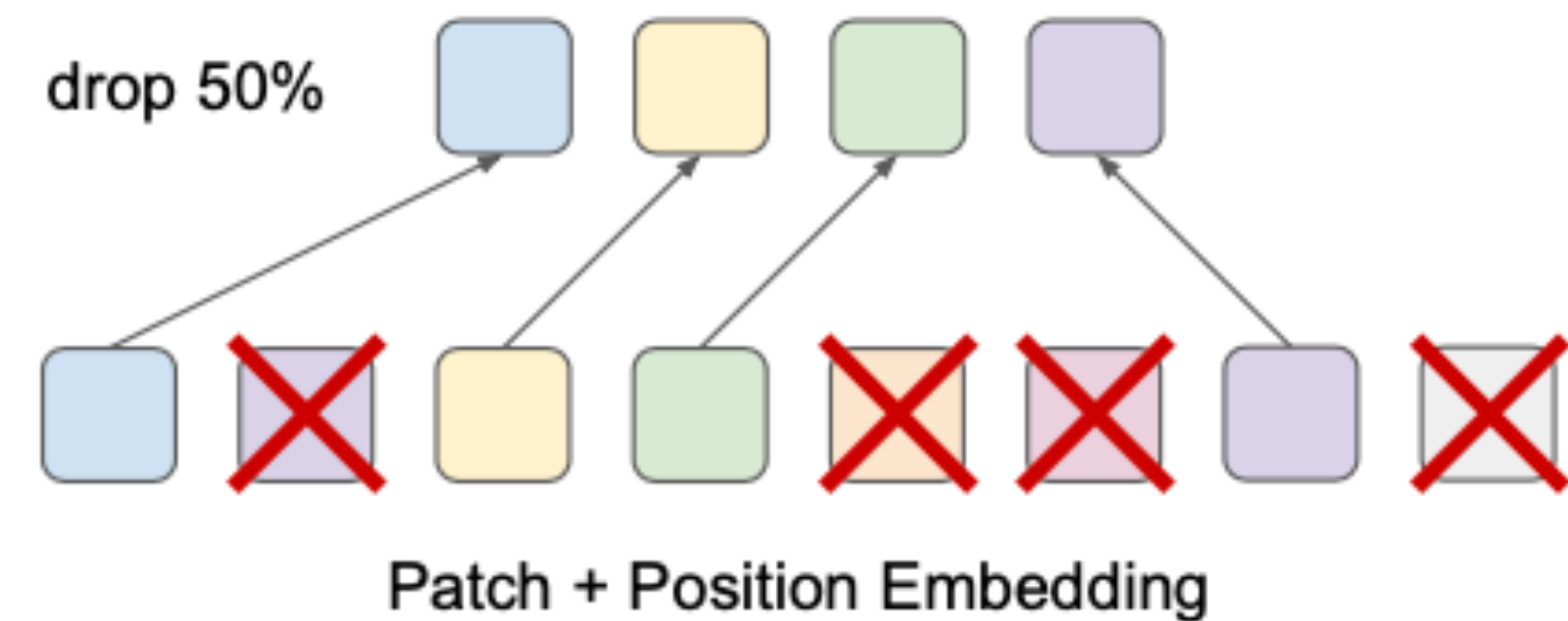
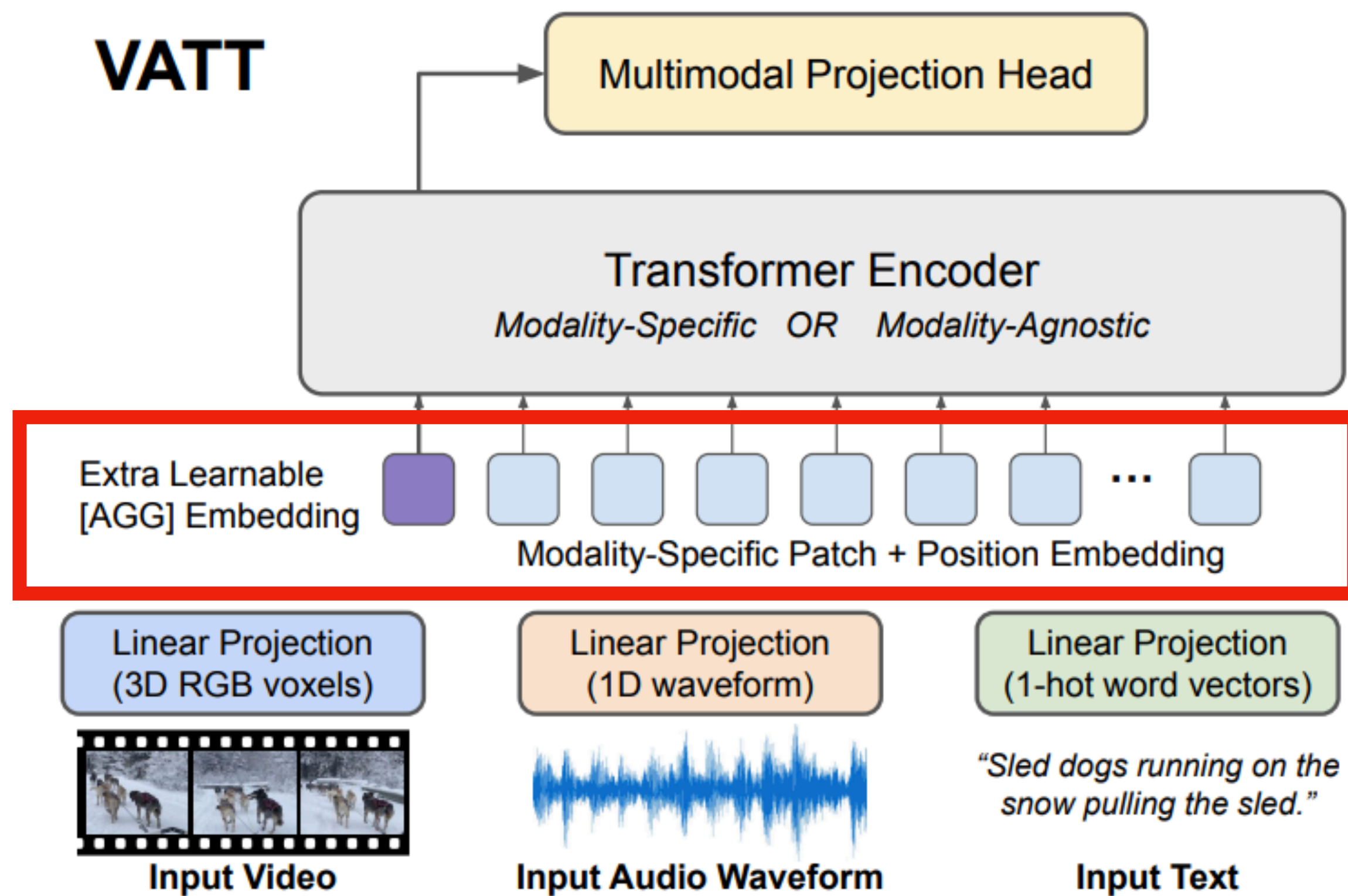
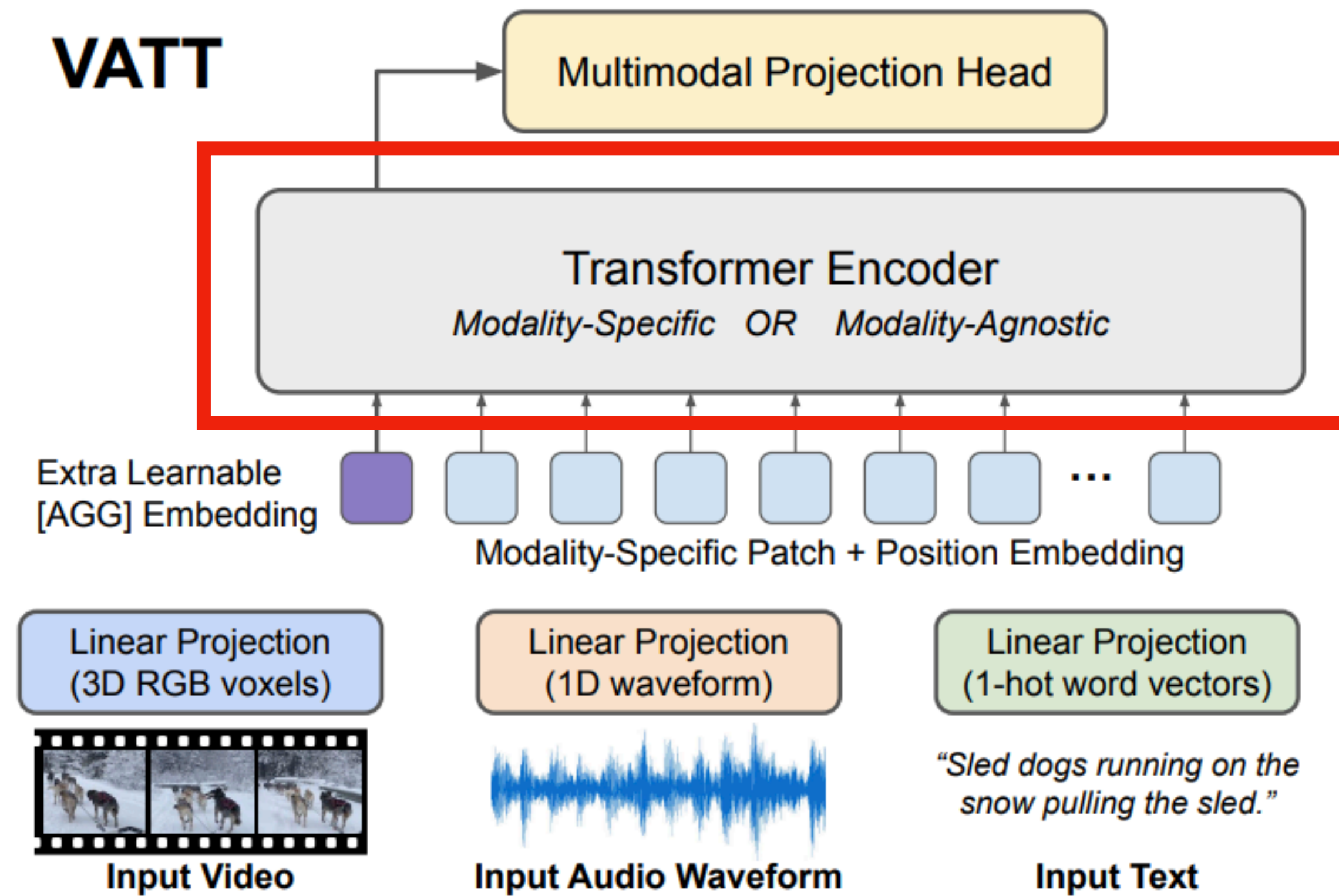


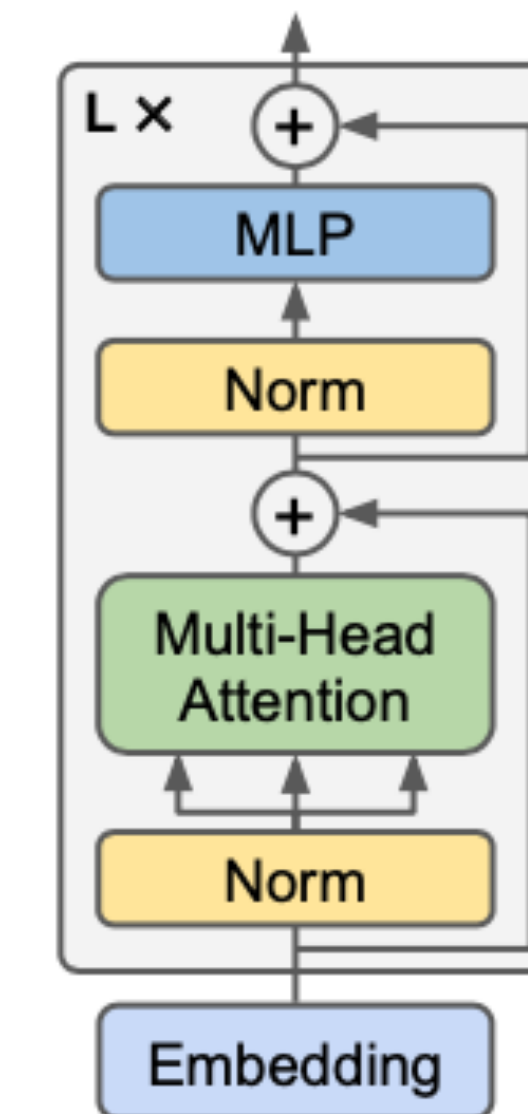
Figure 2. **DropToken**. During training, we leverage the high redundancy in multimodal video data and propose to randomly drop input tokens. This simple and effective technique significantly reduces training time with little loss of quality.

- video, audio의 token를 랜덤하게 drop

Transformer Encoder

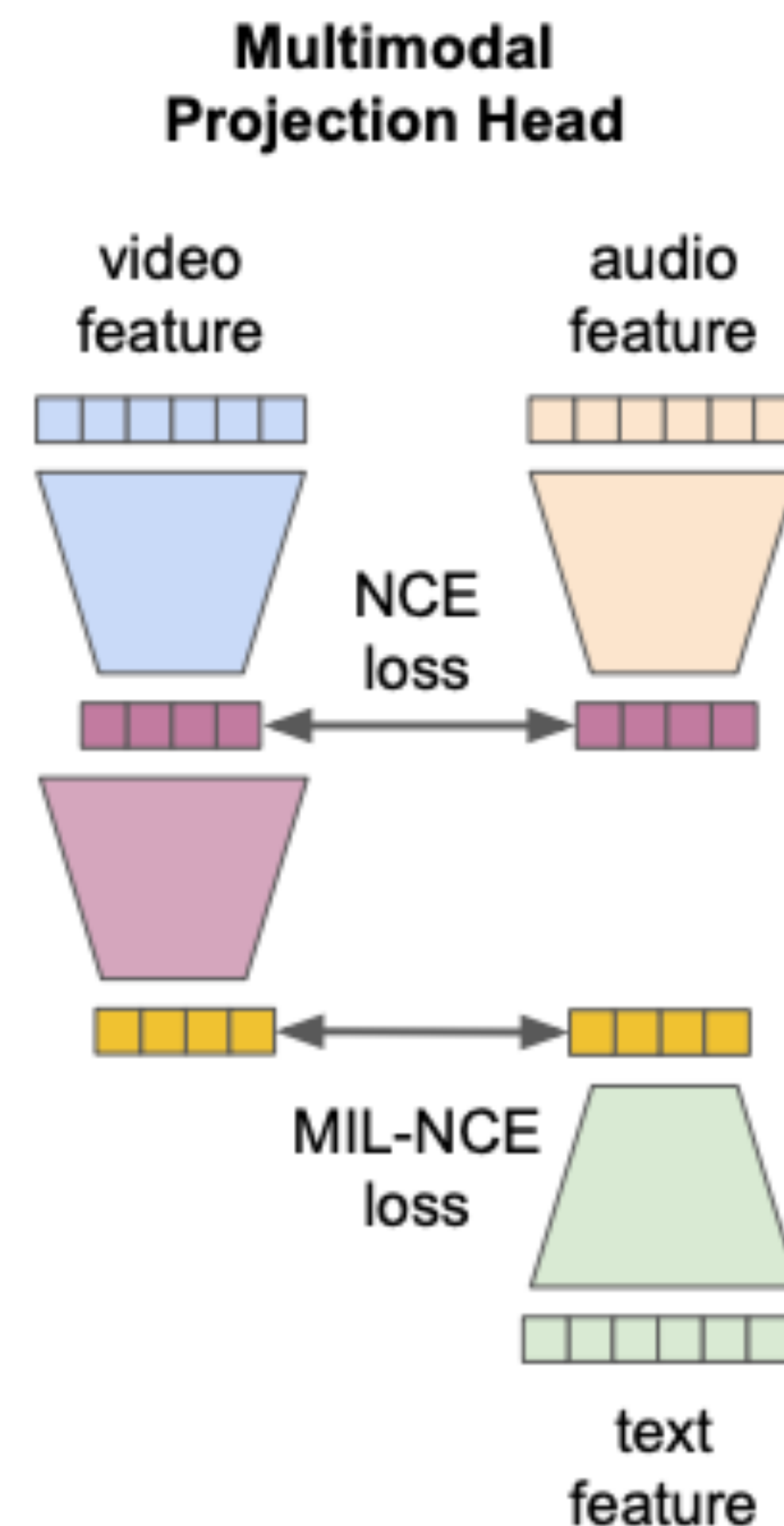
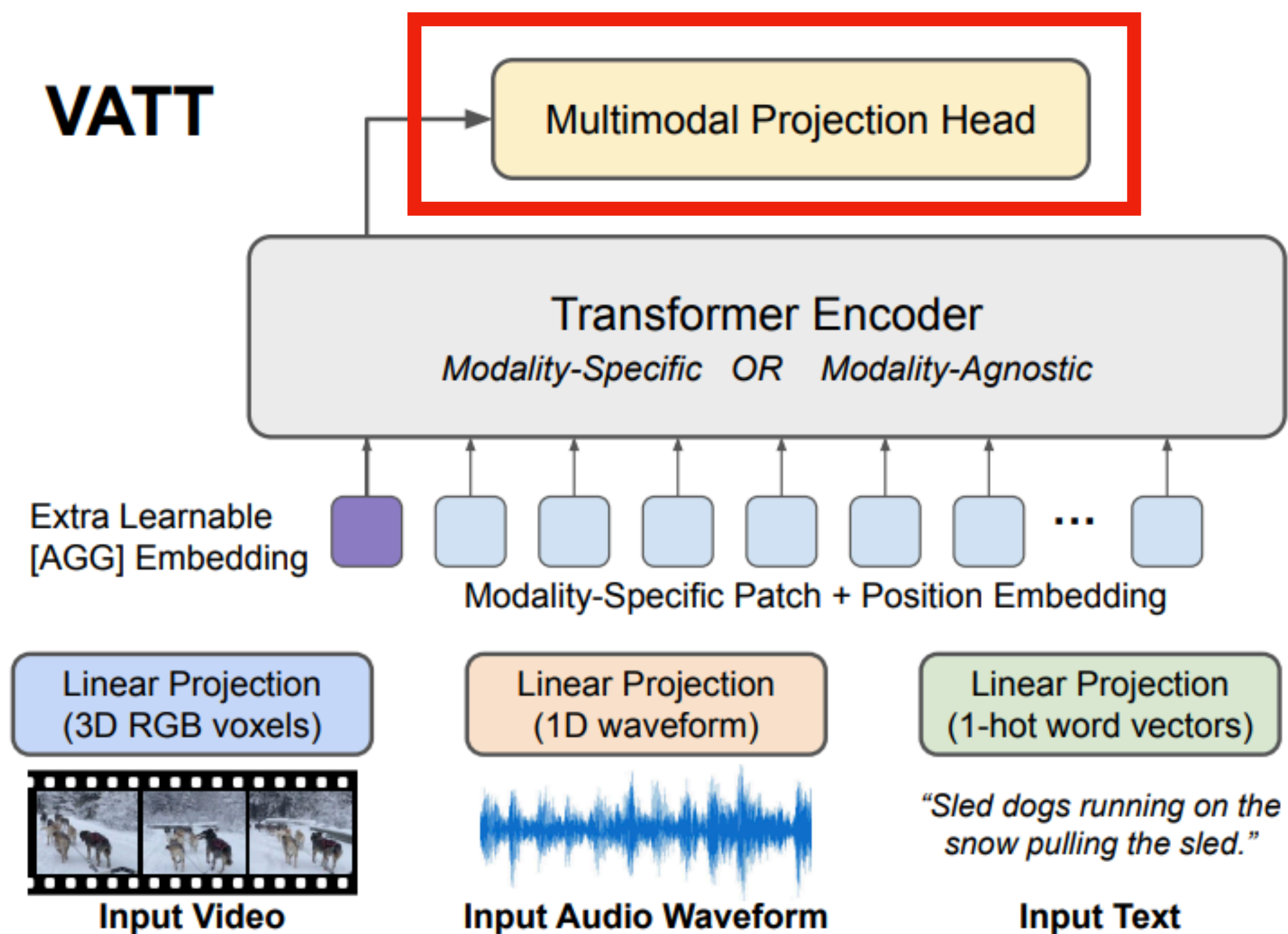


Transformer Encoder




- Modality-Specific : 3개의 Transformer
- Modality-Agnostic : 1개의 Transformer

Projection Head & Loss



- video : 2-layer (d-512)
- audio : 1 layer (d-512)
- text : 1 layer (d-256)

Projection Head & Loss



\mathcal{P} Positive candidates


.60 it's quite a simple technique for

.53 beginners to learn and basically all I

.63 do is squeeze out three little circles

.49 then with the back of a teaspoon

.47 simply press the teaspoon into the



\mathcal{P} Positive candidates


.50 main body of the laptop cover the

.63 duct tape with aluminum cover all

.61 remaining gaps edges with aluminum

.56 tape use the leftover poster board to

.50 create the keyboard keys I made my



\mathcal{P} Positive candidates

.67 spinach what's the name

.57 keep it simple you just want to add

.58 fresh herbs maybe some oregano

.59 you can add cilantro basil they give

.50 it a couple more copies and when you

Figure 3: Selected video and narration pairs from five positive candidates on HowTo100M held-out samples using MIL-NCE.

$$\text{NCE}(\mathbf{z}_{v,va}, \mathbf{z}_{a,va}) = -\log \left(\frac{\exp(\mathbf{z}_{v,va}^\top \mathbf{z}_{a,va} / \tau)}{\sum_{i=1}^B \exp(\mathbf{z}_{v,va}^{i\top} \mathbf{z}_{a,va}^i / \tau)} \right), \quad (4)$$

$$\text{MIL-NCE}(\mathbf{z}_{v,vt}, \{\mathbf{z}_{t,vt}\}) = -\log \left(\frac{\sum_{\mathbf{z}_{t,vt} \in \mathcal{P}(\mathbf{z}_{v,vt})} \exp(\mathbf{z}_{v,vt}^\top \mathbf{z}_{t,vt} / \tau)}{\sum_{\mathbf{z}_{t,vt} \in \mathcal{P}(\mathbf{z}_{v,vt}) \cup \mathcal{N}(\mathbf{z}_{v,vt})} \exp(\mathbf{z}_{v,vt}^\top \mathbf{z}_{t,vt} / \tau)} \right), \quad (5)$$

$$\mathcal{L} = \text{NCE}(\mathbf{z}_{v,va}, \mathbf{z}_{a,va}) + \lambda \text{MIL-NCE}(\mathbf{z}_{v,vt}, \{\mathbf{z}_{t,vt}\}), \quad (6)$$

Pre-training Datasets

- Howto100M : 1.2M videos (multiple clips with audio, scripts), 136M video-audio-text
- AudioSet : 10s clips sample from 2M videos from Youtube (video + audio) none-text

Downstream Tasks and Datasets

- **Video action recognition :**
 - UFC101 (101 classes, 13,320 videos)
 - HMDB51 (51 classes, 6,766 videos)
 - Kinetics-400 (400 classes, 234,584 videos)
 - Kinetics-600 (600 classes, 366,016 videos)
 - Moments in Time (339 classes, 791,297 videos)
- **Audio event classification :**
 - ESC50 (50 classes, 2000 audio clips)
 - AudioSet (527 classes, 2M audio clips)
- **Zero-shot video retrieval :**
 - YouCook2 (3.1k video-text pairs)
 - MSR-VTT (1k video-text pairs)
- **Image classification :**
 - ImageNet-1000k (1000 classes, 1.2M)

Network setup in VATT

Model	Layers	Hidden Size	MLP Size	Heads	Params
Small	6	512	2048	8	20.9 M
Base	12	768	3072	12	87.9 M
Medium	12	1024	4096	16	155.0 M
Large	24	1024	4096	16	306.1 M

Table 1. Details of the Transformer architectures in VATT.

- **Modality-agnostic** : Medium model
- **Modality-specific** :
 - video : audio : text = Base : Base : Small (BBS)
 - video : audio : text = Medium : Base : Small (MBS)
 - video : audio : text = Large : Base : Small (LBS)

Details

- **Video**
 - train : 32 x 224 x 224 x 3 (10fps) + augmentation
- **Audio**
 - train : Video랑 같은 부분 48kHz
- Patch size : 4 x 16 x 16 x 3 (video), 128 (audio)
- DropToken : 50%
- optimizer : adam, lr : 1e-4 to 5e-5 (quarter period cosine schedule)
- batch size 2048, 500k steps in total, TPU v3 256개로 3일 학습

Results

Video Action Recognition

METHOD	TOP-1	TOP-5	TFLOPs
ARTNet [98]	69.2	88.3	6.0
I3D [16]	71.1	89.3	-
R(2+1)D [30]	72.0	90.0	17.5
MFNet [60]	72.8	90.4	-
Inception-ResNet [2]	73.0	90.9	-
blVNet [32]	73.5	91.2	0.84
A ² -Net [22]	74.6	91.5	-
TSM [61]	74.7	-	-
S3D-G [102]	74.7	93.4	-
Oct-I3D+NL [21]	75.7	-	0.84
D3D [88]	75.9	-	-
GloRe [23]	76.1	-	-
I3D+NL [98]	77.7	93.3	10.8
ip-CSN-152 [92]	77.8	92.8	-
MoViNet-A5 [51]	78.2	-	0.29
CorrNet [17]	79.2	-	6.7
LGD-3D-101 [75]	79.4	94.4	-
SlowFast [34]	79.8	93.9	7.0
X3D-XXL [33]	80.4	94.6	5.8
TimeSFormer-L [10]	80.7	94.7	7.14
VATT-Base	79.6	94.9	9.09
VATT-Medium	81.1	95.6	15.02
VATT-Large	82.1	95.5	29.80
VATT-MA-Medium	79.9	94.9	15.02

Table 2. Results for video action recognition on Kinetics-400.

METHOD	TOP-1	TOP-5
I3D-R50+Cell [99]	79.8	94.4
LGD-3D-101 [75]	81.5	95.6
SlowFast [34]	81.8	95.1
X3D-XL [33]	81.9	95.5
TimeSFormer-HR [10]	82.4	96.0
MoViNet-A5 [51]	82.7	95.7
VATT-Base	80.5	95.5
VATT-Medium	82.4	96.1
VATT-Large	83.6	96.6
VATT-MA-Medium	80.8	95.5

Table 3. Results for video action recognition on Kinetics-600.

METHOD	TOP-1	TOP-5
TSN [97]	25.3	50.1
R3D-50 [82]	27.2	51.7
TRN [115]	28.3	53.4
I3D [16]	29.5	56.1
blVNet [31]	31.4	59.3
SRTG-R3D-101[87]	33.6	58.5
AssembleNet-101 [82]	34.3	62.7
MoViNet-A5 [51]	39.1	-
VATT-Base	38.7	67.5
VATT-Medium	39.5	68.2
VATT-Large	41.1	67.7
VATT-MA-Medium	37.8	65.9

Table 4. Results for video action recognition on Moments in Time.

- video action recognition에서 SOTA 달성
- model이 클 때 결과도 좋고 FLOPS도 증가
- model-agnostics도 충분히 경쟁력 있음

Results

Audio Event Classification

METHOD	mAP	AUC	d-prime
DaiNet [25]	29.5	95.8	2.437
LeeNet11 [59]	26.6	95.3	2.371
LeeNet24 [59]	33.6	96.3	2.525
Res1dNet31 [52]	36.5	95.8	2.444
Res1dNet51 [52]	35.5	94.8	2.295
Wavegram-CNN [52]	38.9	96.8	2.612
VATT-Base	39.4	97.1	2.895
VATT-MA-Medium	39.3	97.0	2.884

Table 5. Results for audio event classification on AudioSet.

- audio event classification에서 SOTA달성

Results

Image Classification

METHOD	PRE-TRAINING DATA	TOP-1	TOP-5
iGPT [18]	ImageNet	66.5	-
ViT-Base [29]	JFT	79.9	-
VATT-Base	-	64.7	83.9
VATT-Base	HowTo100M	78.7	93.9

Table 6. Finetuning results for ImageNet classification.

- unlabeled-data로 pre-train해도 준수한 성능

Results

Zero-shot Video Retrieval

METHOD	BATCH	EPOCH	YouCook2		MSR-VTT	
			R@10	MedR	R@10	MedR
MIL-NCE [64]	8192	27	51.2	10	32.4	30
MMV [1]	4096	8	45.4	13	31.1	38
VATT-MBS	2048	4	45.5	13	29.7	49
VATT-MA-Medium	2048	4	40.6	17	23.6	67

Table 7. Zero-shot text-to-video retrieval. We evaluate our text and video Transformers of the final pre-training checkpoint.

- batch size와 epoch을 키우면 더 성능이 좋아짐을 확인

Results

LRC & SVM

METHOD	UCF101	HMDB51	ESC50
MIL-NCE [64]	83.4	54.8	-
AVTS [54]	-	-	82.3
XDC [3]	-	-	84.8
ELo [73]	-	64.5	-
AVID [84]	-	-	89.2
GDT [71]	-	-	88.5
MMV [1]	91.8	67.1	88.9
VATT-Medium + SVM	89.2	63.3	82.5
VATT-Medium + LRC	89.6	65.2	84.7
VATT-MA-Medium + LRC	84.4	63.1	81.2

Table 8. Linear evaluation results for video action recognition on UCF101 and HMDB51 and audio event classification on ESC50. MA refers to the modality-agnostic backbone.

- fine-tune을 하지 않아도 성능이 준수함
- pre-train 과정에서 의미있는 feature를 학습

Results

Visualization

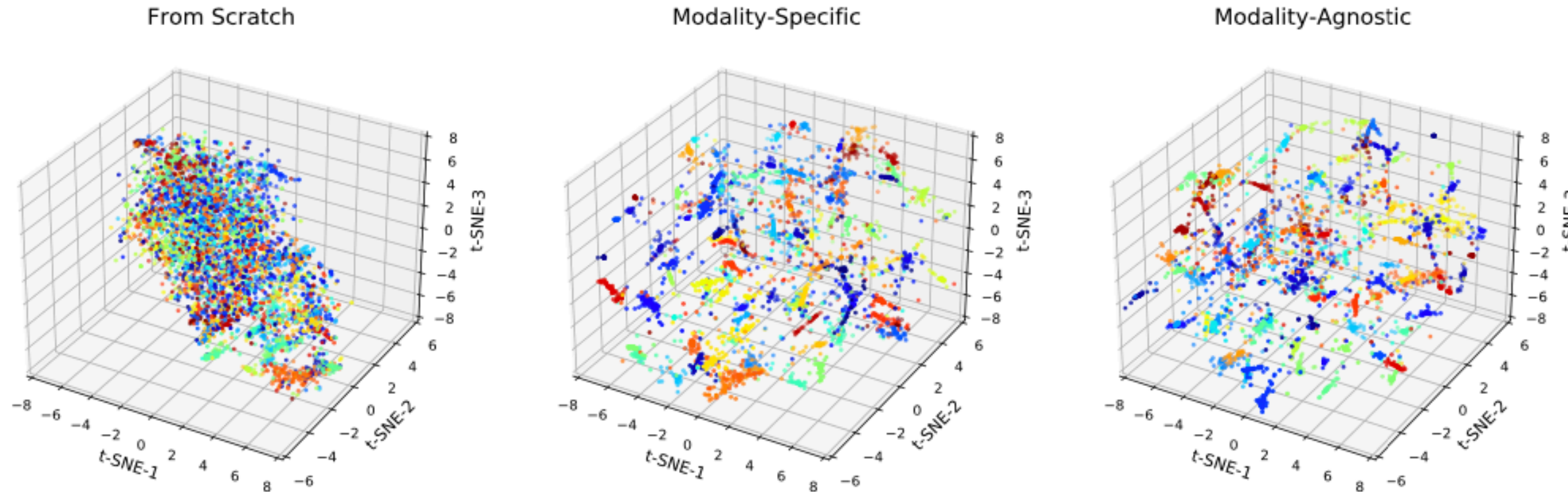


Figure 3. t-SNE visualization of the feature representations extracted by the vision Transformer trained from scratch on Kinetics-400 validation set, the modality-specific VATT's vision Transformer after fine-tuning, and the modality-agnostic Transformer after fine-tuning. For better visualization, we show 100 random classes from Kinetics-400.

- 하나의 모델로도 잘 구분한 결과 보여줌
- domain에 상관없이 유사한 것끼리 임베딩

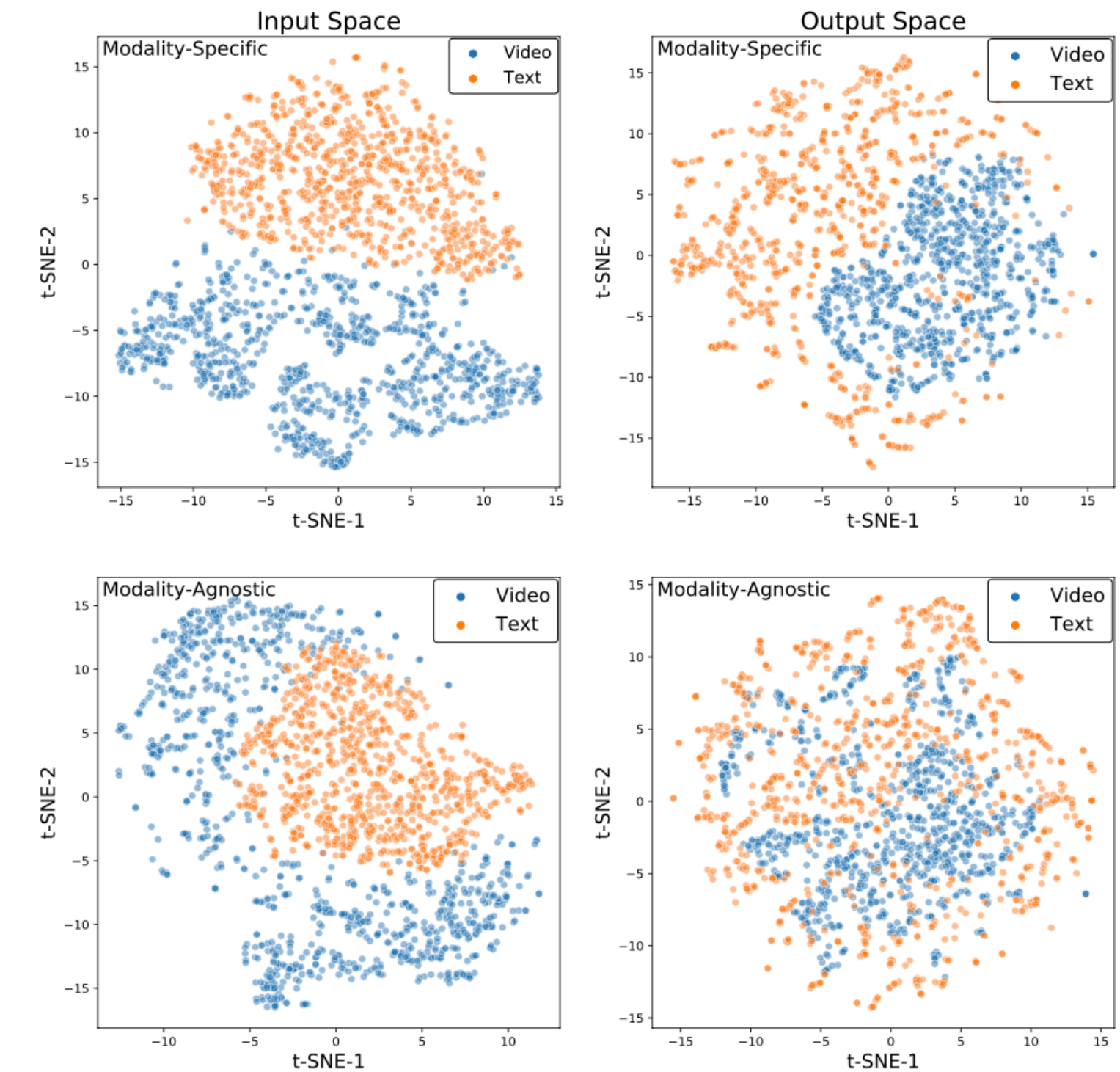


Figure 4. t-SNE visualization of the input space vs. output space for modality-specific and modality-agnostic backbones when different modalities are fed.

Ablation Study

DropToken

	DropToken Drop Rate			
	75%	50%	25%	0%
Multimodal GFLOPs	188.1	375.4	574.2	784.8
HMDB51	62.5	64.8	65.6	66.4
UCF101	84.0	85.5	87.2	87.6
ESC50	78.9	84.1	84.6	84.9

Table 13. Linear classification top-1 accuracy vs. sampling rate vs. inference GFLOPs in the Medium-Base-Small (MBS) setting.

Resolution/ FLOPs	DropToken Drop Rate			
	75%	50%	25%	0%
32 × 224 × 224	-	-	-	79.9
Inference (GFLOPs)	-	-	-	548.1
64 × 224 × 224	-	-	-	80.8
Inference (GFLOPs)	-	-	-	1222.1
32 × 320 × 320	79.3	80.2	80.7	81.1
Inference (GFLOPs)	279.8	572.5	898.9	1252.3

Table 14. Top-1 accuracy of video action recognition on Kinetics400 using high-resolution inputs coupled with DropToken vs. low-resolution inputs.

Conclusion

- Self-supervised multimodal representation learning framework
- Using basic Transformer
- Proposed DropToken
- SOTA in video action recognition and audio event classification

Reference

- MIL-NCE : <https://arxiv.org/pdf/1912.06430.pdf>
- Multimodal Versatile Network : <https://arxiv.org/pdf/2006.16228v2.pdf>
- ViT : <https://arxiv.org/abs/2010.11929>
- VATT : <https://arxiv.org/pdf/2104.11178.pdf>

Pre-training setup

- optimizer : Adam,
 - initial lr : $1e-4$
 - 10k warmup steps
 - 500k steps in total
 - quarter-period cosine schedule ($1e-4$ to $5e-5$)

Setup

Video fine-tuning setup

- optimizer : SGD
 - momentum : 0.9
 - initial rate : 0.005
 - half-period cosine schedule (0.005 to 0)
 - 2.5k warmup steps, 100k steps
- batch size : 64
- label smoothing 0.1
- video frame resolution : 320 x 320
- uniformly sample 4 clips to cover the entire 10s

Setup

Audio fine-tuning setup

- optimizer : SGD
 - momentum : 0.9
 - initial lr : 0.2
 - half-period cosine schedule (0.2 to 0)
 - 2.5k warmup steps, 50k steps in total
- batch size : 1,024
- mixup augmentation
- 6.4s with 24kHz

Setup

Image fine-tuning setup

- optimizer : SGD
 - momentum : 0.9
 - initial lr : $8e-2$ (cosine learning rate decay)
- batch size : 512
- 50 epochs
- label smoothing 0.1
- no weight decay
- 384 x 384 input resolution

Setup

Linear evaluation setup

- optimizer : adam
- initial lr : $5e-4$
- batch size : 64
- 50k training steps

Setup

zero-shot retrieval setup

- 224 x 224 central crops for 32 frames with a temporal stride of 2 at 25fps