

LLMs can see and hear
without any training

MILS

- Multimodal Iterative LLM Solver
- training을 하지 않고 multimodal 을 구현한다는 점에서 매우 신기
- 개념이 어렵지는 않음

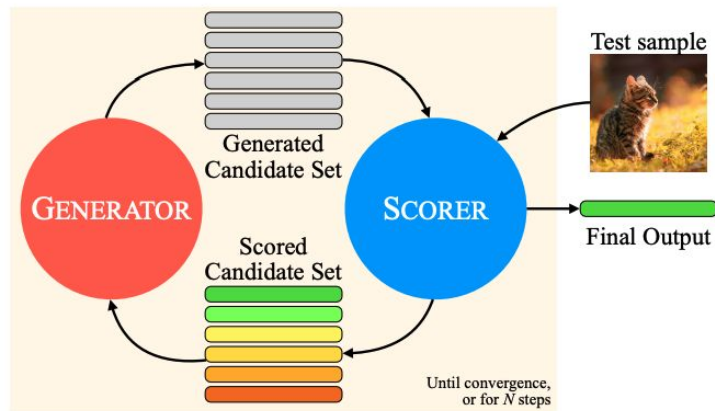
MILS

- Generator
 - Task를 풀기 위한 Candidate outputs를 생성하는 역할을 함(LLM)
 - Text를 input으로 받아서, task를 포함한 description을 생성함
- Scorer
 - Generator가 생성한 description과 input image를 쌍으로 받아서 text와 image 간의 score를 뽑아냄(CLIP 같은 image-text multimodal 을 사용할 수도 있음)

MILS

- 과정

1. Generator가 text를 생성
2. Scorer가 image와 생성된 text와의 score를 측정
3. Score 가 높은 것들 몇 가지를 선택
4. 선택된 Score와 생성된 text의 관계를 Generator에 입력
5. 2 ~ 4를 반복
6. 특정 조건(score, number of iteration)을 만족하면 final output을 받음



성능




			
MILS	<i>Boston Terrier's sharp teeth firmly gripping a rope toy tightly.</i>	<i>Close-up of a bird near a ship's cabin window.</i>	<i>Cluttered desk with a large monitor and a window.</i>
MeaCap (Zeng et al., 2024)	<i>The image depicts that dog licking a man who is in inappropriately close proximity.</i>	<i>The image depicts that observation boat in the water below us, " he said.</i>	<i>The image depicts that workspace during an cluttered desk with a large monitor and a online career as a consultant.</i>

Figure 3: **Image Captioning using MILS**, compared to existing state-of-the-art zero-shot approach, MeaCap (Zeng et al., 2024). MILS, while being a much simpler approach, produces more accurate and syntactically correct captions to the image.

Method	BLEU ₄	CIDEr	METEOR	SPICE
ZeroCap (Tewel et al., 2022)	2.6	14.6	11.5	5.5
ConZIC (Zeng et al., 2023)	1.3	13.3	11.2	5.0
CLIPRe (Li et al., 2023c)	4.6	25.6	13.3	9.2
MeaCap _{TF} (Zeng et al., 2024)	7.1	42.5	16.6	11.8
MeaCap _{TF} [*] (Zeng et al., 2024)	4.5	26.0	14.1	9.4
MILS	8.0	33.3	15.0	9.6

Table 1: **Zero-shot image captioning** on MSCOCO (Karpathy & Fei-Fei, 2015). Despite being far simpler than existing approaches, MILS performs competitively on all automatic metrics, and especially METEOR and SPICE which take into account the semantic similarity. ^{*} refers to results we obtained by running the provided code.

성능

Method	Training Data	CIDEr	METEOR
Nagrani <i>et al.</i> (Nagrani et al., 2022)	HowTo100M (Miech et al., 2019)	0.5	8.23
Nagrani <i>et al.</i> (Nagrani et al., 2022)	VideoCC3M (Nagrani et al., 2022)	8.2	11.3
MILS	–	2.3	14.4

Table 2: **Zero-shot video captioning** on MSR-VTT (Xu et al., 2016). MILS outperforms (Nagrani et al., 2022) when trained on HowTo100M, and is competitive to it when trained on the much cleaner VideoCC3M dataset, outperforming it on METEOR. We grayed (Nagrani et al., 2022) since it is trained for video captioning, while MILS is not.

Method	BLEU ₄	ROUGE _L	METEOR	SPICE
ZerAuCap (Salewski et al., 2023)	2.9	25.4	9.4	5.3
MILS	2.7	23.1	12.4	7.6

Table 3: **Zero-shot audio captioning** on Clotho (Drossos et al., 2020) dataset. MILS performs competitively to the existing zero-shot audio captioning approach ZerAuCap, even outperforming it on semantics-aware metrics like METEOR and SPICE, while being simpler and applicable to many other modalities and tasks.

T2I

- Generator를 LLM이 아니라 Text to Image generator로 써서, 성능을 비교함(여기선 FLUX.1)
 - 여기서 Generator를 어떻게 바꾸냐에 따라서 Audio Generator가 될 수도 있음



Figure 6: **Style Transfer.** Using Gram Matrix distance (Gatys, 2015) as the SCORER, MILS can discover the edit prompt required to apply a given style to an image.

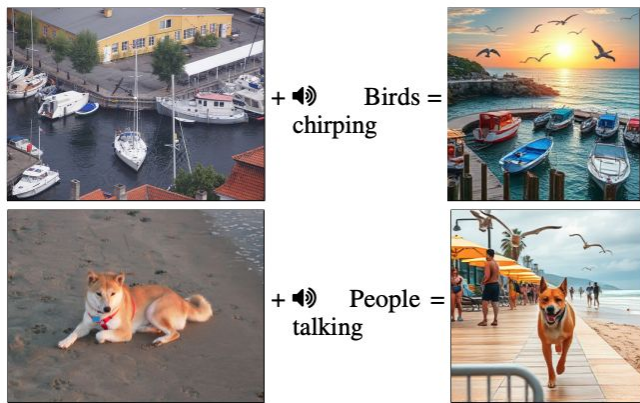


Figure 7: **MILS enables cross-modal arithmetic** by inverting modalities into text, combining them, and mapping them back to an image.