Zero-Shot Object Counting

2025년도 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

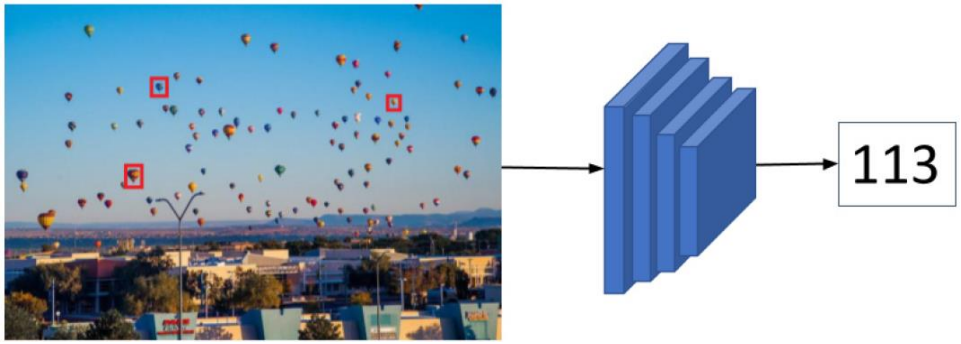
Presented By
안성욱

Outline

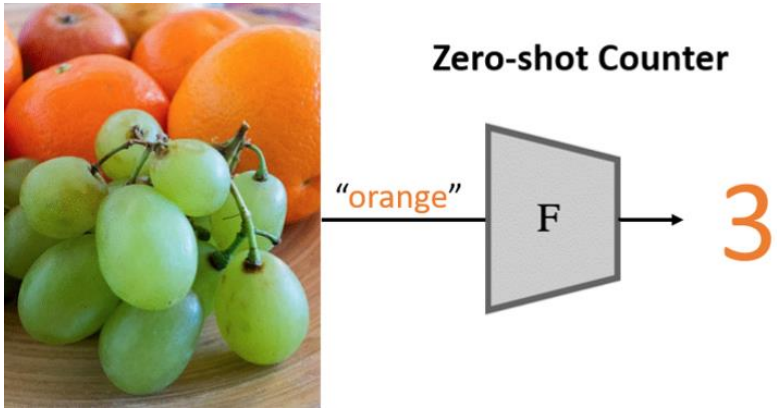
- Introduction
  - Object Counting
- RichCount: Expanding Zero-Shot Object Counting with Rich Prompts
  - arXiv 2025
- T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting
  - CVPR 2025

Introduction to Object Counting

- What is Object Counting?
  - Few-shot object counting



  - Zero-shot object counting

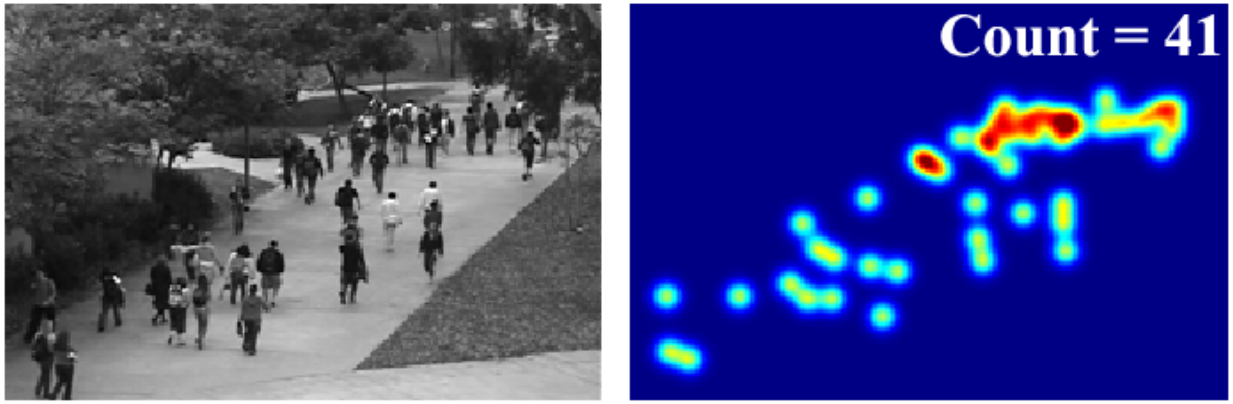


Introduction to Object Counting

- Importance of Object Counting
  - 적용 사례
    - 상품 계수 및 불량품 탐지
      - AI 기반 object counting으로 실시간 제품 계수
    - 생산 부품 제고 관리
      - AI 카메라를 통한 부품 개수 집계, 재고 관리 자동화
    - 불량률 모니터링 및 품질 향상
      - 제품 계수와 동시에 불량품 자동 식별, 실시간 모니터링
      - 생산 품질 향상 가능
  - 인력 절감, 생산성 향상, 품질 관리 강화, 비용 절감

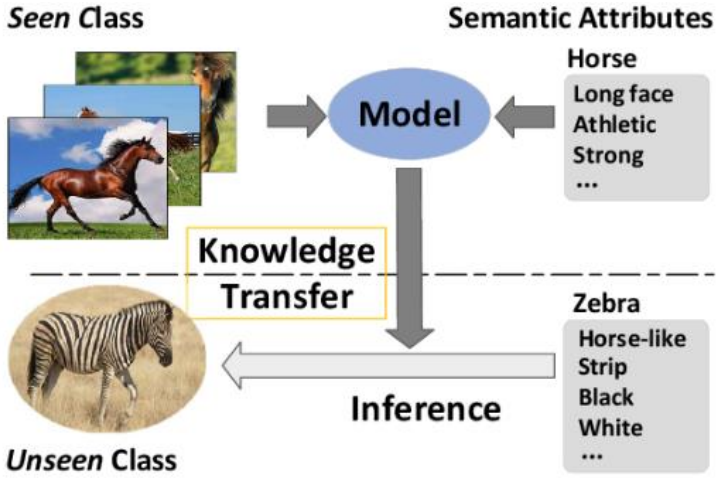
Introduction to Object Counting

- Limitation of Traditional Object Counting
  - Annotation 의존성
    - Supervised learning은 class 별로 대량의 bounding box/density map label 요구
    - 시간/비용 낭비
  - Generalization 성능
    - Train data에 없는 object category (unseen categories) 계수 불가능



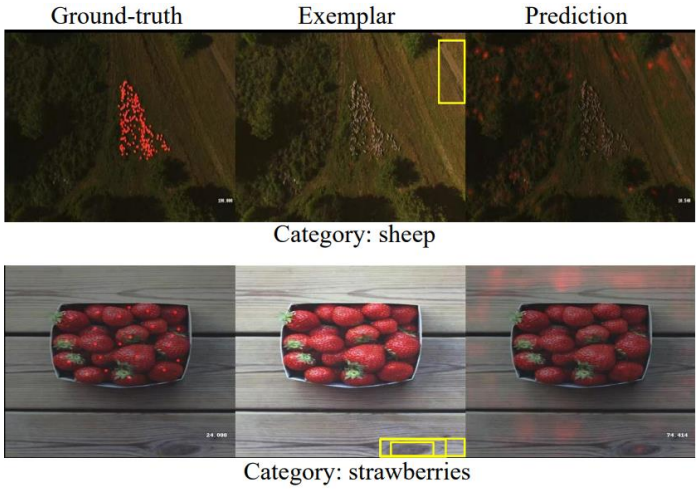
Introduction to Object Counting

- Zero-Shot Learning
  - 핵심 개념
    - Seen classes에서 unseen classes로 knowledge transfer
    - Semantic embedding space 활용
      - Object의 visual/text feature 연결
  - Object counting 적용
    - Annotation 비용 절감, 실시간 novel object를 count 가능
    - Open-world 시나리오 대응력 향상



Introduction to Object Counting

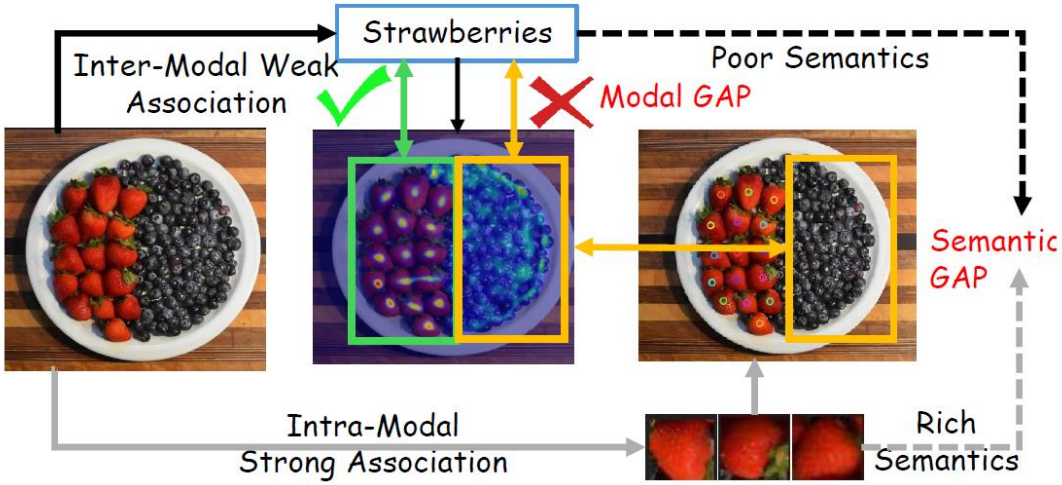
- Zero-Shot Object Counting
  - Challenges
    - Density map 생성의 한계
      - Unseen object는 pixel 단위 label의 부재로 전통적인 density regression이 불가능함
    - Visual-semantic gap
      - Text description과 실제 object의 표현이 불일치하는 경우가 많음
    - Background noise
      - Unseen object와 유사한 background를 오검출



- **RichCount: Expanding Zero-Shot Object Counting with Rich Prompts (arXiv 2025)**
- **T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting (CVPR 2025)**

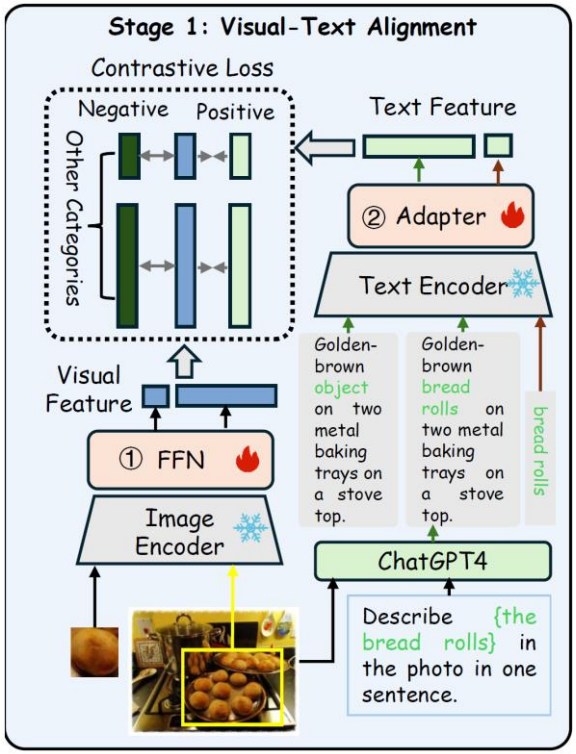
RichCount[1)]

- Introduction
  - 기존 zero-shot object counting의 핵심 문제
    - Modal gap
      - Text prompt와 visual feature 간 불일치
    - Semantic 정보가 풍부하지 않음
      - 단순 카테고리 label (e.g., “사과” 등)만 사용
      - Object 정보 부재
    - Prompt 유연성 부족
      - 복합적인 text를 입력했을 경우 처리가 불가능함



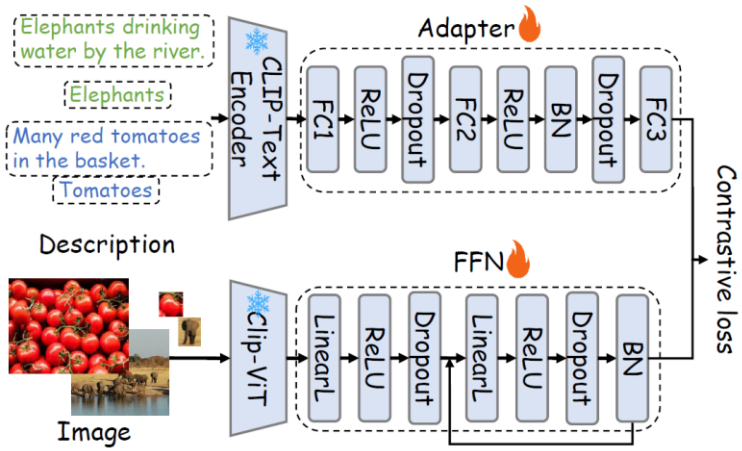
RichCount[1)]

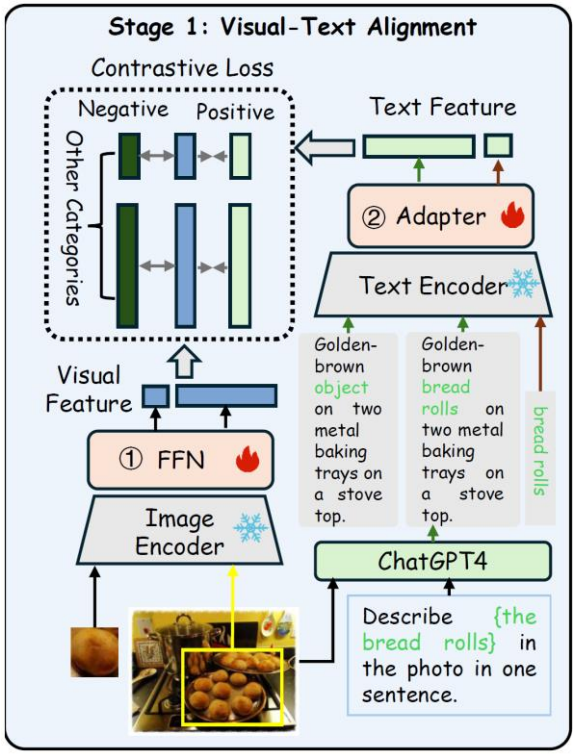
- Method
  - RichCount는 기존 zero-shot counting의 한계를 해결하기 위해 2-stage train 제안
  - Stage1: Visual-Text Alignment
    - Description Augmentation
      - ChatGPT-4로 object의 detail을 설명하는 문장 생성
      - 색상, 형태, 위치 등 정보의 추가
        - ✓Visual-semantic gap 해소
    - Feature Enhancement via FFN
      - CLIP visual encoder 뒤에 FFN 추가
      - Category embedding과 image embedding 출력



RichCount[1)]

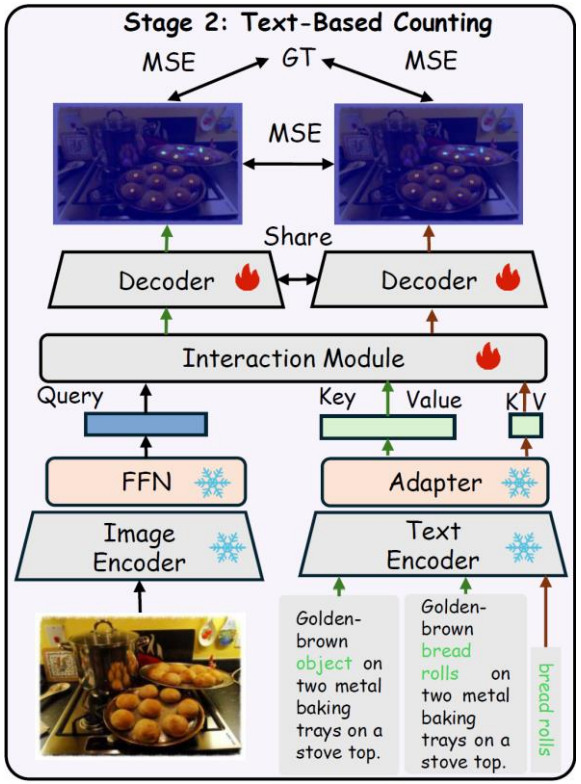
- Method
  - RichCount는 기존 zero-shot counting의 한계를 해결하기 위해 2-stage train 제안
  - Stage1: Visual-Text Alignment
    - Adapter Training for Textual Refinement
      - Adapter를 활용하여 text encoder 개선
    - Contrastive Loss
      - Visual-semantic 정보 align

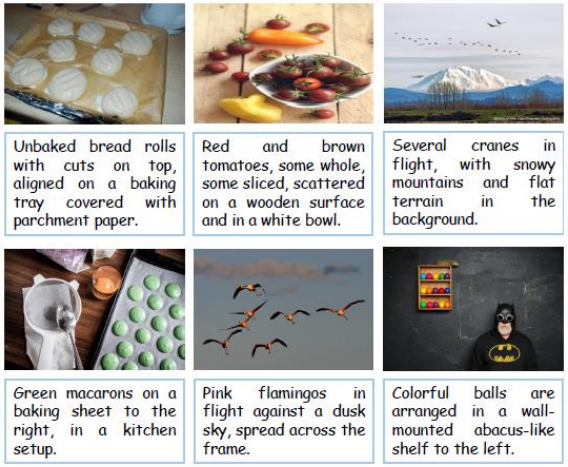




RichCount[1)]

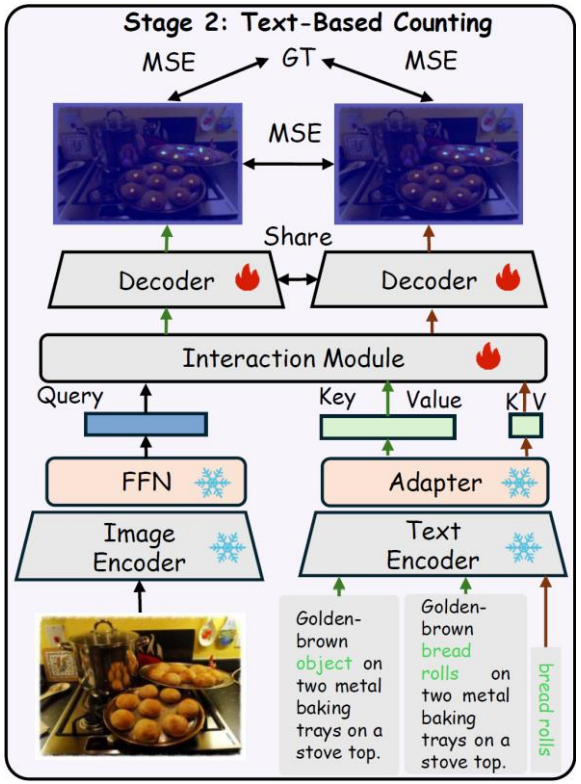
- Method
  - RichCount는 기존 zero-shot counting의 한계를 해결하기 위해 2-stage train 제안
  - Stage2: Text-Based Counting
    - Feature Fusion
      - Interaction module
        - ✓Q, K, V를 받아 fusion하는 역할을 수행함
      - Multi-prompt 지원
        - ✓카테고리, 세부 설명, object 일반화 설명





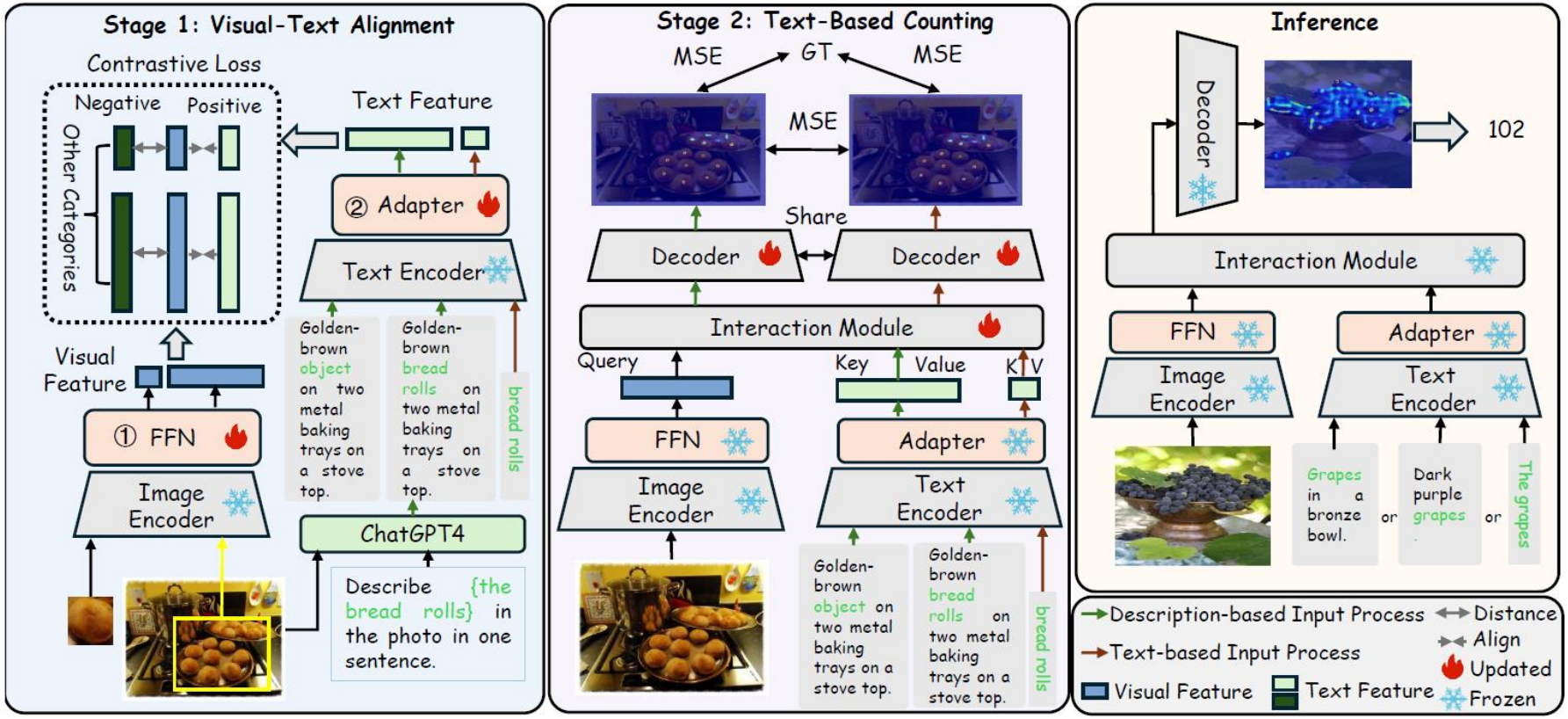
RichCount[1)]

- Method
  - RichCount는 기존 zero-shot counting의 한계를 해결하기 위해 2-stage train 제안
  - Stage2: Text-Based Counting
    - Density Map Generation
      - Decoder에서 세 개의 density map 출력
        - ✓카테고리(t), 세부 설명(d), object 일반화 설명(d')
    - Consistency loss
      - $\mathcal{L}_t = \sum_a \mathcal{L}_D(D^a, D^g) + \sum_{(a,b)} \mathcal{L}_D(D^a, D^b)$
        - ✓D^g =*ground truth density map*
        - ✓$a \in \{t, d, d'\}$, $(a, b) \in \{t, d, d'\}$
        - ✓출력된 density maps 각각 비교
        - ✓GT와 각 density map 비교



RichCount[1)]

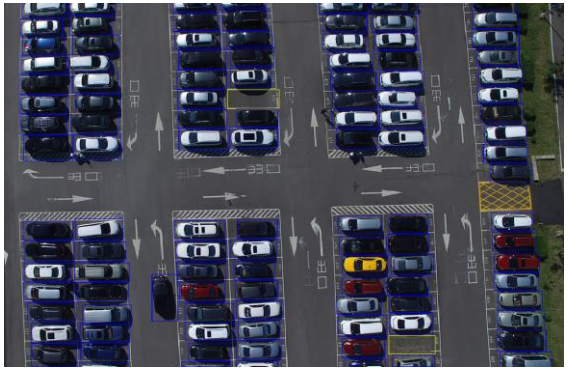
- Method
  - Inference



RichCount[1)]

- Experimental Result
  - Datasets
    - FSC-147
      - 다양한 object category를 제공하는 zero-/few-shot counting 연구의 표준 benchmark
    - CARPK
      - 드론으로 주차장 상공에서 촬영한 dataset
    - ShanghaiTech
      - 대규모 crowd counting benchmark dataset

< FSC-147 >



< CARPK >

< ShanghaiTech >

RichCount[1)]

- Experimental Result
  - Quantitative results on FSC-147

Scheme	Method	Venue	Exemplar	Val Set		Test Set		Avg	
				MAE	RMSE	MAE	RMSE	MAE	RMSE
Reference-less	FamNet [22]	CVPR'21	None	32.15	98.75	32.27	131.46	32.21	115.11
	RCC [7]	CVPR'22	None	17.49	58.81	17.12	104.53	17.31	81.67
	CounTR [15]	BMVC'23	None	18.07	71.84	**14.71**	106.87	**16.39**	89.36
	LOCA [6]	ICCV'23	None	**17.43**	**54.96**	16.22	**103.96**	16.83	**79.46**
Few-shot	FamNet [22]	CVPR'21	Visual Exemplars	24.32	70.94	22.56	101.54	23.44	86.24
	CFOCNet [32]	WACV'22	Visual Exemplars	21.19	61.41	22.10	112.71	21.65	87.06
	CounTR [15]	BMVC'22	Visual Exemplars	13.13	49.83	11.95	91.23	12.54	70.53
	PseCo [9]	CVPR'23	Visual Exemplars	15.31	68.34	13.05	112.86	14.18	90.60
	LOCA [6]	ICCV'23	Visual Exemplars	10.24	32.56	10.97	56.97	10.61	44.77
	CACViT [29]	AAAI'24	Visual Exemplars	9.13	10.63	48.96	37.95	10.94	52.99
	CountGD [2]	NeurIPS'24	Visual & Text	**7.10**	**26.08**	**5.74**	**24.09**	**6.42**	**16.25**
Zero-shot	ZSC [30]	CVPR'23	Text	26.93	88.63	22.09	115.17	24.51	101.90
	VA-Count [35]	ECCV'24	Text	17.87	73.22	17.88	129.31	17.87	101.26
	VLCount [12]	AAAI'24	Text	18.06	65.13	17.05	106.16	17.56	85.65
	CounTX [1] †	BMVC'23	Text	**16.99**	61.67	17.29	112.50	17.15	87.09
	CLIP-Count [11] †	ACM MM'23	Text	19.85	67.69	17.19	103.44	18.52	85.57
	CLIP-Count [11] †	ACM MM'23	Description	19.52	67.80	17.49	104.60	18.51	86.20
	RichCount (Ours)		Text	18.65	58.55	16.37	102.48	17.51	80.51
	RichCount (Ours)		Description	17.68	**57.24**	**15.78**	**99.65**	**16.73**	**78.45**

RichCount[1)]

- Experimental Result
  - Cross-dataset evaluation (FSC-147 → CARPK)

Method	Venue	Exemplar	FSC → CARPK	
			MAE	RMSE
FamNet [22]	CVPR'21	Visual	28.84	44.47
BMNet [26]	CVPR'22	Visual	14.41	24.60
BMNet+ [26]	CVPR'22	Visual	10.44	13.77
RCC [7]	CVPR'22	Text	21.38	26.61
CLIP-Count [11]	ACM MM'23	Text	13.59	18.30
RichCount (Ours)		Description	**9.91**	**13.28**

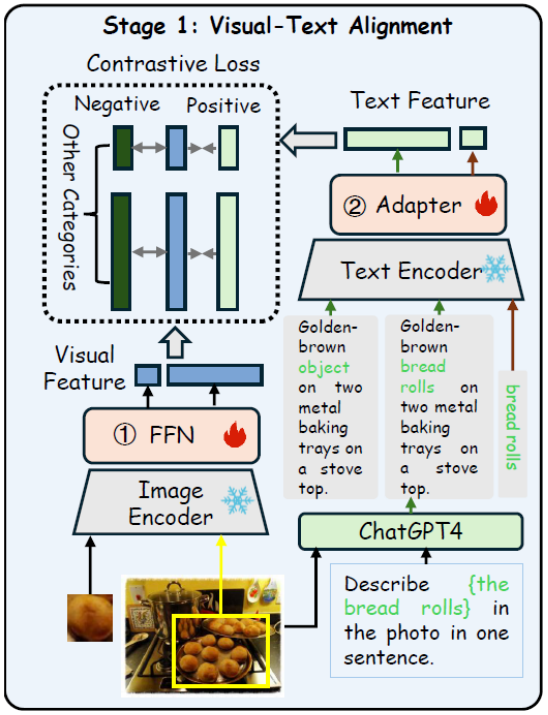
  - Cross-dataset evaluation

	Type	Method	ShanghaiTech Part B		ShanghaiTech Part A	
			SHB		SHA	
			MAE	RMSE	MAE	RMSE
Trained on SHA	Specific	MCNN [34]	85.20	142.30	221.40	357.80
		CrowdCLIP [14]	69.60	80.70	217.00	322.70
Trained on FSC-147	Generic	RCC [7]	66.60	104.80	240.10	366.90
		CLIP-Count [11]	47.92	80.48	197.47	319.75
		RichCount (Ours)	**44.77**	**75.62**	**193.39**	**314.35**

RichCount[1)]

- Experimental Result
  - Ablation study on FSC-147

FFN	Ada	Des	$\mathcal{L}_c$	Val Set		Test Set		Average	
				MAE	RMSE	MAE	RMSE	MAE	RMSE
○	○	○	○	19.85	67.69	17.19	103.44	18.52	85.57
○	●	○	●	18.48	62.72	17.02	**99.37**	17.75	81.05
○	●	●	●	17.79	**57.13**	16.35	102.32	17.07	79.73
●	○	○	○	18.21	68.51	18.54	101.69	18.38	85.10
●	●	○	○	18.13	61.32	17.57	104.36	17.85	82.84
●	○	○	●	18.19	60.58	18.57	106.18	18.38	83.38
●	●	●	○	17.85	63.02	17.58	101.71	17.72	82.37
●	●	○	●	17.99	60.65	17.25	99.68	17.62	80.17
●	●	●	●	**17.68**	57.24	**15.78**	99.65	**16.73**	**78.45**



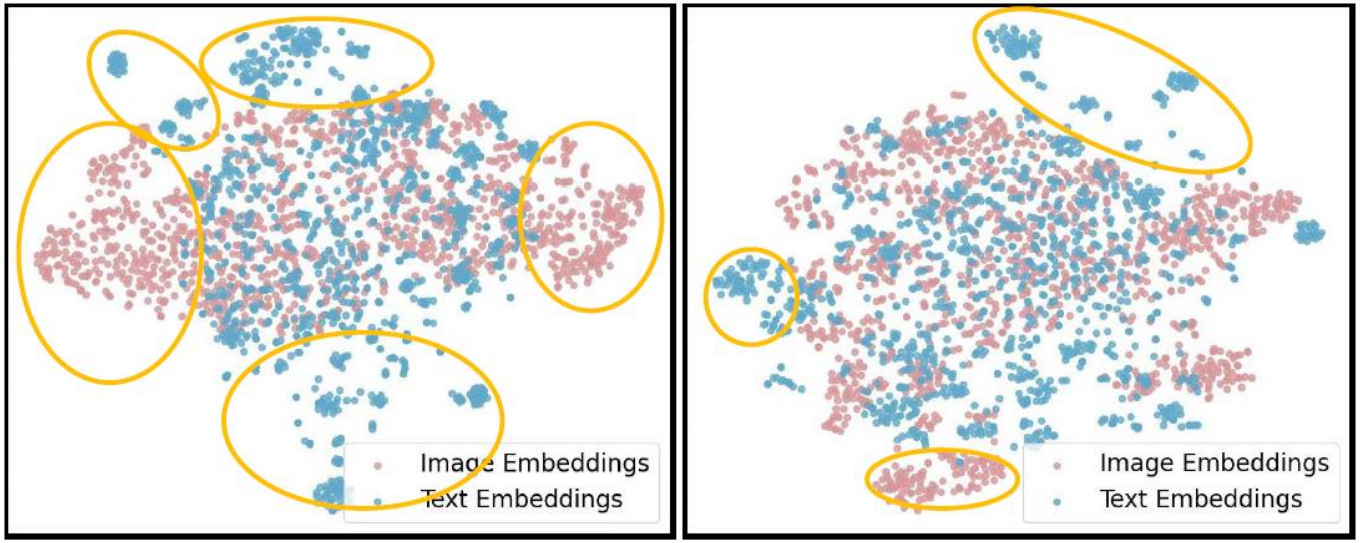
RichCount[1)]

- Experimental Result
  - Impact of image descriptions on counting performance on FSC-147

Method	Exemplar	Val Set		Test Set	
		MAE	RMSE	MAE	RMSE
Text	Text	<u>17.99</u>	60.65	17.25	99.68
Claude [3]	Text	18.42	62.59	17.13	102.20
GPT-4o [19]	Text	18.27	62.13	17.59	100.36
GPT-4	Text	18.65	<u>58.55</u>	<u>16.37</u>	102.48
Claude [3]	Claude	18.05	59.36	16.63	100.65
GPT-4o [19]	GPT-4o	18.08	65.29	16.85	**98.84**
GPT-4 [19]	GPT-4	**17.68**	**57.24**	**15.78**	<u>99.65</u>

RichCount[1)]

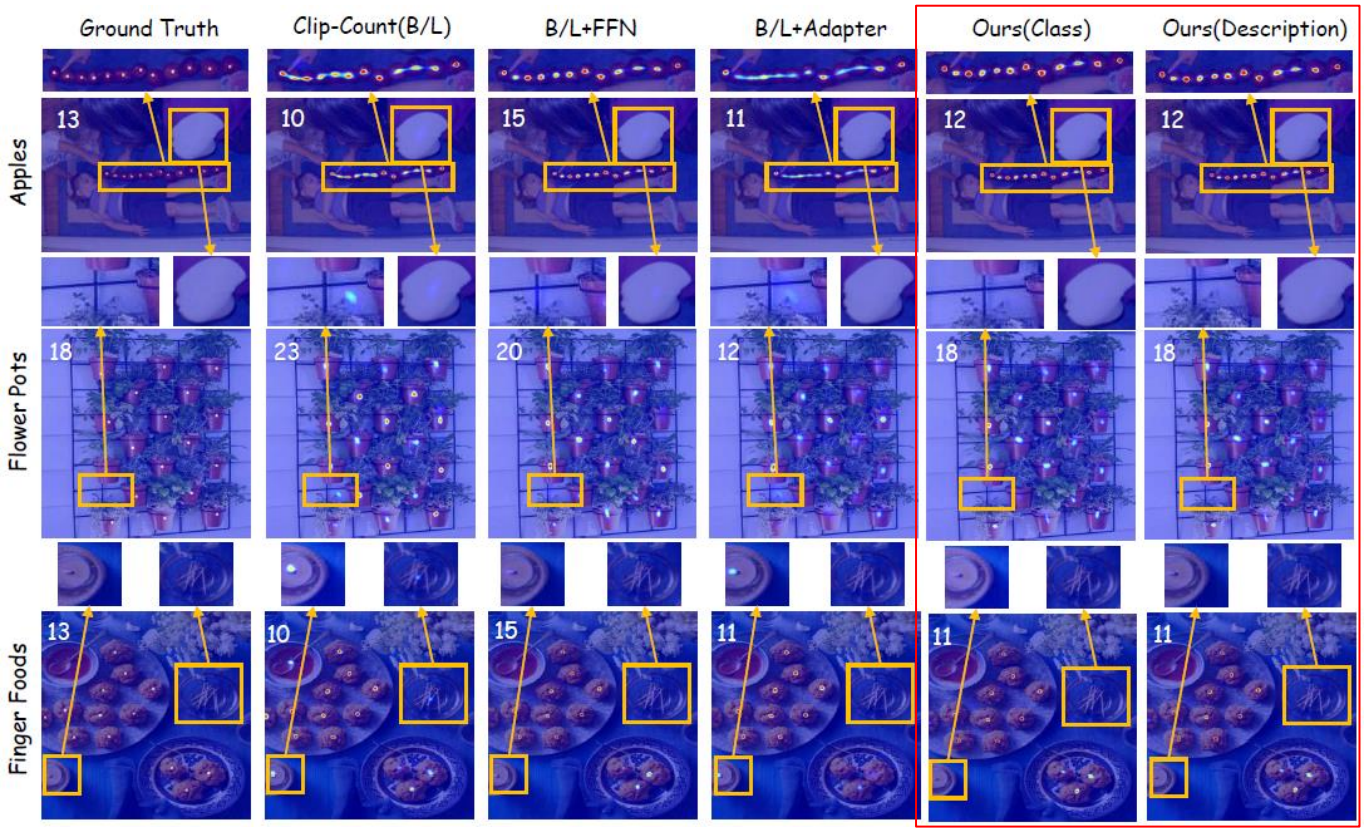
- Experimental Result
  - Analysis of image-text alignment
    - Misaligned cluster를 노란색 원으로 표시



(a) Clipcount(B/L) (b) EACount(Ours)

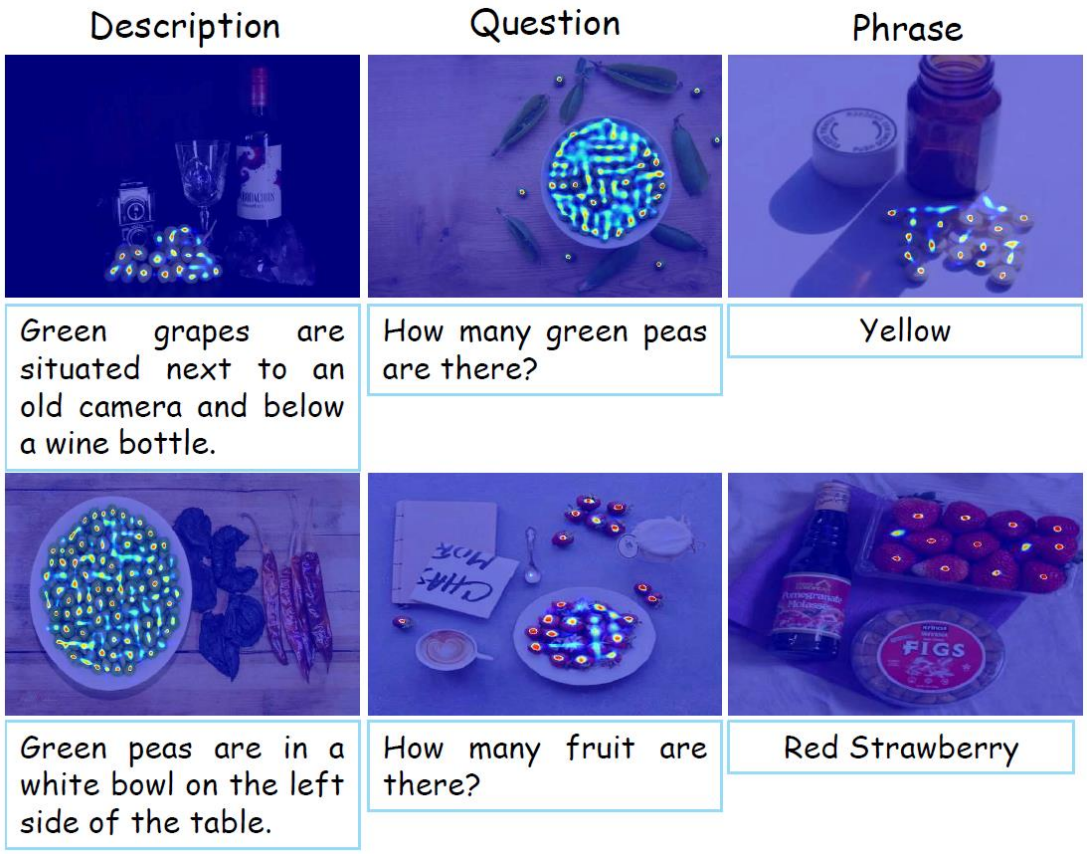
RichCount[1)]

- Experimental Result
  - Zero-shot density maps on FSC-147



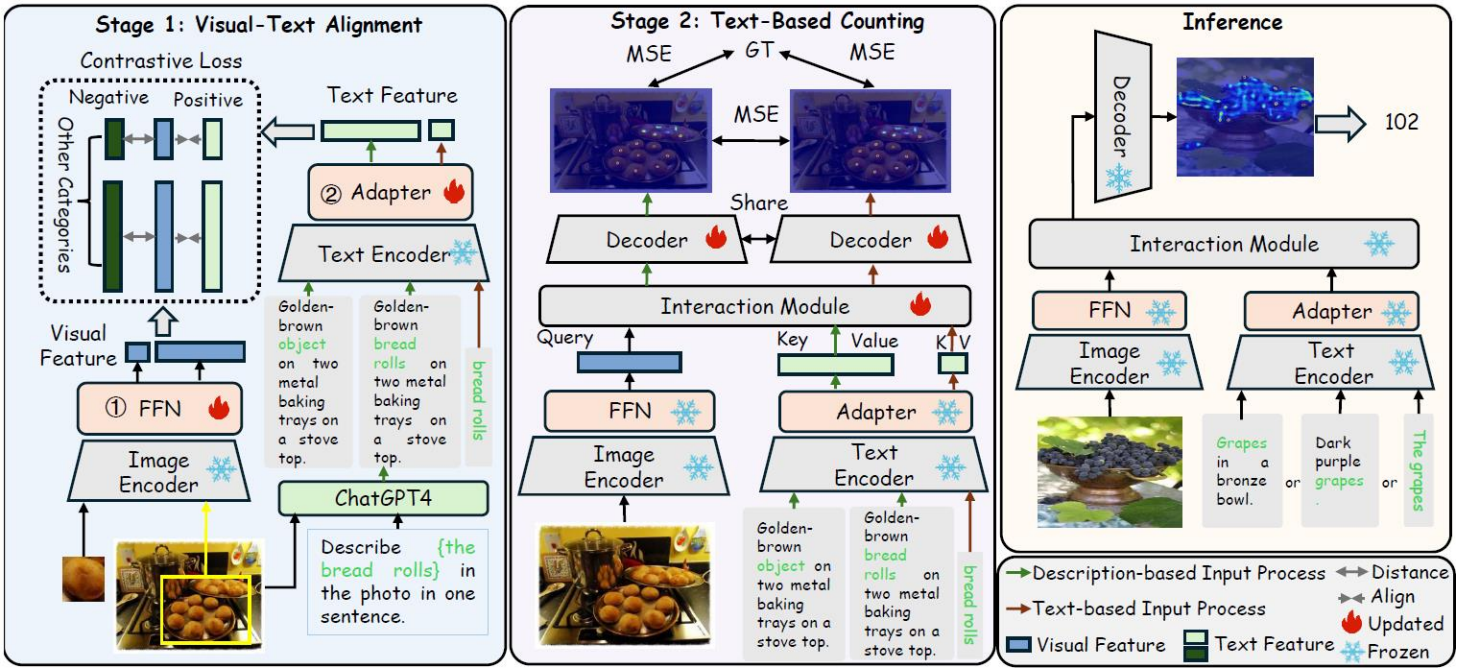
RichCount[1)]

- Experimental Result
  - Illustration of density maps generated from various texts



RichCount[1)]

- Conclusion
  - 다양한 text prompt (detail/일반화 설명 등)를 활용해 zero-shot counting의 한계 극복
  - Visual-semantic alignment 강화로 정확성과 일반화 능력 동시 향상
  - 주요 benchmark에서 SOTA 성능 달성

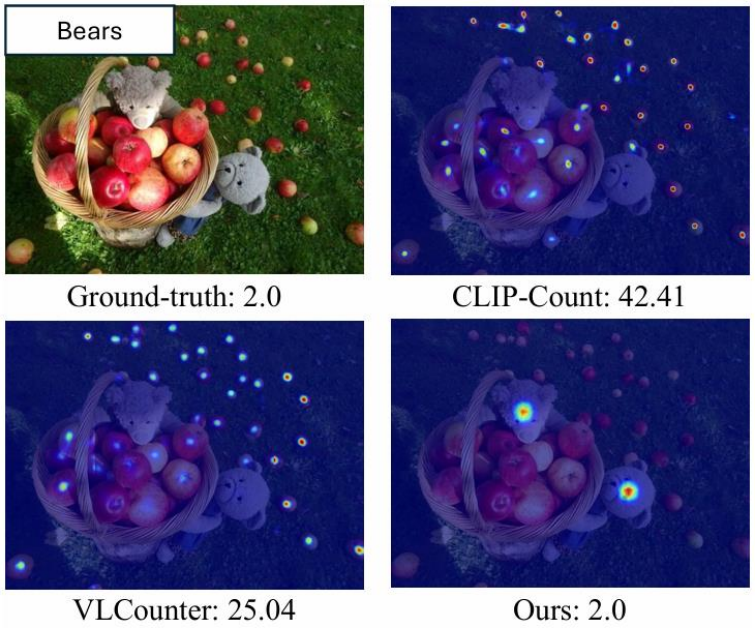


- RichCount: Expanding Zero-Shot Object Counting with Rich Prompts (arXiv 2025)

- **T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting (CVPR 2025)**

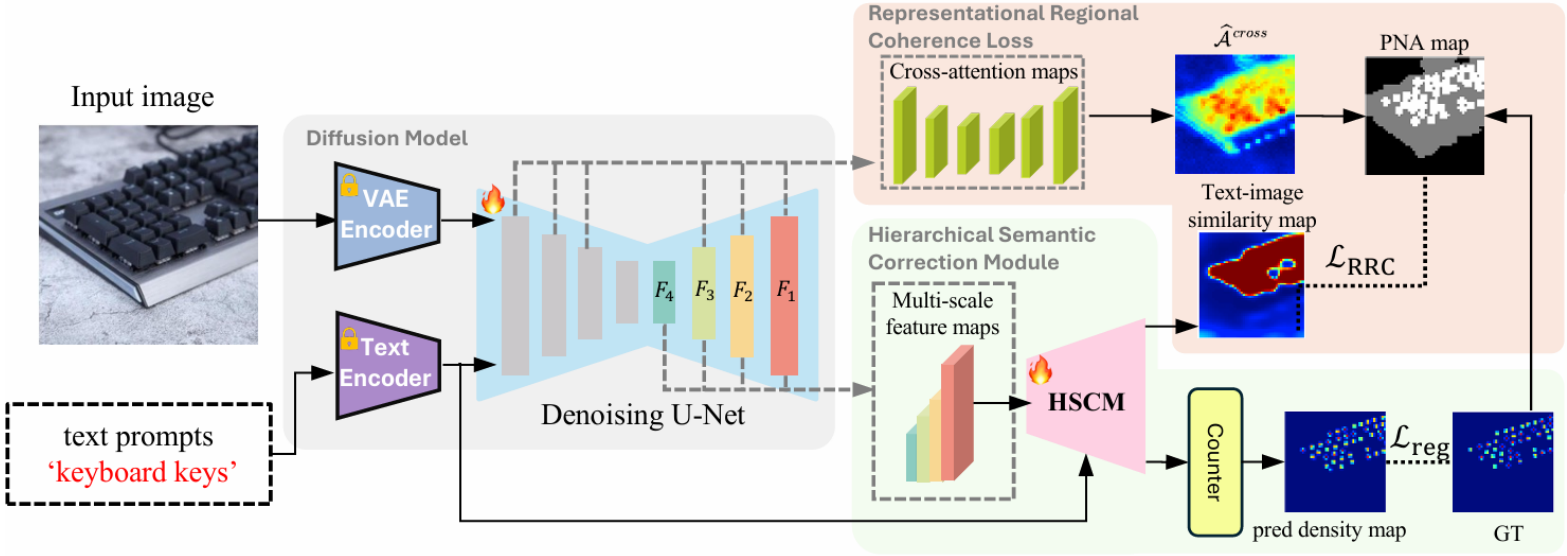
T2ICount[1)]

- Introduction
  - 기존 방법의 핵심 문제: Text insensitivity
    - CLIP 기반 모델의 한계
      - Global semantic 정보에 집중 → pixel 수준 object counting에 부적합
      - Text prompt보다 image 내 다수 object에 편향됨
    - Text prompt에 민감하게 반응하면서 다양한 object를 정확히 count하는 framework 제안



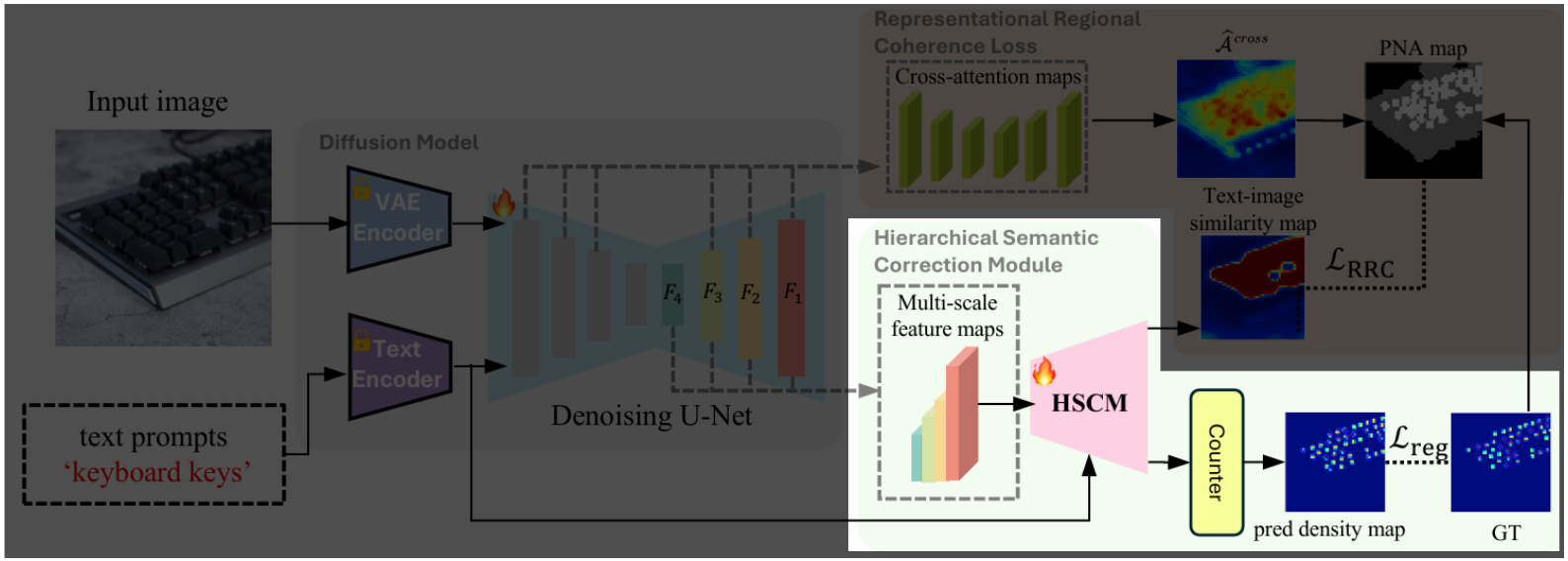
T2ICount[1)]

- Method
  - T2ICount는 위와 같은 문제를 해결하기 위해 pre-trained diffusion model을 활용함
  - 전체 구조
    - Stable diffusion 기반의 pre-trained model 사용
      - Image-text alignment 능력, pixel level 표현력 보유
    - Single-step denoising
      - 효율적인 inference를 위해 (speed↑, text sensitivity↓)



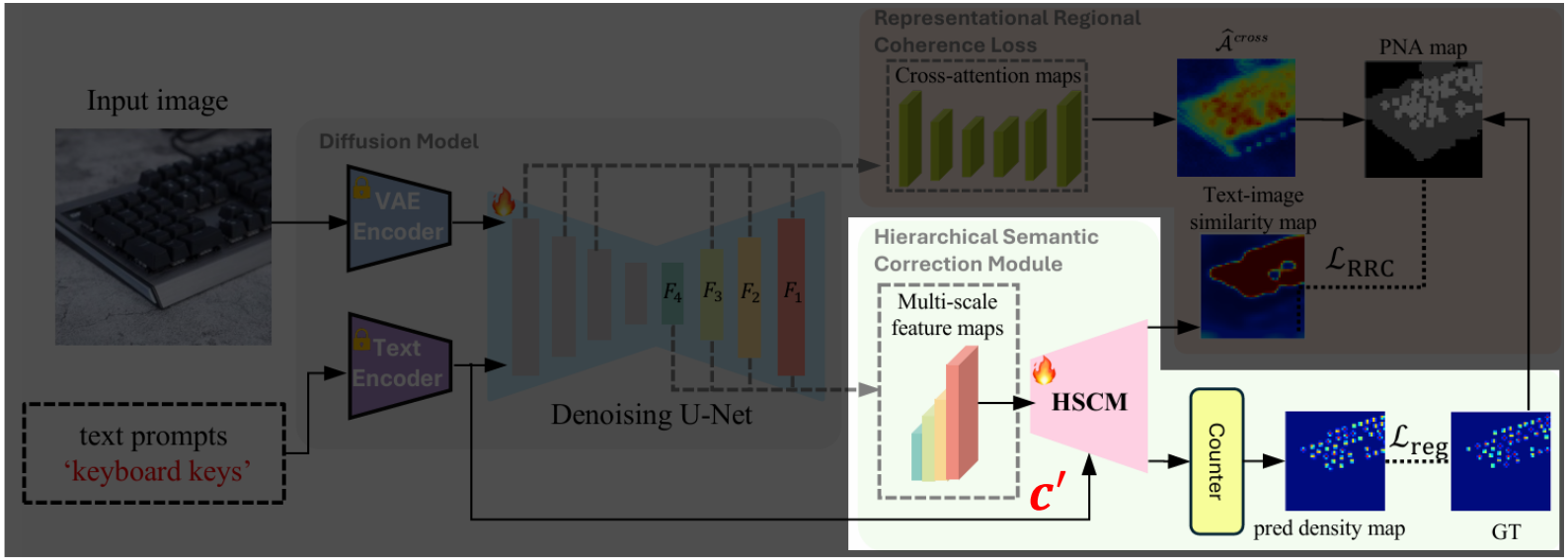
T2ICount[1)]

- Method
  - Hierarchical Semantic Correction Module (HSCM)
    - Three refinement processes
      - 1. Multi-scale feature fusion
        - ✓U-Net decoder 부분의 feature map (F_4~F_1)를 up-sampling 및 concatenating
        - ✓$F_i' = Conv(Concat(Up(V_{i+1}), F_i))$
          - Output = F_3', F_2', F_1'



T2ICount[1)]

- Method
  - Hierarchical Semantic Correction Module (HSCM)
    - Three refinement processes
      - 2. Semantic Enhancement Module (SEM)
        - ✓Image-text cross-attention을 적용하여 visual 정보와 text 정보 align, V_i 생성
        - ✓Cosine similarity map 생성
          - $S_i = \frac{V_i \cdot c'}{\|V_i\| \|c'\|}$ → Pixel-level alignment 품질 향상



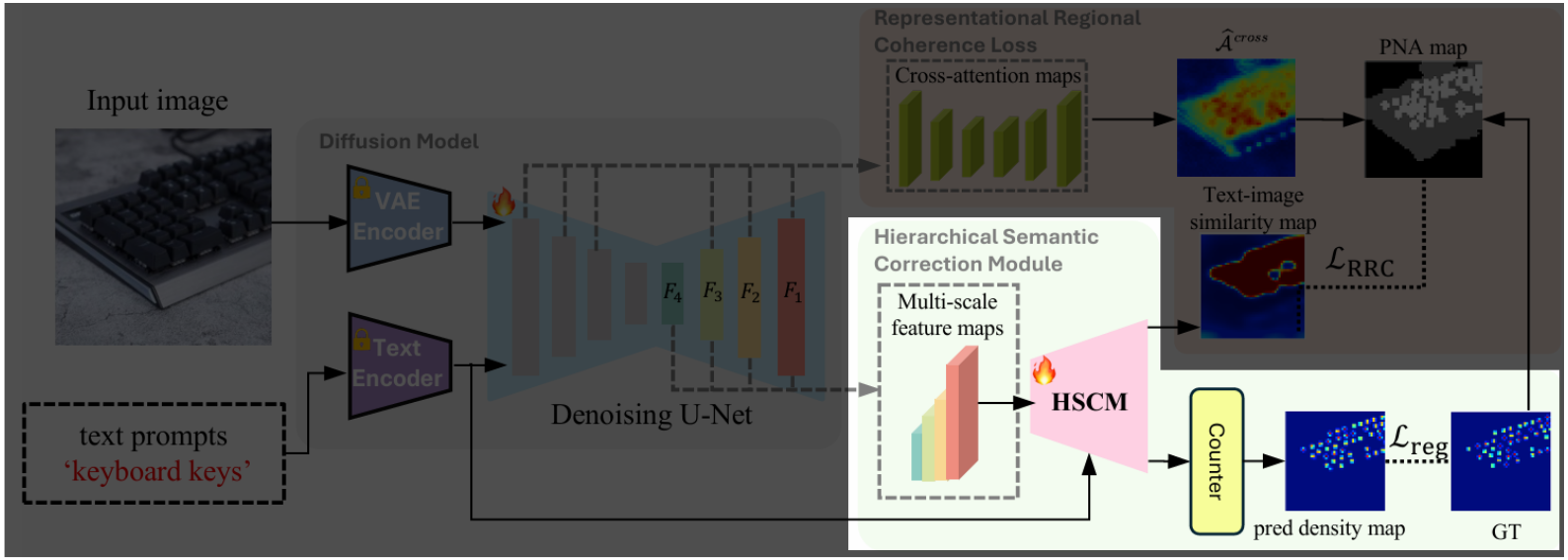
T2ICount[1)]

- Method
  - Hierarchical Semantic Correction Module (HSCM)
    - Three refinement processes
      - 3. Semantic Correction Module (SCM)

✓Low-resolution → high-resolution으로 hierarchical refinement, feature map 재구성

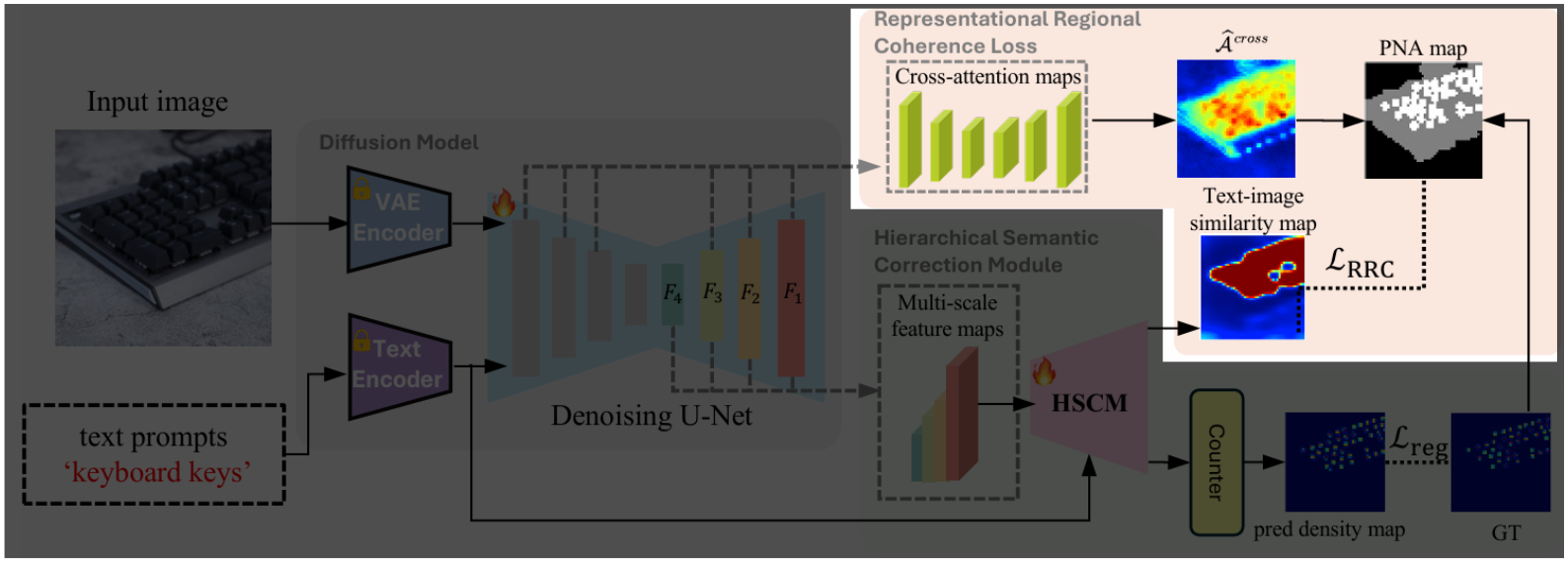
✓$F_i' \leftarrow F_i' + Up(V_{i+1} \odot S_{i+1})$

✓Low-resolution에서 대략적 object 영역 식별 후 high-resolution에서 detail 정교화



T2ICount[1)]

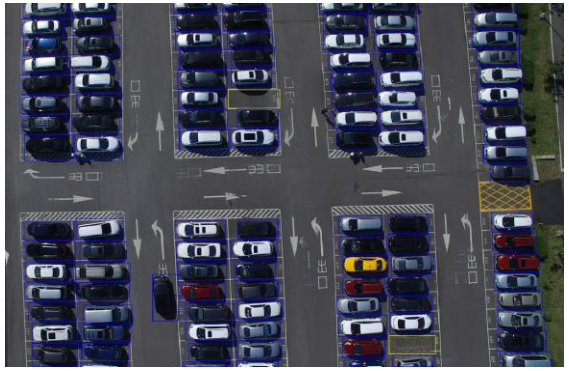
- Method
  - Representational Regional Coherence Loss ($\mathcal{L}_{RRC}$)
    - Point-level annotation만으로 신뢰할 수 있는 supervision signal를 생성하는 것이 목표
      - PNA (Positive-Negative-Ambiguous) map 생성
        - ✓Positive: GT $density \geq T$
        - ✓Negative: $\hat{A}^{cross} \leq \theta$
        - ✓Ambiguous: 나머지 (ambiguous 영역은 background로 강제하지 않음)



T2ICount[1)]

- Experimental Result
  - Datasets
    - FSC-147
      - 다양한 object category를 제공하는 zero-/few-shot counting 연구의 표준 benchmark
    - FSC-147-S (본 논문 contribution)
      - Minor object counting 평가를 위한 re-annotation subset
    - CARPK
      - 드론으로 주차장 상공에서 촬영한 dataset

< FSC-147 >



< CARPK >

T2ICount[1)]

- Experimental Result
  - Quantitative results on FSC-147

Scheme	Method	Venue	Shot	Val Set		Test Set	
				MAE	RMSE	MAE	RMSE
Few-shot	FamNet [24]	CVPR'21	3	24.32	70.94	22.56	101.54
	BMNet [27]	CVPR'22	3	15.74	58.53	14.62	91.83
	LOCA [31]	ICCV'23	3	**10.24**	**32.56**	**10.97**	**56.97**
	SAM [28]	WACV'24	3	-	-	19.95	132.16
	PseCo [38]	CVPR'24	3	15.31	68.34	13.05	112.86
	GMN [14]	ACCV'19	1	29.66	89.81	26.52	124.57
	FamNet [24]	CVPR'21	1	24.32	70.94	22.56	101.54
	BMNet [27]	CVPR'22	1	19.06	67.95	16.71	103.31
Reference-less	FamNet [24]	CVPR'21	0	32.15	98.75	32.27	131.46
	LOCA [31]	ICCV'23	0	**17.43**	**54.96**	16.22	**103.96**
	RCC [5]	CVPR'23	0	17.49	58.81	17.12	104.53
Zero-shot	Patch-selection [33]	CVPR'23	0	26.93	88.63	22.09	115.17
	CLIP-Count [7]	ACM MM'23	0	18.79	61.18	17.78	106.62
	CounTX [1]	BMVC'23	0	17.10	65.61	15.88	106.29
	VLCounter [8]	AAAI'24	0	18.06	65.13	17.05	106.16
	PseCo [38]	CVPR'24	0	23.90	100.33	16.58	129.77
	DAVE [16]	CVPR'24	0	15.48	**52.57**	14.90	103.42
	VA-Count [39]	ECCV'24	0	17.87	73.22	17.88	129.31
	GeCo [15]	NeurIPS'24	0	14.81	64.95	13.30	108.72
	T2ICount (Ours)	-	0	**13.78**	58.78	**11.76**	**97.86**

T2ICount[1)]

- Experimental Result
  - Quantitative results on FSC-147-S

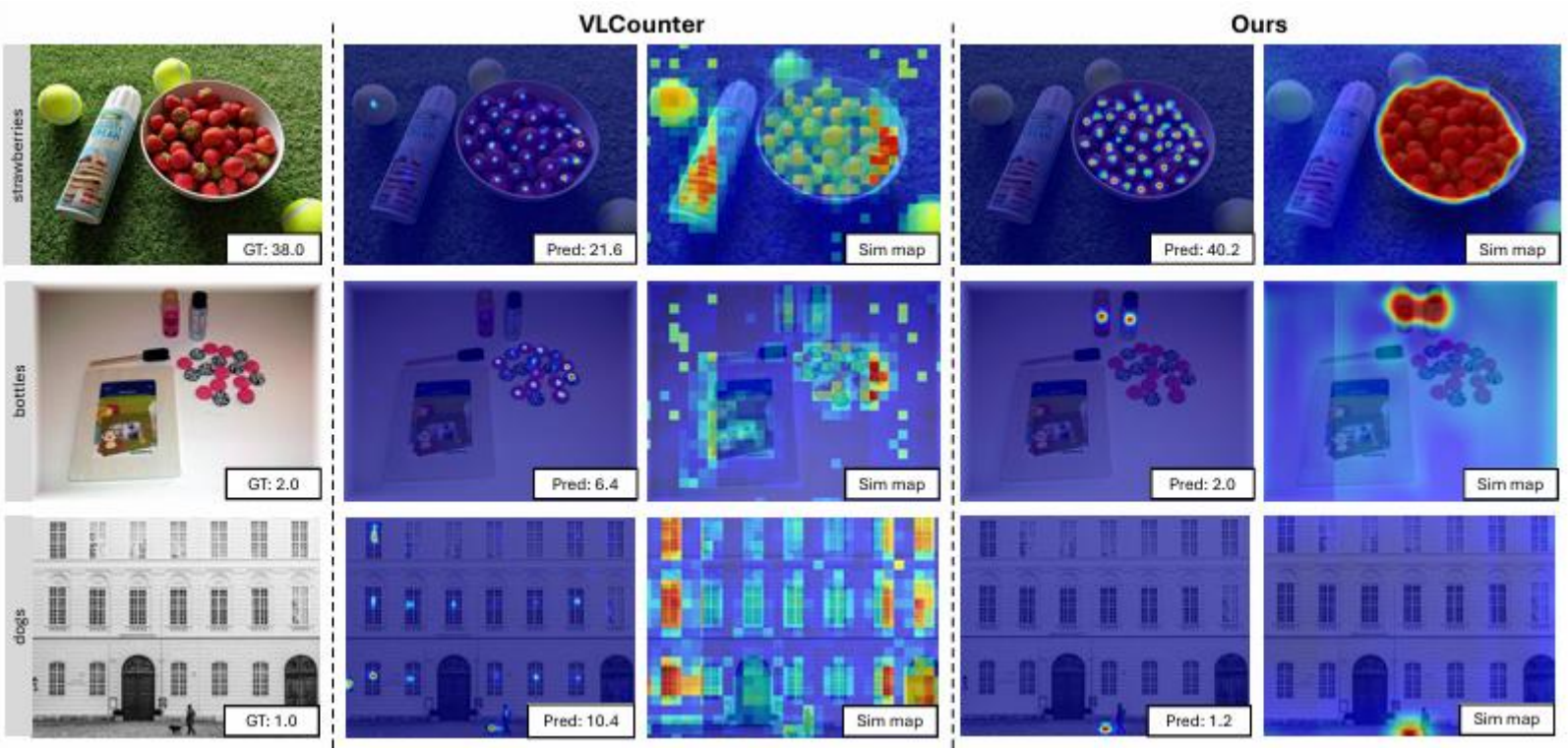
Method	MAE	RMSE
CLIP-Count [7]	48.42	108.04
CounTX [1]	<u>31.30</u>	98.80
VLCounter [8]	35.24	75.46
PseCo [38]	39.01	<u>61.34</u>
DAVE [16]	49.32	108.47
T2ICount (Ours)	**4.69**	**8.06**

  - Quantitative results on CARPK

Method	MAE	RMSE
RCC [5]	21.38	26.61
CLIP-Count [7]	11.96	16.61
CounTX [1]	**8.13**	**10.87**
Grounding DINO [13]	29.72	31.60
VA-Count [39]	10.63	<u>13.20</u>
T2ICount (Ours)	<u>8.61</u>	13.47

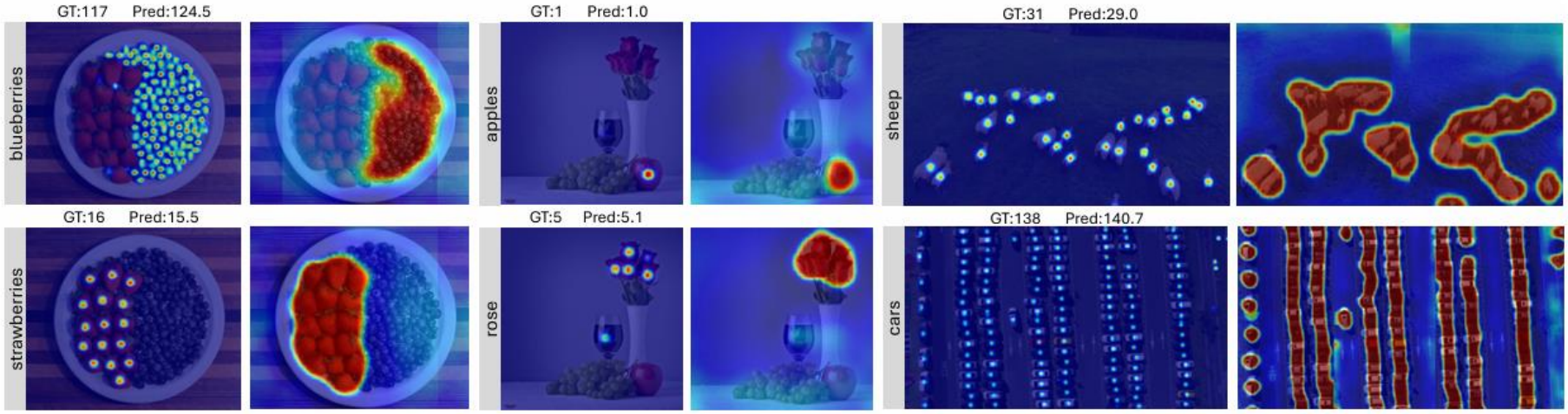
T2ICount[1)]

- Experimental Result
  - Qualitative results on FSC-147-S



T2ICount[1)]

- Experimental Result
  - Qualitative results of T2ICount
    - Density map (left)
    - Similarity map (right)



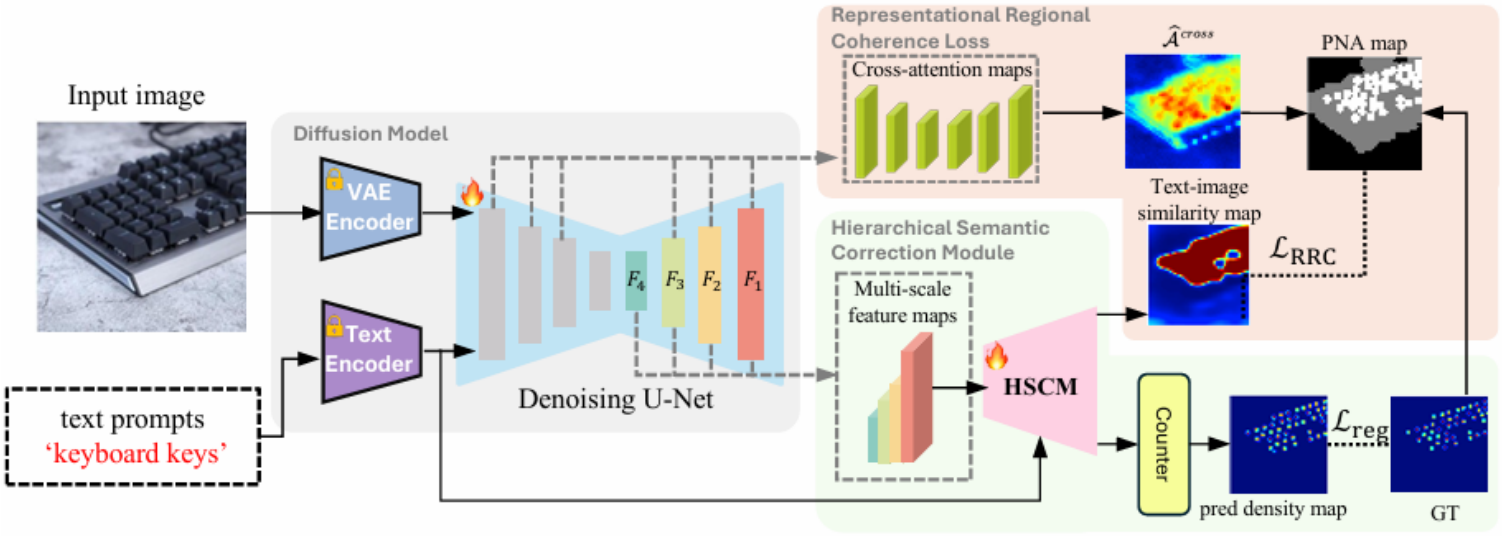
T2ICount[1)]

- Experimental Result
  - Ablation study on the key components of T2ICount

	Test		FSC-147-S	
	MAE	RMSE	MAE	RMSE
Baseline (B)	14.66	111.62	24.34	64.74
B + $\mathcal{L}_{RRC}$	14.55	106.21	9.59	22.25
B + $\mathcal{L}_{RRC}$ + HSCM	**11.76**	**97.86**	**4.69**	**8.06**

T2ICount[1)]

- Conclusion
  - Text sensitivity 문제를 근본적으로 해결한 zero-shot counting framework 제안
  - HSCM, $\mathcal{L}_{RRC}$로 image-text alignment 및 minor object counting 능력 향상
  - FSC-147-S라는 새로운 benchmark dataset 제안
    - Model의 text 기반 counting 능력 검증 가능



감사합니다