T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting

Supplementary Material

1. Implementation Details

1.1. Hierarchical Semantic Correction Module

We illustrate the workflow of the proposed Hierarchical Semantic Correction Module in fig 1 for clarity. This module progressively integrates multi-scale features from the denoising U-Net decoder while enhancing cross-modal alignment between visual and textual representations.

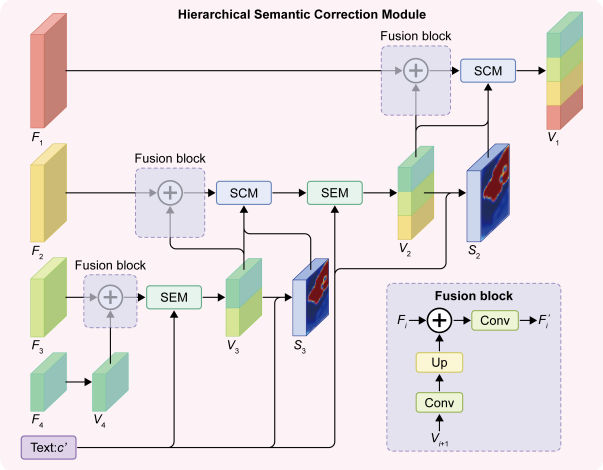

Figure 1. The workflow of the proposed Hierarchical Semantic Correction Module

Semantic Enhance Module: Text-to-image and image-to-text attention mechanisms operate through scaled dot-product attention computations: text-to-image attention maps textual queries to visual key-value pairs to locate relevant image regions, while image-to-text attention uses visual queries to attend to textual key-value pairs, enabling effective cross-modal feature alignment and information exchange between modalities.

1.2. Regression Loss

We detail the regression loss used in our framework here, which is proposed by CUT [18]. The loss function consists of two terms: $\mathcal{L}_{\text{SSIM}}$ which optimizes the structural similarity in the foreground region, and the $\mathcal{L}_2$-like total variation loss, $\mathcal{L}_{\text{TV}}$ which maintains the consistency of background regions. It can be expressed as:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{SSIM}} + \beta \cdot \mathcal{L}_{\text{TV}}, \tag{11}$$

where

$$\mathcal{L}_{\text{SSIM}} = \frac{1}{M}\sum_{m=1}^{M}\left(1-\texttt{SSIM}\big(\texttt{AP}_m(D\odot S^{\text{gt}}), \texttt{AP}_m(D^{\text{gt}}\odot S^{\text{gt}})\big)\right), \tag{12}$$

where $\texttt{SSIM}$ represents the structural similarity index, accessing the discrepency between the predicted (D) and ground truth (D^{gt}) density maps, with $\texttt{AP}_m$ indicating the average pooling operation that scales down an image to $\frac{1}{2^{m-1}}$ of its original size. The Hadamard product ($\odot$) enables element-wise multiplication, focusing the model's attention on dense regions identified by the binary segmentation map $S^{\text{gt}} : D^{\text{gt}} \geq 1e-3$. In addition, M is set as 3. The total variation loss, $\mathcal{L}_{\text{TV}}$ alongside the hyper-parameter, β, set to 0.1, ensuring the model accurately captures crowd density variations across the entire image.

2. Discussion

While conventional diffusion models typically employ a multi-step denoising process, we strategically select z_1 (small t) for denoising in our single-step denoising framework. This choice is driven by a key insight: the noise addition process inherently introduces uncertainty and potentially disrupts the original image structure. As shown in Theorem 1, the upper bound of difference between the clean image z_0 and diffused image z_t is increasing by the t. With $t = 1$, we prioritize the retention of original image semantics for accurate object counting while maintaining determinism.

Theorem 1 *The L2 distance between the clean image and noised image satisfies the bound with a probability of at least* $1-\delta$*, we have*

$$\|z - z_t\| \leq (1-\sqrt{\bar{\alpha}_t})\|x\| + \sqrt{1-\bar{\alpha}_t}C$$

where $C := \sqrt{d + 2\sqrt{d\log\frac{1}{\delta}} + 2\log\frac{1}{\delta}}$

Proof:

$$\begin{aligned}\|z - z_t\| &= \|z - \sqrt{\bar{\alpha}_t}z - \sqrt{1-\bar{\alpha}_t}\epsilon\| \\ &\leq \|z - \sqrt{\bar{\alpha}_t}z\| + \sqrt{1-\bar{\alpha}_t}\|\epsilon\| \\ &= (1-\sqrt{\bar{\alpha}_t})\|z\| + \sqrt{1-\bar{\alpha}_t}\|\epsilon\|\end{aligned} \tag{13}$$

As we know $\|\epsilon\|^2$ is Chi-squared distribution, by concentration inequality [2], we have

$$P(\|\epsilon\|^2 \geq d + 2\sqrt{dr} + 2r) \leq e^{-r} \tag{14}$$

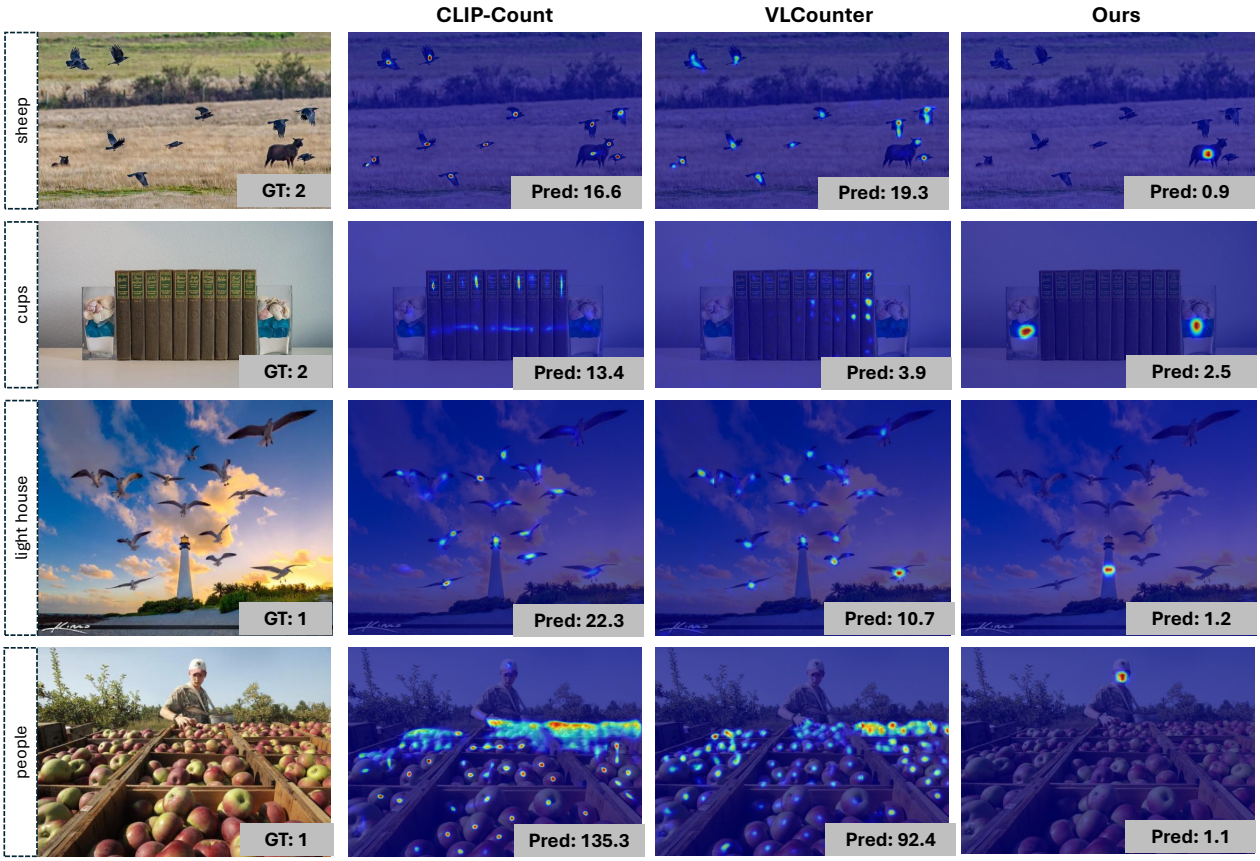

Figure 2. Qualitative comparison of T2ICount with CLIP-Count [8] and VLCounter [9] on the FSC-147-S protocal.

Let $e^{-r} = \delta$, we have

$$P\left(\|\epsilon\| \geq \sqrt{d + 2\sqrt{d\log\frac{1}{\delta}} + 2\log\frac{1}{\delta}}\right) \leq \delta \qquad (15)$$

Therefore, we have

$$\|z - z_t\| \leq (1 - \sqrt{\bar{\alpha}_t})\|z\| + \sqrt{1 - \bar{\alpha}_t}C \qquad (16)$$

with probability at least $1 - \delta$

3. Ablation Study

Effects of fine-tuning the U-Net: We empirically validate whether the weights of the denoising U-Net can remain frozen during training. The results are provided in Table 1. We find that when the U-Net is entirely frozen, the model shows poor performance after training. This is likely because the diffusion model was previously trained for image generation and not for counting tasks. Therefore, there is a domain gap between the pre-trained model and our specific task, which requires the U-Net to be fine-tuned to learn domain-specific knowledge.

Table 1. Ablation study on freezing or fine-tuning the UNet

U-Net Frozen?	Validation		Test	
	MAE	RMSE	MAE	RMSE
Yes	35.16	104.92	37.50	133.47
No	**13.78**	**58.78**	**11.76**	**97.86**

Ablation on the timestep t: Apart from the theoretical analysis of using features from timestep $t = 1$, we empirically demonstrate how different timesteps influence counting results in Tab 2. Generally, later (small) timesteps in the reverse diffusion process demonstrate comparable performance while preserving more informative features than earlier timesteps (large).

Ablation on γ in $\mathcal{L}_{\text{RRC}}$: We investigate the optimal balance between positive and negative samples in $\mathcal{L}_{\text{RRC}}$ by fine-tuning γ. As shown in Tab 3, our experiments demonstrate that doubling the weight for positive samples ($\gamma = 2$) in the loss function provides the most effective supervision, outperforming other weighting strategies.

Table 2. Ablation study on the timestep t

t	Validation		Test	
	MAE	RMSE	MAE	RMSE
1	**13.78**	**58.78**	**11.76**	**97.86**
50	14.83	60.17	13.16	100.22
100	15.54	68.93	14.92	116.28
150	15.63	61.66	15.41	108.30
250	16.86	64.44	15.40	111.81
700	18.79	78.67	18.55	117.47

Table 3. Ablation study on the γ

γ	Validation		Test	
	MAE	RMSE	MAE	RMSE
0.5	15.82	64.81	15.45	102.96
1	14.42	63.37	13.00	109.76
2	**13.78**	**58.78**	**11.76**	**97.86**
4	16.75	68.55	15.91	125.36

4. Qualitative results

We provide with more qualitative comparisons between our proposed model and two existing approaches (CLIP-Count [8] and VLCounter [9]) on the protocol FSC-147-S. Our method demonstrates superior text sensitivity, accurately responding to specific counting queries, while these two methods tend to count all objects indiscriminately, regardless of the textual prompt.