/TemplateVersion (2026.1)

CountCluster：用于文本到□像生成的□象□量指□□交叉注意力□聚□

Joohyeon Lee, Jin-Seop Lee, Jee-Hyong Lee*

Sungkyunkwan University, Suwon, South Korea
{ leeju001, wlstjq0602, john } @skku.edu

Abstract

基于□散的文本到□像生成模型在□像□量和多□性方面表□出色。然而 ,□□仍然□以生成准□反映□入提示中指定□量的□象的□像。已□提出了□□方法 ,□些方法依□于外部□□模□□行迭代□化 ,或依□于□已□□的□□或□在特征中得出的□量表示。然而 ,□□在准□反映指定□象□量方面仍然存在局限性 ,□且忽略了一□重要的□□特性——生成□像中的□象□例□量在去□□程的早期□□步□大程度上已被□定。□了正□反映□像生成的□象□量 ,早期□□步中□象交叉注意力□的高激活□域□□□入□象□量相匹配 ,同□每□□域□□□分□。□解□□□□□□ ,我□提出了CountCluster ,一□根据□入中指定的□象□量□□象交叉注意力□□行聚□的方法 ,无需依□任何外部工具或□外的□□。□方法在推理□根据注意力分□□□象交叉注意力□□分□ k □聚□ ,定□一□理想的分布 ,其中每□聚□在空□上良好分□ ,□□化□在□□以□此目□分布□□。□□有方法相比 ,我□的方法在□象□□准□性上平均提高了18.5 % p ,□在各□提示中展示了□□的□量控制性能。代□□在：https://github.com/JoohyeonL22/CountCluster □布。

介□

生成模型在文本到□像生成□域取得了□著的□展 ,其中□□□容完全□自然□言描述中合成。□管早期方法基于GANs (Goodfellow et al. 2014; Radford, Metz, and Chintala 2016; Karras, Laine, and Aila 2019) ,最近□散模型 (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Rombach et al. 2022) 在□像□量和多□性方面表□出更□越的性能。□着□散模型的□展 ,各□□用包括□像合成 (Gu et al. 2022; Saharia et al. 2022; Ramesh et al. 2021) 、修□ (Meng et al. 2022; Lugmayr et al. 2022; Nichol et al. 2022; Couairon et al. 2023) 和□□□□ (Kawar et al. 2023; Tumanyan et al. 2023; Cao et al. 2023) 在□□□景中被越□越多地采用。□一□□突□出生成忠□反映用□意□的□像（包括指定的□□□量）的日益增□的需求。然而 ,□散模型在生成准□反映□入文本中描述的每□□象的□切□量的□像方面仍然存在困□（例如 , “四□□□的照片”) (Paiss et al. 2023; Hu et al. 2023; Wen et al. 2023) 。

最近 ,已□□行了一些□究□解□生成□像中反映物□□量的限制。□有的方法可以分□□□：基于迭代□化的方法和基于□量表示引□的方法。基于迭代□化的方法 (Binyamin et al. 2024; Kang et al. 2025) 在□像生成□程中□□物□□例的□量 ,□通□重新生成□像或迭代□化□像□重新□□□□□□行改□。基于□量表示引□的方法 (Zafar, Wolf, and Schwartz 2024; Zhang, Chen, and Lee 2023) 通□定□新符□□□□□量表示或提取控制□例□量的□在向量□□□□量信息。然后 , 他□利用□量表示□引□其他物□□□生成相同□量的□例。□些方法□示出相□于一□□□的□定□散方法 (Rombach et al. 2022; Podell et al. 2024) 的改□ ,然而 , □□仍然□以生成具有准□□量物□的□像。重要的是 , □□忽略了□量控制中最基本的因素 ,□生成□像中□例□量主要是在□散□程的早期□□步□中□定的。

□于□散模型□□ ,在□象跨注意力□中高度激活的□域——□些□域表明□相□□象□□的強烈□注——□生成□像中□象的位置密切相□ ,如 (Hertz et al. 2023) 所示。如 fig. 1 中的“lemons”□□跨注意力□所示 ,

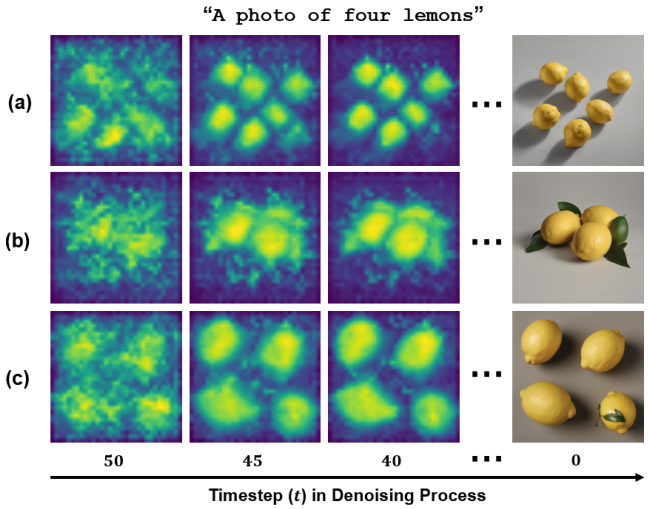

Figure 1: 使用 SDXL (Podell et al. 2024) 生成□像以□□提示“照片中有四□□□”。左□□示了去□□程中□□步 $t = 50, 45, 40$ 的“□□”交叉注意力□ ,而右□□示了□□步 $t = 0$ 的最□生成□像。(a)、(b) 和 (c) 是在相同□件下使用不同□子生成的□果。其中 ,(a) 和 (b) 是□例□量不正□的失□案例 ,而 (c) 是□例□量正□的成功案例。可以□察到 ,□例的位置和□量大多在去□□程的早期□□步中□定。

*Corresponding author

生成□程的早期□□步中 ,物□的位置和□量基本已□定。□早期□□步中□象跨注意力□的高度激活□域的簇□超□□入提示中指定的□象□量□ ,生成的□像往往包含更多的□象 ,如 fig. 1 (a)所示。相反 ,□激活□域之□的□界模糊且重□□ ,在生成□程中可能□合□一些□象 ,□致生成的□象□例□少 ,如 fig. 1 (b)所示。因此 ,□□保生成□像中的□象□量反映文本提示中指定的□量 ,早期□□步中□象跨注意力□中高度激活□域的簇□□准□地□□期□象□□相匹配 ,□且每□□域也□充分□□□□ ,如 fig. 1 (c)所示。

此外 ,□有的方法 (Binyamin et al. 2024; Kang et al. 2025; Zafar, Wolf, and Schwartz 2024; Zhang, Chen, and Lee 2023) 依□于物□□□和布局生成等外部工具 ,或者需要□□□量表示□□□□例□□信息。因此 ,□□□□生大量的□算成本和□□□□ ,使□量控制技□的□□部署具有挑□性。因此 ,有必要在不需要□外模□和□□的情□下□□□量信息。

在本文中 ,我□提出了CountCluster ,□是一□在生成正□□量的□象□例的同□ ,最大限度地□少推理成本和□□的方法 ,□方法不依□于任何外部工具或□外的□□。所提出的方法通□聚□□象的交叉注意力□□指□去□□程 ,以反映□期的□例□量。

□□□□一目□ ,我□的方法旨在□保在去□□程中 ,物□交叉注意力□中高度激活的□域足□分□ ,□匹配□入提示中指定的□例□量。□此 ,我□基于物□交叉注意力□中的注意力得分形成 k □聚□ ,□定□一□目□分布 ,□目□分布鼓□□些聚□在空□上良好分□。然后 ,我□□□一□□失函□ ,引□每□聚□的分布□目□分布□□ ,□在推理□程中基于□目□□用无□□的□在□化。最□ ,交叉注意力□中高度激活的□域□根据指定□量□□分□ ,□而生成具有正□□例□量的□像。

□□有方法相比 ,□方法在□例□□准□率上平均提高了18.5 % p。□□展示了在各□提示下生成□量控制正□的□□□景的能力。重要的是 ,□□通□在去□□程中引□□象的交叉注意力□□操作 ,而不需要任何□外的□□或外部模□ ,□而比以前的方法具有更高的效率和更容易的□用性。

□管最近取得了□展 ,文本到□像模型在准□反映提示中明□指定的□象□量方面仍常常失□。□了解□□一局限性 ,人□提出了各□方法□提高生成□像中□象□量的准□性。其中 ,迭代□化方法在□像生成□程中□□象□例的□量 ,□基于□□的□量反□□化或再生成□像。

□一□□究路□定□了□□□□以□式地□□模型□行□量控制 (Zafar, Wolf, and Schwartz 2024) ,或者利用□量表示引□的方法 (Zhang, Chen, and Lee 2023) ,□ CLIP (Radford et al. 2021) 嵌入空□中提取□□□□信息的□□表示向量。

□然□□□方法都增強了物□□量控制 ,但□□未能解□一□根本的□□性限制:在□像生成□程中交叉注意力□中被激活□域的□量往往□□期的物□□量不匹配。因此 ,□管在□量控制方面有所改□ ,生成的□像依然常常无法忠□地表示指定□量的物□。

无□□□在□化已被□□探索 ,以精准控制□□□文字到□像生成模型的□出。特□是 ,利用交叉注意力□的各□方法已被提出。

□□方法的□点是能□在不修改□□□□散模型的情□下□□□象的存在性。然而 ,□□主要□注的是二□制的存在□否 ,因此在精□控制□如□象□量等定量□性□面□□□性限制。

在□散模型中 ,□物□□□□□的交叉注意力□在□定生成□像中物□的□量和位置方面起着至□重要的作用。□物□交叉注意力□未能形成足□明□的激活□域 ,或者高度激活□域的□量□□入提示中指定的物□□量 (k) 不匹配□ ,生成的□像往往□出□物□□□□□。

□了解□□些□□ ,我□提出了一□无□□方法CountCluster ,□方法□保□象跨注意力□中高度激活□域的聚□□□入中指定的□象□量相匹配 ,同□鼓□□□分□的激活□域。□方法包括以下三□步□:(1)□跨注意力□分割□ k □簇 ,(2)定□一□理想的目□分布 ,其中每□簇□□一□□特的激活□域 ,(3)□行□在□化以使簇分布□定□的目□□□。

□了生成一□包含□入文本中指定□量的物□□例的□像 ,至□重要的是在物□交叉注意力□中□得 k □良好分□的激活□域。□此 ,我□□物□交叉注意力□□行□一化 ,□根据注意力得分□其□分□ k □簇。

首先 ,□考□那些注意力分□大于或等于□定□□□ τ 的□丁□行聚□ ,而低于 τ 的□被排除。如 fig. 1 中的交叉注意□可□化所示 ,只有在□准化交叉注意□中得分足□高的□丁才□□像中□□物□的生成直接相□。此□程防止了无□背景□域或□□被包含在聚□中 ,□□保聚□□□注于□物□生成有□□的有意□激活□域。

□后 ,□□注意力得分最高的前 k □□丁作□聚□中心。□了□保每□聚□代表一□不同的□象而不□其他合□ ,□所□的聚□中心施加了最小距□限制。因此 ,我□定□以下最小距□□束 d :

$$\|\boldsymbol{c}_i - \boldsymbol{c}_j\|_2 \geq d = \frac{H}{k} \quad \text{for all } i \neq j \tag{1}$$

□里 , $\boldsymbol{c}_i$ 和 $\boldsymbol{c}_j$ 表示□□不同聚□中心的位置 ,而 H 指的是注意力□的高度(或□度)。由于本□究生成正方形□像 ,我□假□一□交叉注意力□ A_c ,其中 $H = W$ 。Algorithm 1 □述了在□些□件下□□聚□中心的□程。一旦□定聚□中心后 ,交叉注意力□中的每□□丁都□被分配到其最近的中心□中□□着 k 激活□域形成簇之后 ,我□定□了一□目□分布 ,以促□每□簇□□于一□□特且彼此分□良好的激活□域。

Algorithm 1: □□象交叉注意力□中□□簇中心

Input : Cross-attention map of object token A_c , number of object instances k , minimum distance d
Output : Cluster centers $\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k\}$

Initialize $\mathcal{C} \leftarrow \{\}$
$\mathcal{S} \leftarrow \text{argsort}_{(x,y)} A_c(x, y)$ in descending order
for each $(x, y) \in \mathcal{S}$ do
  if $\forall \boldsymbol{c} \in \mathcal{C}, \|(x, y) - \boldsymbol{c}\|_2 \geq d$ then
    Add (x, y) to $\mathcal{C}$
  end if
  if $|\mathcal{C}| = k$ then
    break
  end if
end for
return $\mathcal{C}$

□□地根据注意力分□□分□域可能□致在□□簇□存在多□子簇 ,□可能□致□□□簇生成多□□象。□□保每□簇只生成一□□象 ,每□簇□必□只有一□主要

密度□域。因此 ,我□定□了一□目□分布 ,其中特征分□□簇中心向□界逐□□少 ,□鼓□□□的交叉注意力分布遵循□□形式。

我□使用二□高斯函□作□目□分布。高斯函□定□□在聚□中心具有□1 ,□在聚□□界衰□至在聚□□程中使用的□□ τ 。□□□□鼓□聚□中的所有片段保持注意力得分高于 τ ,□是□生成□象做出□□所需的最低水平。□□□□旨在□保每□聚□□注于□一□象 ,同□避免□原始交叉注意力□分布的□著失□。□了□建□□高斯分布 ,我□根据聚□中心和聚□□界之□的距□ r □定 σ ,如下所示: □中 ,每□聚□的目□分布定□如下: □里 ,$\boldsymbol{x}$ 表示注意力□中每□片段的坐□ ,$\boldsymbol{\mu}$ 代表聚□中心的位置。

一旦在交叉注意力□中定□了每□聚□及其□□的目□分布 ,我□□□了一□□失函□ ,以鼓□每□聚□的分布接近其目□分布。通□使用KL散度 (Kullback and Leibler 1951) □量交叉注意力□中每□聚□的注意力分布 $Q(x)$ □其□□目□分布 $P(x)$ 的相似性。

□着聚□□量的增加 ,□的KL□失增加 ,而每□□□聚□占据的□像面□□少 ,□□向于□少每□聚□的KL□失。□了考□□一影□ ,□在聚□□量□化□保持□KL□失在一□可比的□模 ,我□通□ $\sqrt{k}$ □□□失□行□一化 ,□定□整□□失如下:

$$
\begin{aligned}
L_{\mathrm{KL}} &= \frac{1}{\sqrt{k}} \sum_{i=1}^{k} D_{\mathrm{KL}}(P_i \| Q_i) \\
&= \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \sum_{\boldsymbol{x}} P_i(\boldsymbol{x}) \log \frac{P_i(\boldsymbol{x})}{Q_i(\boldsymbol{x})}
\end{aligned} \tag{2}
$$

□□失鼓□交叉注意力□形成□□入中的指定□象□□相□□的激活□域 ,□而有助于生成具有准□□象□量的□像。最后 ,□□量 $\boldsymbol{z}_t$ 在去□□程中如下更新:

□了定量□□模型反映物□□量的能力 ,我□通□整合先前物□□□□究中的提示集□□建一□基准 (Binyamin et al. 2024; Kang et al. 2025) 。每□提示的格式□“a photo of [count] [object]” ,其中[count]是一□□“two”到“ten”的□字□□。此□置使得可以□不同物□□□和□量□□行系□□□。

□管基准是用相□□□、注重□量的提示□建的 ,但提出的方法不限于□一集合。如我□的分析所述 ,□可以有效地□展到涉及任意□量、多□物□□型和其他□性的□□提示。

基□方法。 我□□我□提出的方法□以下五□基□方法□行比□ ,包括□□□□用于反映□象□量控制的方法。每□基□都是使用SD2.1或SDXL□□的 ,□果根据基□模型架□□行分□。

SD2.1 (Rombach et al. 2022) 和 SDXL (Podell et al. 2024) 是□□□的□散模型 ,□□□提示中生成□像 ,但□有□式的□量控制。Counting Guidance (Kang et al. 2025) 使用 RCC (Hobley and Prisacariu 2022) 在每□□□步□□物□□量 ,□根据□量差□更新□在表示以指□□像生成。CountGen (Binyamin et al. 2024) 在生成期□□□物□□量 ,如果□□□量□提示中指定的□量不匹配 ,□重新生成□像。□此 ,□使用提出的 ReLayout 模型生成□量□整的掩□ ,然后使用□掩□指□重新生成。Zafar 等人 (Zafar, Wolf, and Schwartz 2024) 使用

Method	Acc.(%) (↑)	MAE. (↓)	RMSE (↓)
SD2.1	23.45	2.235	3.898
Counting Guidance	23.57	2.256	4.522
Ours (SD2.1)	41.87	1.228	2.067
SDXL	27.25	4.350	9.656
Zafar et al.	26.90	1.722	2.675
CountGen	44.80	1.530	3.293
Ours (SDXL)	63.27	0.820	4.670

Table 1: 在□像生成中比□CountCluster(我□的方法)□□有方法的目□□□准□性。使用□□模型□量生成□像中的目□□量 ,□使用准□率、MAE和RMSE指□□□□果。

CLIP-Count (Jiang, Liu, and Chen 2023) □□生成□像中的物□□例 ,□□一□□□□□以反映提示中指定的□量 ,□在推理□使用此□□。

定性□果

Figure 2 展示了由所提出的方法和基准生成的□像的定性比□ ,□明了每□方法在反映□象□量方面的□□效果。

□了□□物□□例□化□□像生成的影□ ,我□生成了□在物□□量上有所不同的提示□像 ,如 fig. 2 (a) 和(b)所示 ,同□保持初始□□量不□。在□□□□中 ,CountCluster准□捕捉到了□提示□□的狗的□量增加 ,生成了一致的□果 ,而□有□著改□□像的整□□□或□格。相比之下 ,□有的方法在生成的□像中 ,□使狗的□量增加 ,表□出的□化也□少甚至□有 ,通常□致□像中的物□□量□提示中描述的不匹配。□表明基□模型在□□□量信息的□□中 ,引入了□弱的交叉注意力 ,而CountCluster□被明□□□□□□□信息更敏感地做出反□。

此外 ,□了□□在不同□象□□□□量控制的一致性 ,我□生成了在提示中指定相同□象□量但□象□型不同的□像 ,如 fig. 2 (c)和(d)所示 ,同□保持初始□在□量不□。CountCluster始□能保持准□的□象□例而不依□于□象□□ ,而□有方法在□量准□性上展示出□著的□化 ,□取□于□象□型。□表明之前的工作更依□于□散模型的□□□先□ ,而不是明□的□量控制。相比之下 ,CountCluster直接操□交叉注意力□ ,以反映提示中指定的□量 ,□且不依□于□象□□。

定量□果

目□□□准□性和□差。 Table 1 提供了 CountCluster 和□有方法在基准□据集上的定量比□。我□的基于 SDXL 的方法(我□的 (SDXL)) 在精度和最低 MAE 上都取得了最高成□ ,比原始 SDXL 提高了 36 % p ,比 CountGen 高出 18.5 % p。□些□果表明 CountCluster 通□交叉注意力□的□□分割有效地控制了□象□□。□然 RMSE 因一些□□模型高□了□象□量的□群情□而高于某些基□ ,但□□指□□□了我□的方法在反映指定□量方面的准□性。

我□的(SD2.1)在准□性上也比原始SD2.1提高了18.4 % p。此外 ,□在所有基于SD2.1的方法中□到了最低的MAE和RMSE ,□一步□□了所提出方法在不同模型架□中通用性和可□展性。

Figure 2: CountCluster在SDXL和其他基□方法上的定性□果。(a)和(b)展示了在相同□□下，□改□文本提示中的□象□量（“六□”→“七□”）的□果。(c)和(d)展示了在相同□□下，□改□目□□象（“□□圈”→“橙子”）的□果。

Method	Counting (%)	Appearing (%)
SD2.1	71.39	99.19
Counting Guidance	71.73	98.89
Ours (SD2.1)	77.95	99.56
SDXL	72.47	99.78
Zafar et al.	74.94	99.11
CountGen	77.01	98.08
Ours (SDXL)	83.16	98.67

Table 2: 使用TIFA□架的VQA□□□果。□□通□□□□分□□□□和出□□型□□行。

Method	Time (s)	Mem. (GB)	Training	External
SD2.1	5.89	7	✗	✗
Counting Guidance	21.27	14	✗	✓
Ours (SD2.1)	7.03	8	✗	✗
SDXL	13.32	15	✗	✗
Zafar et al.	9.07	16 (26 *)	✓	✗
CountGen	88.04	71	✗	✓
Ours (SDXL)	30.96	20	✗	✗

Table 3: 每□方法的推理□□和□存需求，以及是否使用了□外的□□或外部模□。*表示在□外的□□□程中所需的要求。

基于VQA的□象□量□□。 Table 2 展示了使用 TIFA □架在生成□像上的物□□□和存在性□□□果，其中 TIFA □架□合了 GPT-4o mini 作□□□□答模型。TIFA □每幅□像□行□□□型的□□：物□□□和物□存在，□分□□算其准□率作□□□得分和出□得分。

TIFA□□□果表明，CountCluster在□□得分方面比□有方法提高了6 % p，□□了其在□□上反映□量信息的有效性。然而，□□分□相□□低，□可以□因于某些情□下生成的□象□□□提示中指定的不一致。例如，□察到一些情□，□然“□表”被□□□“□□”或“手表”。□些□□表明，□管CountCluster在□量控制方面表□出色，但在需要□格□□一致性的□□中可能□于劣□。

推理效率和□外□源需求。 Table 3 提出了一□在各□方法中比□推理□□、□存消耗、□外□□需求以及推理□外部模□依□的方法。所有□量均在使用四□NVIDIA RTX 3090 GPU的基准上□行。

由于使用梯度下降更新□在表示，CountCluster比SD2.1和SDXL需要更□的□像生成□□。□□□□在SDXL中尤□明□，因□交叉注意□的□量增加，□致□

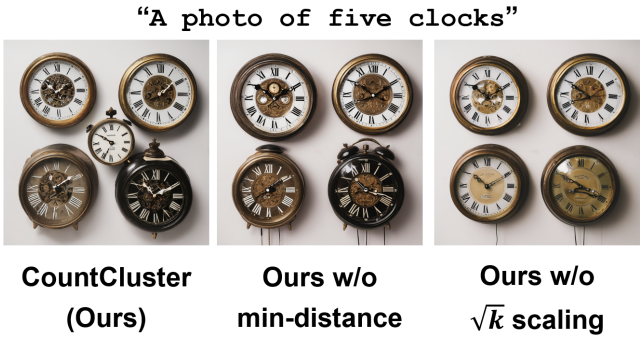

Figure 3: 使用提示□“一□五□□表的照片”生成的□像。(左)CountCluster。(中)□有聚□中心之□的最小距□□束。(右)□用□失□放 ,使用 k 而不是 $\sqrt{k}$ 。

Method	Acc.(%) (↑)	MAE. (↓)	RMSE (↓)
Ours (SDXL)	63.27	0.820	4.670
w/o min-distance	47.13	1.430	9.067
w/o $\sqrt{k}$ scaling	44.33	1.249	3.212

Table 4: CountCluster□□□件的消融□究。

算和□存需求大幅增加 ,□且速度明□降低。相比之下 ,由于□像再生 ,CountGen的推理速度比CountCluster慢□ 2.8× 倍 ,而由于反□使用外部模□ ,Counting Guidance□慢□ 3× 倍。CountGen□□的推理□□□因于在四□GPU上□行部署□出□的□□ ,表明CountGen是□□高性能□GPU或□化的分布式□境□□的。

此外 ,Zafar 等人提出的方法通□省略无分□器指□(Ho and Salimans 2022) □□□更快的生成速度;然而 ,□需要□□一□□□□□ ,□大□需要 214 秒的□□□□和 26 GB 的□存 ,□而限制了其□用性。

□□底分析所提出方法的有效性 ,□行了消融□究以□□簇中心之□最小距□□束的影□以及基于簇□量的KL□失□放方法的影□。□□利用了□□模型□量 ,□之前的□□一致。

群集中心之□的最小距□□束的影□。 所提出的方法□聚□中心施加了最小距□□束 d ,以防止□□□象跨越多□聚□。□□□果表明 ,□□用此最小距□□束□ ,生成的□像能□准□反映提示所□□的□象□量 ,如在 fig. 3 中所示。□是因□每□聚□□□一□□立的□象。

相比之下 ,如果□有最小距□□束 ,一□□□的物□可能□跨越多□簇 ,□致物□□量少于□□的簇□ ,□降低物□□□的准□性。如 table 4 所示 ,□□致准□率下降了 16 %

基于 $\sqrt{k}$ 的□失□放效果。 CountCluster 通□□KL散度□失的□和除以 $\sqrt{k}$ □□ k □□行□一化。□□□一化□□了□着□象□□ k 的增加 ,每□□占据的空□□域□小的□□ ,□而□致□KL□失增加而每□□的□失□少。

□□KL□失通□除以 k □行□一化□ ,□着聚□□量的增加 ,□失□度□少 ,□致□在表示的欠□化 ,□而□生的□象□量少于□入中指定的□量 ,如 fig. 3 所示。此外 ,table 4 □示相□于□用 $\sqrt{k}$ □放□ ,□量反射准□

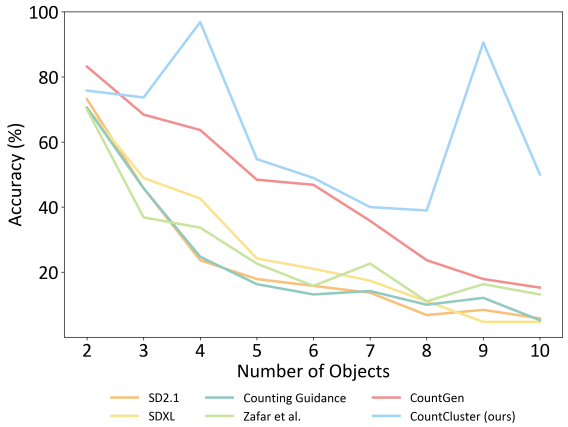

Figure 4: 根据物□□量□化比□生成□像中的□量准□性。

性下降了19□百分点。□表明 $\sqrt{k}$ □放适□地平衡了KL□失 ,有助于更□定和准□地反映□象□量。

分析

□象□□的定量准□性分析。 Figure 4 展示了基于基准□据集中指定□象□量的方法□的□量准□性比□。□□象□量□□□□ ,CountCluster 的准□性略低于CountGen ,但在□象□量□三□或更多□ ,比所有□有方法都表□出更高的准□性。□得注意的是 ,在□象□量□四和九□的准□性特□高 ,□可以□因于在生成分辨率□平方(512 × 512 ,1024 × 1024)的□像 ,□分□以2 × 2和3 × 3□格排列□ ,更容易分□激活□域。

□有的方法在物□□量□九□也表□出相□□高的准□性 ,□表明□□□□散模型可能□空□模式(例如3 × 3□格)□□熟悉。

□了□□CountCluster在□□提示上的有效性 ,我□□行了□外的定性分析。?? □□□上展示了 CountCluster 在□些□□提示中的□量控制□果。

生成的□像表明 ,交叉注意力地□可以可□地控制 ,以保持准□的□象□量 ,无□提示□成或□象□性的□化。□表明CountCluster具有強大的□量控制能力 ,可以在各□提示中□好地泛化。

在本文中 ,我□介□了CountCluster ,□是一□无需□□的方法 ,能□在基于□散的文本生成□像中□□象□量□行精□控制。由于在去□□程的早期□□步中生成□像中的□象□量大多被□定 ,我□提出的方法引□□象的交叉注意力□形成□□分□的簇 ,□指定的□象□量相匹配。在不需要任何外部模□或□外□□的情□下 ,我□的方法在□象□量精度上取得了□著的改□ ,同□保持了□低的推理成本。□些□果突出了注意力□干□在生成模型中的□量控制方面的有效性和□用性。

References

Amini-Naieni, N.; Han, T.; and Zisserman, A. 2025. COUNTGD: multi-modal open-world counting. In Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Binyamin, L.; Tewel, Y.; Segev, H.; Hirsch, E.; Rassin, R.; and Chechik, G. 2024. Make It Count: Text-to-Image Generation with an Accurate Number of Objects. arXiv preprint arXiv:2406.10210.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 22560–22570.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. ACM Trans. Graph., 42(4).

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2023. DiffEdit: Diffusion-based semantic image editing with mask guidance. In The Eleventh International Conference on Learning Representations.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat GANs on image synthesis. In Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713845393.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10696–10706.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In The Eleventh International Conference on Learning Representations.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., Advances in Neural Information Processing Systems, volume 33, 6840–6851. Curran Associates, Inc.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598.

Hobley, M.; and Prisacariu, V. 2022. Learning to Count Anything: Reference-less Class-agnostic Counting with Weak Supervision. arXiv:2205.10203.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20406–20417.

Jiang, R.; Liu, L.; and Chen, C. 2023. CLIP-Count: Towards Text-Guided Zero-Shot Object Counting. In Proceedings of the 31st ACM International Conference on Multimedia, MM '23, 4535–4545. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Kang, W.; Galim, K.; Koo, H. I.; and Cho, N. I. 2025. Counting guidance for high fidelity text-to-image synthesis. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 899–908. IEEE.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In Conference on Computer Vision and Pattern Recognition 2023.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. The Annals of Mathematical Statistics, 22(1): 79–86.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11461–11471.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In International Conference on Learning Representations.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, 16784–16804. PMLR.

OpenAI; :; Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; Kirillov, A.; Nichol, A.; Paino, A.; Renzin, A.; Passos, A. T.; Kirillov, A.; Christakis, A.; Conneau, A.; Kamali, A.; Jabri, A.; Moyer, A.; Tam, A.; Crookes, A.; Tootoochian, A.; Tootoonchian, A.; Kumar, A.; Vallone, A.; Karpathy, A.; Braunstein, A.; Cann, A.; Codispoti, A.; Galu, A.; Kondrich, A.; Tulloch, A.; Mishchenko, A.; Baek, A.; Jiang, A.; Pelisse, A.; Woodford, A.; Gosalia, A.; Dhar, A.; Pantuliano, A.; Nayak, A.; Oliver, A.; Zoph, B.; Ghorbani, B.; Leimberger, B.; Rossen, B.; Sokolowsky, B.; Wang, B.; Zweig, B.; Hoover, B.; Samic, B.; McGrew, B.; Spero, B.; Giertler, B.; Cheng, B.; Lightcap, B.; Walkin, B.; Quinn, B.; Guarraci, B.; Hsu, B.; Kellogg, B.; Eastman, B.; Lugaresi, C.; Wainwright, C.; Bassin, C.; Hudson,

C.; Chu, C.; Nelson, C.; Li, C.; Shern, C. J.; Conger, C.; Barette, C.; Voss, C.; Ding, C.; Lu, C.; Zhang, C.; Beaumont, C.; Hallacy, C.; Koch, C.; Gibson, C.; Kim, C.; Choi, C.; McLeavey, C.; Hesse, C.; Fischer, C.; Winter, C.; Czarnecki, C.; Jarvis, C.; Wei, C.; Koumouzelis, C.; Sherburn, D.; Kappler, D.; Levin, D.; Levy, D.; Carr, D.; Farhi, D.; Mely, D.; Robinson, D.; Sasaki, D.; Jin, D.; Valladares, D.; Tsipras, D.; Li, D.; Nguyen, D. P.; Findlay, D.; Oiwoh, E.; Wong, E.; Asdar, E.; Proehl, E.; Yang, E.; Antonow, E.; Kramer, E.; Peterson, E.; Sigler, E.; Wallace, E.; Brevdo, E.; Mays, E.; Khorasani, F.; Such, F. P.; Raso, F.; Zhang, F.; von Lohmann, F.; Sulit, F.; Goh, G.; Oden, G.; Salmon, G.; Starace, G.; Brockman, G.; Salman, H.; Bao, H.; Hu, H.; Wong, H.; Wang, H.; Schmidt, H.; Whitney, H.; Jun, H.; Kirchner, H.; de Oliveira Pinto, H. P.; Ren, H.; Chang, H.; Chung, H. W.; Kivlichan, I.; O'Connell, I.; O'Connell, I.; Osband, I.; Silber, I.; Sohl, I.; Okuyucu, I.; Lan, I.; Kostrikov, I.; Sutskever, I.; Kanitscheider, I.; Gulrajani, I.; Coxon, J.; Menick, J.; Pachocki, J.; Aung, J.; Betker, J.; Crooks, J.; Lennon, J.; Kiros, J.; Leike, J.; Park, J.; Kwon, J.; Phang, J.; Teplitz, J.; Wei, J.; Wolfe, J.; Chen, J.; Harris, J.; Varavva, J.; Lee, J. G.; Shieh, J.; Lin, J.; Yu, J.; Weng, J.; Tang, J.; Yu, J.; Jang, J.; Candela, J. Q.; Beutler, J.; Landers, J.; Parish, J.; Heidecke, J.; Schulman, J.; Lachman, J.; McKay, J.; Uesato, J.; Ward, J.; Kim, J. W.; Huizinga, J.; Sitkin, J.; Kraaijeveld, J.; Gross, J.; Kaplan, J.; Snyder, J.; Achiam, J.; Jiao, J.; Lee, J.; Zhuang, J.; Harriman, J.; Fricke, K.; Hayashi, K.; Singhal, K.; Shi, K.; Karthik, K.; Wood, K.; Rimbach, K.; Hsu, K.; Nguyen, K.; Gu-Lemberg, K.; Button, K.; Liu, K.; Howe, K.; Muthukumar, K.; Luther, K.; Ahmad, L.; Kai, L.; Itow, L.; Workman, L.; Pathak, L.; Chen, L.; Jing, L.; Guy, L.; Fedus, L.; Zhou, L.; Mamitsuka, L.; Weng, L.; McCallum, L.; Held, L.; Ouyang, L.; Feuvrier, L.; Zhang, L.; Kondraciuk, L.; Kaiser, L.; Hewitt, L.; Metz, L.; Doshi, L.; Aflak, M.; Simens, M.; Boyd, M.; Thompson, M.; Dukhan, M.; Chen, M.; Gray, M.; Hudnall, M.; Zhang, M.; Aljubeh, M.; Litwin, M.; Zeng, M.; Johnson, M.; Shetty, M.; Gupta, M.; Shah, M.; Yatbaz, M.; Yang, M. J.; Zhong, M.; Glaese, M.; Chen, M.; Janner, M.; Lampe, M.; Petrov, M.; Wu, M.; Wang, M.; Fradin, M.; Pokrass, M.; Castro, M.; de Castro, M. O. T.; Pavlov, M.; Brundage, M.; Wang, M.; Khan, M.; Murati, M.; Bavarian, M.; Lin, M.; Yesildal, M.; Soto, N.; Gimelshein, N.; Cone, N.; Staudacher, N.; Summers, N.; LaFontaine, N.; Chowdhury, N.; Ryder, N.; Stathas, N.; Turley, N.; Tezak, N.; Felix, N.; Kudige, N.; Keskar, N.; Deutsch, N.; Bundick, N.; Puckett, N.; Nachum, O.; Okelola, O.; Boiko, O.; Murk, O.; Jaffe, O.; Watkins, O.; Godement, O.; Campbell-Moore, O.; Chao, P.; McMillan, P.; Belov, P.; Su, P.; Bak, P.; Bakkum, P.; Deng, P.; Dolan, P.; Hoeschele, P.; Welinder, P.; Tillet, P.; Pronin, P.; Tillet, P.; Dhariwal, P.; Yuan, Q.; Dias, R.; Lim, R.; Arora, R.; Troll, R.; Lin, R.; Lopes, R. G.; Puri, R.; Miyara, R.; Leike, R.; Gaubert, R.; Zamani, R.; Wang, R.; Donnelly, R.; Honsby, R.; Smith, R.; Sahai, R.; Ramchandani, R.; Huet, R.; Carmichael, R.; Zellers, R.; Chen, R.; Chen, R.; Nigmatullin, R.; Cheu, R.; Jain, S.; Altman, S.; Schoenholz, S.; Toizer, S.; Miserendino, S.; Agarwal, S.; Culver, S.; Ethersmith, S.; Gray, S.; Grove, S.; Metzger, S.; Hermani, S.; Jain, S.; Zhao, S.; Wu, S.; Jomoto, S.; Wu, S.; Shuaiqi; Xia; Phene, S.; Papay, S.; Narayanan, S.; Coffey, S.; Lee, S.; Hall, S.; Balaji, S.; Broda, T.; Stramer, T.; Xu, T.; Gogineni, T.; Christianson, T.; Sanders, T.; Patwardhan, T.; Cunninghman, T.; Degry, T.; Dimson, T.; Raoux, T.; Shadwell, T.; Zheng, T.; Underwood, T.; Markov, T.; Sherbakov, T.; Rubin, T.; Stasi, T.; Kaftan, T.; Heywood, T.; Peterson, T.; Walters, T.; Eloundou, T.; Qi, V.; Moeller, V.; Monaco, V.; Kuo, V.; Fomenko, V.; Chang, W.; Zheng, W.; Zhou, W.; Manassra, W.; Sheu, W.; Zaremba, W.; Patil, Y.; Qian, Y.; Kim, Y.; Cheng, Y.; Zhang, Y.; He, Y.; Zhang, Y.; Jin, Y.; Dai, Y.; and Malkov, Y. 2024. GPT-4o System Card. arXiv:2410.21276.

Paiss, R.; Ephrat, A.; Tov, O.; Zada, S.; Mosseri, I.; Irani, M.; and Dekel, T. 2023. Teaching CLIP to Count to Ten. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3170–3180.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In The Twelfth International Conference on Learning Representations.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, 8748–8763. PMLR.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In Meila, M.; and Zhang, T., eds., Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.;

Cho, K.; and Oh, A., eds., Advances in Neural Information Processing Systems, volume 35, 36479–36494. Curran Associates, Inc.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1921–1930.

Wen, S.; Fang, G.; Zhang, R.; Gao, P.; Dong, H.; and Metaxas, D. 2023. Improving Compositional Text-to-image Generation with Large Vision-Language Models. arXiv:2310.06311.

Zafar, O.; Wolf, L.; and Schwartz, I. 2024. Iterative Object Count Optimization for Text-to-image Diffusion Models. arXiv:2408.11721.

Zhang, R.; Chen, Y.; and Lee, K. 2023. Zero-shot Improvement of Object Counting with CLIP. In R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models.

□充材料□

CountCluster：利用交叉注意力□聚□的无□□□象□量引□用于文本到□像生成

□□□□

在本□究中，我□□施了自己方法□□定了所有超□□，□考了Attend-and-Excite (Chefer et al. 2023) 的□散管道代□。提出的□量控制方法基于□□□的SD2.1 (Rombach et al. 2022) 和SDXL (Podell et al. 2024) 模型。利用 inflect □□提示中自□提取□象□□及其□量。

在□像生成□程中，□散模型的引□尺度 (Ho and Salimans 2022) 固定□ 7.5。在所提出的方法中，高斯平滑使用大小□ 3 和 $\sigma = 0.5$ 的核□用于□象交叉注意力□，□后□行最小-最大□一化。在聚□步□中，我□□定了注意力分□□□ $\tau = 0.3$。

在去□□程的前10□□□步（$t \in 50, \ldots, 40$）中，更新了□像□在空□。□于我□的模型（SD2.1），比例因子 α 被□置□ 40，而□于我□的模型（SDXL），□被□置□ 75,000。此外，在 t_{50} □，□行了迭代□化，直到□失降到低于 0.2，在 t_{40} □，□化□□□行，直到□失□到 0.15 或更低，以□一步□化□在空□。

□于□像生成，CountGen (Binyamin et al. 2024) 在推理□使用了四□ NVIDIA RTX 3090 GPU □行□行，而所有其他□□均在一□ RTX 3090 GPU 上□行。此外，□于 Zafar 等人的方法，□□□□的□□使用了□□ RTX 3090 GPU。

□于我□的□□，我□通□整合 Counting Guidance (Kang et al. 2025) 和CountGen (Binyamin et al. 2024) 中提出的提示□建了一□基准提示集，使用的提示格式□“[□量]□[物□]的照片”。□里，[count]是一□□二到十的英文□□，[object]□ ?? 中列出的19□□定□□□中□□。

□□指□

□了□□生成□像中物□的□量，我□使用了 CountGD (Amini-Naieni, Han, and Zisserman 2025)，自□□量每幅□像中□提示中指定物□相□□的□例□量。□ CountGD □得的□□□例□量□后□提示中提供的目□□量□行比□，以□算以下□□指□：准□率 (Acc)、平均□□□差 (MAE) 和均方根□差 (RMSE)。每□指□的公式如下：

$$\text{Acc} = \sum_{i=1}^{n} \frac{\mathbf{1}[y_i = \hat{y}_i]}{n} \tag{3}$$

$$\text{MAE} = \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{n} \tag{4}$$

□里，y_i 表示在第 i □提示中指定的目□□象□□，$\hat{y}_i$ 是 CountGD □□的第 i □生成□像中相□□象□例的□量。n 代表生成□像的□□。在我□的□□中，我□□□了每□方法在□共 1,710 □□像上的表□，□些□像是通□□ 171 □□□和□象□合□用 10 □□机□子 (0 到 9) 生成的。

此外，我□采用了TIFA□架 (Hu et al. 2023)，□利用 GPT-4o mini (OpenAI et al. 2024) 作□VQA模型，□合□□□□得分和出□得分。TIFA得分□算□GPT-4o mini基于生成的□像提供正□答案的VQA□□比例：

$$\text{TIFA Score} = \sum_{i=1}^{n} \frac{\mathbf{1}[A_i = \hat{A}_i]}{n} \tag{5}$$

□里，A_i 表示 i -th□□的□□答案，$\hat{A}_i$ 代表GPT-4o mini生成的答案，n 是□答□的□□。

在TIFA□架中□□使用的□□答案□以 ?? 和 ?? □例：

□了提供□外的定性比□，我□□CountCluster□□□□□散模型SD2.1 (Rombach et al. 2022) 和SDXL (Podell et al. 2024) □行比□，以及□□有的□象□量控制方法，包括Counting-guidance (Kang et al. 2025)、CountGen (Binyamin et al. 2024) 和Zafar等人的方法 (Zafar, Wolf, and Schwartz 2024) □行比□。

如 fig. 6 和 fig. 7 所示，□有方法生成的□像中的□例□量常常□提示□格不符，或者□□□象□有明□分□。相比之下，CountCluster 生成的□像不□能准□生成指定□量的□例，而且每□□例都能□楚□分。

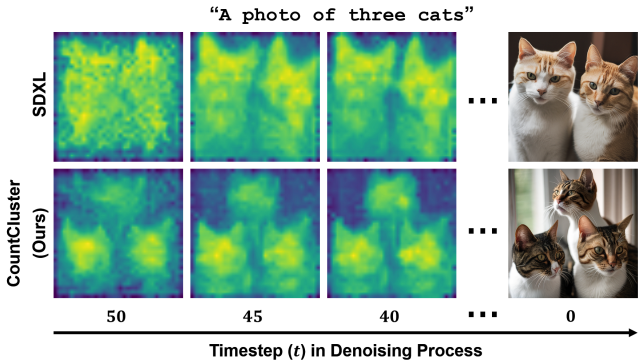

Figure 5: 在早期□□步的□象交叉注意力□和□自SDXL□CountCluster的生成□像

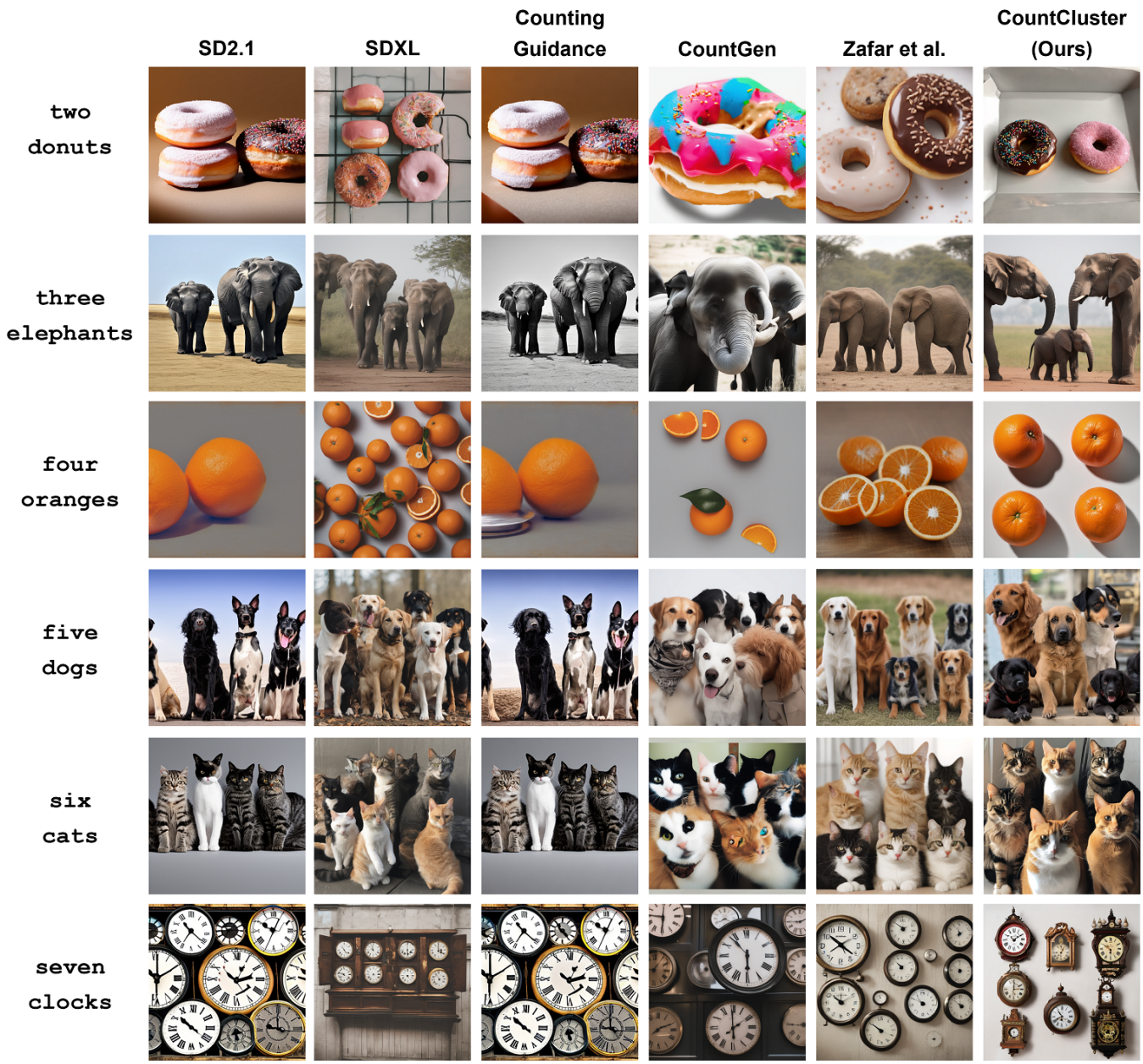

Figure 6: □CountCluster、SD2.1、SDXL、Counting Guidance、CountGen和Zafar等□行更多□性比□

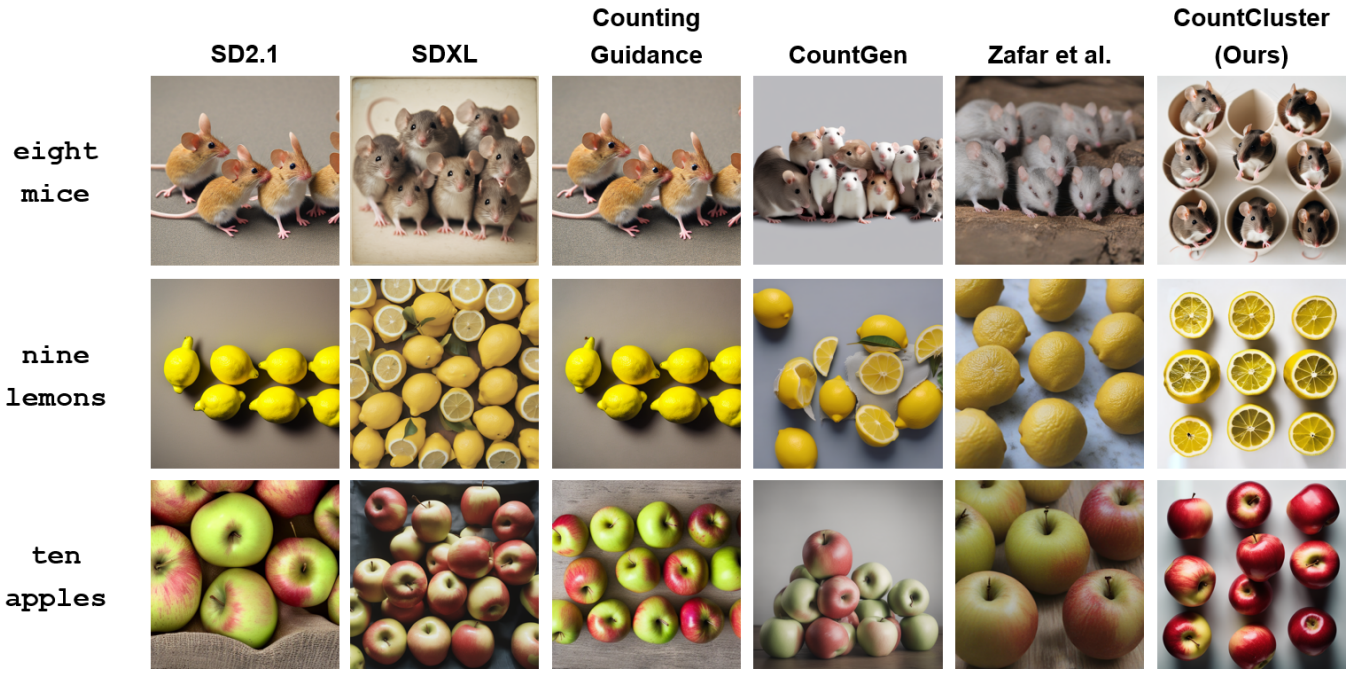

Figure 7: CountCluster和SD2.1、SDXL、Counting Guidance、CountGen以及Zafar等方法的更多定性比□。