β-CLIP: Text-Conditioned Contrastive Learning for Multi-Granular Vision-Language Alignment

Supplementary Material

1. Experiments on ViT-L/14

We evaluate the scalability of β-CLIP using the larger CLIP ViT-L/14 @224px backbone, while keeping all other training settings fixed. Results are reported in Tables 1 and 2.

Fine-Grained Retrieval. On FG-OVD, the CE variant benefits significantly from higher granularity ($K = 36$), achieving 28.3–28.7% on the Hard split (+13.0 over CLIP) and 70.9–72.2% on Trivial (+32–33 over CLIP). We observe consistent gains of +8–10 when increasing K from 6 to 36, validating the effectiveness of the soft-target CE formulation for multi-granular alignment. In this setting, $\beta = 0.75$ consistently outperforms $\beta = 0.5$ by a small margin. Interestingly, the BCE variant is not as effective on FG-OVD despite the larger backbone.

Among methods that train on image–text pairs without region crops or hard negatives, β-CLIP CE achieves the highest performance, outperforming EVA-CLIP (18.3 Hard) and LongCLIP (9.6 Hard). It also performs better than FineCLIP (22.8 Hard), which uses cropped regions during training. The current state-of-the-art method, FG-CLIP (48.4 Hard), combines region crops with targeted hard-negative mining. Even so, β-CLIP CE, trained solely on decomposed long captions, reaches a significant fraction of its performance without explicit region level supervision.

Long-Text Retrieval The pattern reverses on long-caption retrieval tasks between CE and BCE. Across the DCI, Urban1K, and ShareGPT4V-1K benchmarks, the BCE variant clearly outperforms CE, achieving 69.1–70.3% T2I on the challenging DCI benchmark (versus 59.9–65.2% for CE) and up to 94.9% on Urban1K. This is consistent with our observations for the ViT-B/16 experiments which also show BCE's particular effectiveness for long-text image retrieval tasks. Increasing hierarchical granularity from K=6 to K=36 remains beneficial for CE on these tasks (+4–9 on DCI), as additional fine-grained phrases improve the fine-grained alignment of the nouns, actions, and spatial relations that are characteristic of DCI captions.

Across both ViT-B/16 and ViT-L/14 backbones, the results follow a consistent pattern. The CE variant outperforms BCE on fine-grained retrieval, whereas BCE excels at long-caption retrieval. In nearly all configurations, β=0.75 also yields small but reliable gains over β=0.5. This outcome reflects how β moderates overlap among the intra-image positives. Increasing from 0.5 to 0.75 sufficiently increases useful contextual alignment from the multi-granular semantics.

Table 1. **Fine-Grained Retrieval Scalability (ViT-L/14).** Comparison on the FG-OVD benchmark. The Cross-Entropy (CE) variant significantly outperforms the Binary Cross-Entropy (BCE) variant. Performance consistently improves with higher hierarchical granularity ($K = 36$) and stronger intra-image contextualization. ($\beta = 0.75$). ‡ denotes FG-CLIP is trained with hard negatives.

Model	FG-OVD			
	Hard	Medium	Easy	Trivial
CLIP	15.4	25.3	25.7	38.8
EVA-CLIP	18.3	38.4	35.2	62.7
LongCLIP	9.6	19.7	16.0	39.8
FineCLIP	22.8	46.0	46.0	**73.6**
FG-CLIP‡	48.4	69.5	71.2	89.7
β-CLIP *CE*				
K=6, β=0.5	19.3	37.9	41.1	62.2
K=36, β=0.5	<u>28.3</u>	<u>48.8</u>	**53.3**	70.9
K=6, β=0.75	20.3	41.1	44.3	67.9
K=36, β=0.75	**28.7**	**49.1**	<u>52.1</u>	<u>72.2</u>
β-CLIP *BCE*				
K=6, β=0.5	15.1	34.6	29.2	32.4
K=36, β=0.5	14.4	29.4	23.7	45.6
K=6, β=0.75	15.0	36.3	33.1	36.8
K=36, β=0.75	15.4	31.5	23.9	45.9

2. Larger Batch Sizes

In contrastive vision-language pre-training, larger batch sizes are largely beneficial because they expose each image to more negative examples, which helps to improve global alignment. For β-CLIP, however, we note that a per-GPU batch size of B does not correspond to a standard $B \times B$ similarity matrix. Because each image is paired with K captions of varying granularity, each producing a text-conditioned image embedding, the resulting similarity matrix is of size $BK \times BK$. This already provides more contrastive signal than a standard CLIP setup with the same nominal batch size.

We ablate β-CLIP ViT-B/16 (β=0.5, K=6 or K=36) backbone on larger batch sizes from 64 to 96, and 112 (effective batch sizes 2048, 3072, 3584 respectively). As shown in Tables 3 and 4, increasing batch size consistently degrades fine-grained retrieval performance, with the CE variant dropping up to 3.4 points on the Hard split, while long-caption retrieval remains relatively stable, occasion-

Table 2. **Long-Text Retrieval Scalability (ViT-L/14).** Comparison on the DCI, Urban1k, and ShareGPT4V-1K long-text retrieval benchmarks. In contrast to fine-grained tasks, the BCE variant consistently outperforms CE. ‡ denotes FG-CLIP is trained with hard negatives.

Model	DCI		Urban1k		ShareGPT4V	
	T2I	I2T	T2I	I2T	T2I	I2T
CLIP	36.4	37.2	90.7	89.2	83.6	86.5
EVA-CLIP	47.8	47.2	-	-	89.4	91.5
LongCLIP	52.5	44.2	-	-	95.6	95.8
FineCLIP	46.2	40.1	-	-	82.7	73.4
TULIP	56.4	55.7	91.1	90.1	**99.0**	**99.0**
SmartCLIP	-	-	90.1	93.0	98.5	97.9
FineLIP	-	-	93.9	94.5	-	-
FG-CLIP‡	66.1	66.7	-	-	96.8	97.4
β-CLIP *CE*						
K=6, β=0.5	59.9	53.2	91.5	91.9	93.4	94.2
K=36, β=0.5	65.2	62.4	93.2	93.2	94.1	94.3
K=6, β=0.75	60.3	55.7	91.9	92.3	93.4	94.2
K=36, β=0.75	64.8	62.6	92.7	92.8	94.1	94.1
β-CLIP *BCE*						
K=6, β=0.5	69.7	67.0	94.3	93.7	94.1	94.3
K=36, β=0.5	69.4	68.2	94.6	**94.6**	94.1	94.3
K=6, β=0.75	69.1	67.0	94.0	94.1	94.1	94.4
K=36, β=0.75	**70.3**	**68.5**	**94.9**	94.3	94.2	94.3

ally showing minor gains or losses.

Additionally, the performance may also be effected due to asymmetric scaling of the two loss terms. The global CLS contrastive loss strengthens with batch size, benefiting disproportionately from the increased negatives for *inter-image* discrimination. Although the fine-grained multi-granular loss also receives more cross-batch negatives, these additional negatives appear to be diluting the signal from the *intra-image* positives. We speculate that this causes the optimization to prioritize separating distinct images over aligning the subtle, multi-granular intra-image features, similar to the effect of reducing K in our method. Thus, given the already-expanded $BK \times BK$ similarity matrix, larger batches strengthen the coarse image-level objective at the expense of fine-grained correspondences.

3. Distance-Calibrated Intra-Image Weights.

Table 5 evaluates the effect of replacing uniform intra-image positive weights with exponentially distance-calibrated weights in β-CLIP. Across CE and BCE, and for both K=6 and K=36, the calibration yields only minor fluctuations in FG-OVD and long-text retrieval, with no consistent gains. These results suggest that once β determines the overall contextualization strength, the precise allocation of mass across intra-image positives has limited effect, and uniform weighting is already sufficient.

Table 3. **Batch Size Ablation (CE Variants).** Effect of increasing batch size on retrieval performance using the ViT-B/16 backbone. Larger batch sizes result in a noticeable degradation in fine-grained retrieval (FG-OVD) capabilities, while global long-text retrieval remains relatively stable.

BS	FG-OVD				U-1K		SV-1K	
	Hard	Medium	Easy	Trivial	T2I	I2T	T2I	I2T
β-*CLIP K=6*, $\beta = 0.5$								
64	29.2	51.3	56.4	81.5	87.9	88.4	93.5	94.0
96	27.7	50.6	57.0	81.4	88.0	89.0	93.1	94.3
112	25.8	48.9	54.8	80.3	88.4	89.5	93.7	94.3
β-*CLIP K=36*, $\beta = 0.5$								
64	30.9	55.4	60.4	80.3	89.0	88.6	93.7	94.0
96	30.4	55.2	61.7	80.9	88.8	90.5	93.9	94.2
112	28.3	53.3	58.8	80.4	89.4	89.2	94.1	94.1

Table 4. **Batch Size Ablation (BCE Variants).** Effect of increasing batch size for the BCE variant using the ViT-B/16 backbone. Increasing the batch size tends to negatively impact fine-grained performance (FG-OVD) while favoring global image-text alignment metrics.

BS	FG-OVD				U-1K		SV-1K	
	H	M	E	T	T2I	I2T	T2I	I2T
β-*CLIP K=6*, $\beta = 0.5$								
64	20.6	42.7	45.3	71.2	91.8	92.0	94.5	94.0
96	18.4	37.3	37.5	70.0	92.5	93.3	94.2	94.3
112	17.0	33.9	31.9	71.1	92.6	92.8	94.2	94.3
β-*CLIP K=36*, $\beta = 0.5$								
64	20.1	38.5	34.2	71.3	91.8	92.3	94.4	94.1
96	17.1	33.3	31.4	73.5	92.3	93.3	94.1	94.2
112	14.8	27.6	24.7	71.0	92.7	93.0	94.0	94.1

Method	FG-OVD				SV-1k		U-1k		
ViT-B/16	Hard	Medium	Easy	Trivial	T2I	I2T	T2I	I2T	Sim
β-*CLIP (CE)*									
K=6, β=0.75	30.6	52.6	58.5	82.1	93.4	93.9	87.3	88.7	0.97
K=36, β=0.75	30.7	54.2	60.0	80.5	93.7	94.1	88.6	88.7	0.98
+ *distance-calibrated weights*									
K=6, β=0.75	29.9	51.7	58.2	81.4	93.3	94.1	88.1	88.2	0.97
K=36, β=0.75	30.3	54.3	60.1	80.6	93.6	94.1	89.1	88.3	0.98
β-*CLIP (BCE)*									
K=6, β=0.75	20.6	42.4	45.7	71.8	94.3	94.0	91.7	91.4	0.94
K=36, β=0.75	19.8	38.0	34.2	72.8	94.3	93.8	91.8	91.8	0.98
+ *distance-calibrated weights*									
K=6, β=0.75	20.1	41.8	45.6	71.3	94.4	94.4	91.3	91.5	0.95
K=36, β=0.75	21.2	40.5	36.3	72.6	94.3	94.1	91.6	91.4	0.98

Table 5. Effect of distance-calibrated intra-image positive weights at β=0.75. Calibrated targets decay with scale distance (e.g., for K=6: $1.00, 0.71, 0.68, 0.65 \ldots$), in contrast to the uniform BCE weights $(1.00, 0.75, 0.750.75 \ldots)$

4. Effect of Varying β at K=6

Table 6 complements the K=36 analysis in the main paper (Table **??**) by showing the effect of β at a lower hierarchy size. The same specificity–contextualization trade-

Method	FG-OVD				SV-1k		U-1k		
ViT-B/16	Hard	Medium	Easy	Trivial	T2I	I2T	T2I	I2T	Sim
β-CLIP (CE)									
K=6, β=0	5.3	10.1	9.9	32.5	**95.0**	**94.8**	**91.8**	**91.0**	0.17
K=6, β=0.25	28.3	50.0	56.1	79.8	93.2	93.9	87.9	88.8	0.93
K=6, β=0.5	29.2	51.3	56.4	81.5	93.5	94.0	87.9	88.4	0.96
K=6, β=0.75	**30.6**	**52.6**	**58.5**	**82.1**	93.4	93.9	87.3	88.7	0.97
K=6, β=1.0	29.7	**52.6**	58.3	81.4	93.5	93.9	87.1	88.3	0.98
β-CLIP (BCE)									
K=6, β=0	16.1	29.9	29.0	65.6	94.0	**94.2**	**92.1**	**92.5**	0.56
K=6, β=0.25	**21.7**	41.8	45.3	70.0	94.1	**94.2**	91.2	92.4	0.92
K=6, β=0.5	20.6	**42.7**	45.3	71.2	**94.5**	94.0	91.8	92.0	0.94
K=6, β=0.75	20.6	42.4	**45.7**	71.8	94.3	94.0	91.7	91.4	0.94
K=6, β=1	18.8	41.7	44.0	**72.6**	94.2	**94.2**	91.6	91.7	0.95

Table 6. Effect of varying β at $K = 6$. β controls the specificity-contextualization tradeoff between diagonal specificity and off-diagonal contextualization.

off holds: increasing β from 0 improves fine-grained performance for both CE and BCE, with only minor drops in long-text retrieval.

5. Qualitative Analysis

We visualize patch-text logit similarities across distinct scenarios: short phrases isolation (Fig. 1) and long-caption grounding (Fig. 2 and Fig. 3). Across these settings, distinct trends emerge when comparing the baselines to the Cross-Entropy (CE) and Binary Cross-Entropy (BCE) variants of β-CLIP.

Baseline Saliency Both the pre-trained CLIP and CLIP fine-tuned on long-text (CLIP FT) exhibit a strong *saliency bias*. Rather than localizing the specific text query, they tend to focus disproportionately on the most visually salient or high-contrast features. This is most evident in the Fig. 3, where baselines consistently focus on the bird's beak even when queried about "wings" or "body." Similarly, in short-concept scenarios (Fig. 1), baseline similarities frequently encompass unrelated background regions like grass, sky, or walls, showing a reliance on global image features rather than explicit region-text correspondences.

Semantic Disentanglement A primary distinction between our proposed loss variants is the density of their heatmaps and their ability to separate semantically related concepts.

- **Cross-Entropy (CE):** The CE variant yields highly sparse and sharp localizations. In Fig. 1, CE effectively suppresses similarity to background regions, localizing objects like "nose", "candle", and the fence for "western setting" with high precision.
- **Binary Cross-Entropy (BCE):** The BCE variant maintains a broader semantic scope. While this results in higher initial background noise (e.g., the curtain in Fig. 2), it allows the model to capture multi-instance concepts better (e.g., "chatting locals" in Fig. 2).

The Effect of Granularity (K). For both variants, but particularly for BCE, increasing the size of the hierarchy K suppresses noisy features better. At lower K, BCE exhibits diffuse similarities comparable to the baselines. However, as K increases to 36, these diffuse regions become noticeably less. This effectively suppresses irrelevant background regions (ex. the "cowboys" and "cows" in Fig. 2) and sharpens object boundaries (ex. textured log in Fig. 3. It is additionally able to localize fine-grained semantics such as the fence with increasing precision as K incrases (ex. "western setting" in Fig. 2).

6. Implementation Details

Our approach introduces a randomly initialized Cross-Attention Transformer Block (8 heads, 512 hidden dim, MLP expansion ratio 4, Pre-Norm) to the standard CLIP model. The CLIP vision and text encoders are initialized from OpenAI's pre-trained weights. We fine-tune the model on the filtered ShareGPT4V-1.2M dataset. Table 8 summarizes key statistics of the data used during training and evaluation. Training is done using 4 NVIDIA A100 (80GB) GPUs. The optimizer settings, learning-rate schedule and data-augmentation pipeline used during fine-tuning are described in Table 7.

Parameter	**Value**
Epochs	10
Batch size	2,048
Optimizer	AdamW (β_1=0.9, β_2=0.98)
Initial LR	1×10^{-9}
LR	1×10^{-5}
LR Cross-Attn	1×10^{-3}
Final LR Cross-Attn	1×10^{-4}
LR schedule	cosine decay
Weight decay	0.01
Warm-up epochs	0.1
Image Augmentation	RandomResizedCrop (0.5–1.0) Normalize: μ=(0.481, 0.458, 0.408) σ=(0.269, 0.261, 0.276)

Table 7. Hyper-parameters for β-CLIP

6.1. Multi-Granular Data Setup

Caption Filtering: We use Detoxify [4] and FalconsAI [3] to filter out content with a toxicity score > 0.1 and > 0.5 respectively. These samples include approximately $2K$ image-text pairs containing zero-tolerance content such as child abuse and pornography.

Caption Preprocessing: We use regex pattern matching to identify and remove repeated substrings (e.g., character-

Dataset	Task	Images	Text / Query
ShareGPT4V-1.2M [2]	Training	1.2M	-
Flickr30k [7]	Coarse-Grained	31k	5 cap $\times$ img
MS-COCO14 (5k) [5]	Coarse-Grained	5k	5 cap $\times$ img
DCI [6]	Long-Text	8k	1 cap $\times$ img
Urban1k [8]	Long-Text	1k	1 cap $\times$ img
ShareGPT4V-1K [2]	Long-Text	1k	1 cap $\times$ img
FG-OVD (ex. Hard) [1]	Fine-grained	1.7k	2.3k positive cap + 23k neg cap

Table 8. Statistics of the training corpus and evaluation benchmarks.

level repetitions) and 'itertools.groupby' to eliminate consecutive word-level repetitions, which are common in generated long captions.

Hierarchical Parsing: To generate the multi-granular queries, we utilize the `spaCy` library with the en_core_web_sm model.

1. **Sentences:** The caption is split using standard sentence tokenization. During training, we sample N sentences (without replacement if enough sentences are available) to serve as coarse-grained queries.
2. **Phrases:** We extract fine-grained phrases using a custom extractor:
   - **Noun Chunks:** We extract base noun phrases and extend them to include spatial indicators (e.g., "...on the left") using custom matchers.
   - **Actions:** We identify action-oriented phrases by matching `VERB+ADP` patterns (e.g., "leaning against").
   - **Spatial Relations:** We parse spatial prepositional phrases anchored by a set of predefined `SPATIAL HEAD` tokens (including directions like *left, right, top, bottom* and positions like *center, middle, near*). We extract phrases matching the pattern `ADP [DET]? SPATIAL HEAD [of]?`, capturing relative positions such as to the left of" or in the center," alongside explicit adverbial relations like "next to."

We filter out phrases shorter than 3 characters or those consisting solely of stop words to ensure semantic meaningfulness.

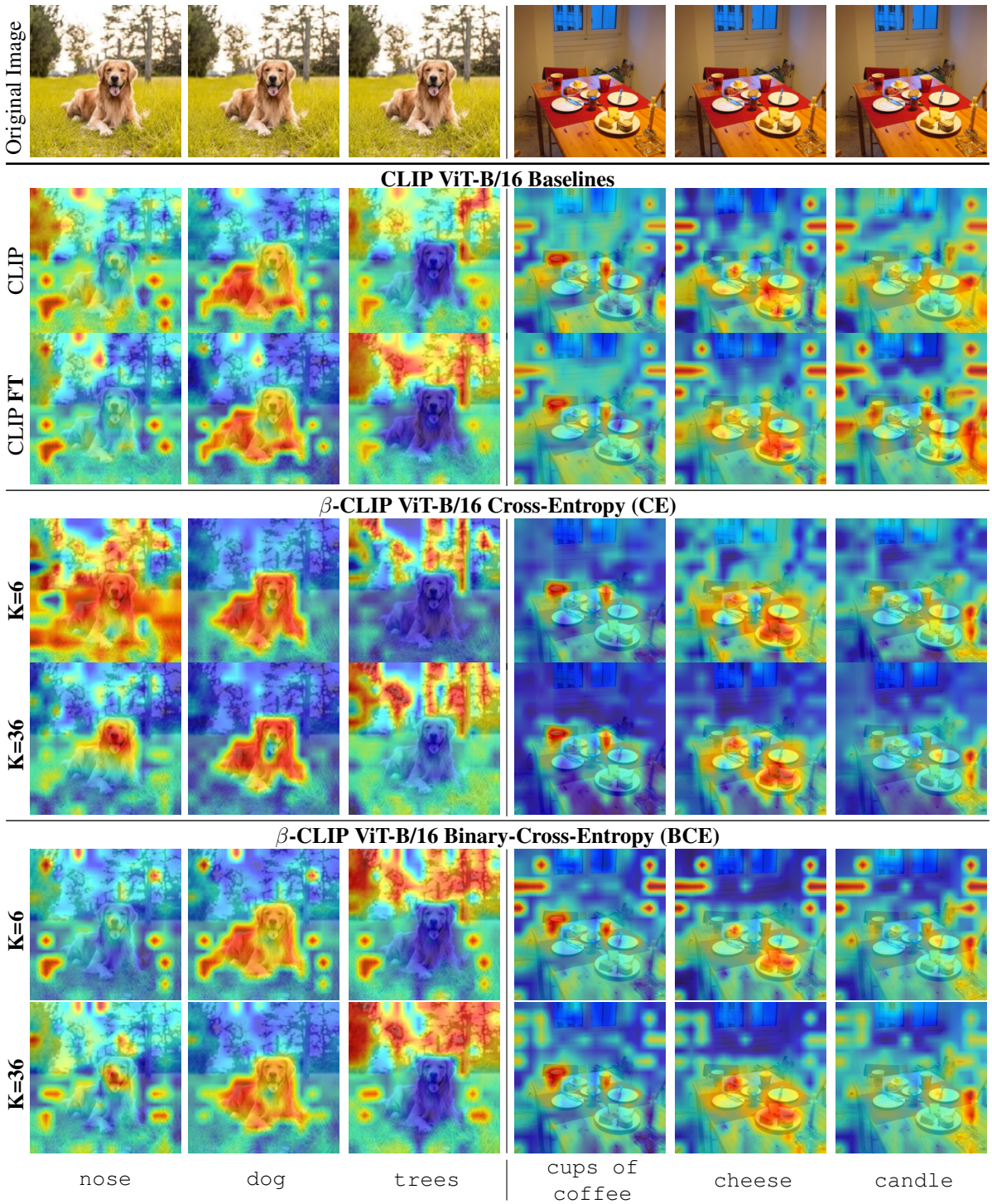

Figure 1. **Heatmaps of Patch-Text Logit Similarities for Short Phrases.** Rows 1-2: Original CLIP and CLIP FT exhibit diffuse similarity patterns, characterized by frequent secondary peaks on semantically unrelated regions such as the grass or background walls. Rows 3-6: β-CLIP successfully disentangles these fine-grained concepts. CE yields the most spatially sparse distributions, concentrating similarity primarily on the regions most relevant to the text. BCE shows higher similarity with background features, comparable to the baselines at lower granularity, K=6. However, as K increases, these irrelevant regions are more effectively suppressed, and the similarities become progressively more concentrated on the semantically relevant regions.

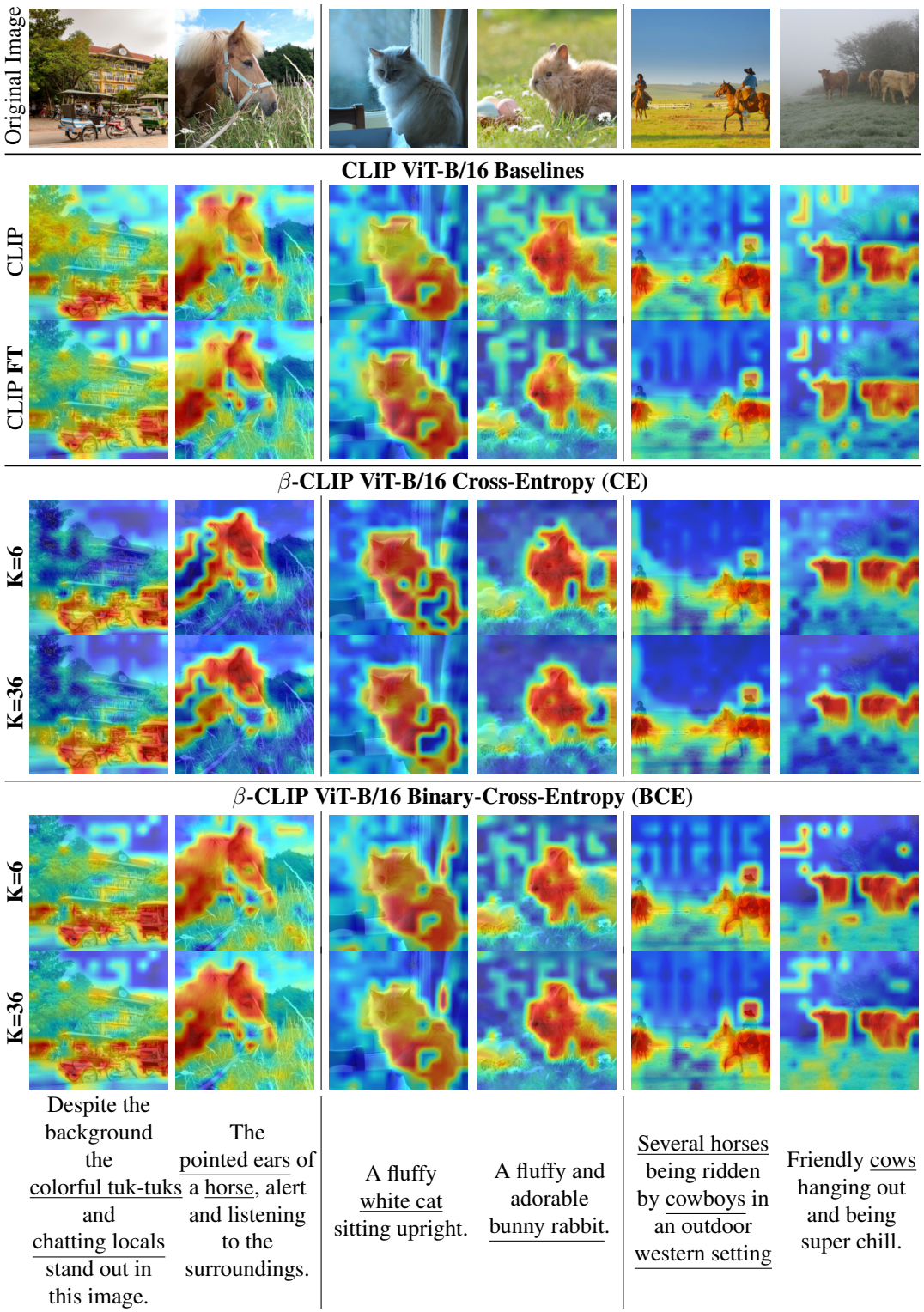

Figure 2. **Heatmaps of Patch-Text Logit Similarities for Long Captions.** Rows 1-2: Original CLIP and CLIP FT exhibit diffuse similarity patterns, assigning high similarity to broader background regions. Rows 3-6: β-CLIP improves alignment with complex semantics. CE yields the sharpest localization, effectively isolating specific details (ex., the "pointed ears") and increasingly so with higher K (ex. "western setting" at K=36). BCE captures a wider semantic scope, grounding multi-instance concepts like the "chatting locals," better than the rest. Increasing K suppresses similarity to irrelevant regions (most evident in the cowboys and cows scenes). Notably, all models exhibit sensitivity to salient unmentioned objects that contextually co-occur with the query, such as the eggs next to the bunny.

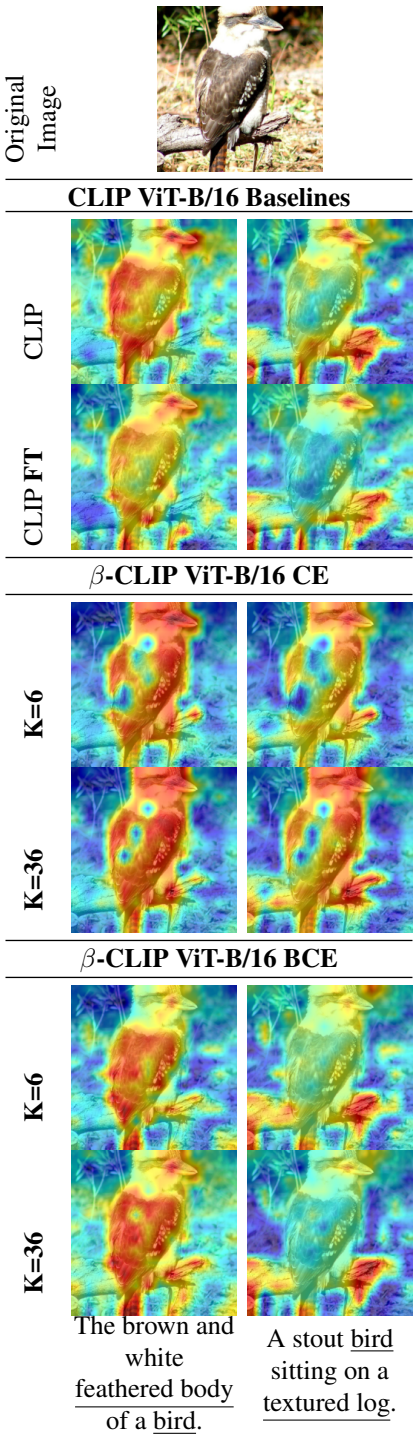

Figure 3. **Heatmaps of Patch-Text Logit Similarities: CE vs BCE.** Rows 1-2: Original CLIP and CLIP FT exhibit diffuse similarity patterns with a disproportionate focus on the beak, regardless of the query. Rows 3-6: β-CLIP comparisons reveal distinct localization behaviors. CE provides significantly higher coverage of the bird's body, and increasingly with higher K. BCE is unable to localize the entire body given the word "bird" in a different context; instead, it focuses strictly on salient features like the beak and tail. However, given a more detailed description, such as "feathered body," it is able to achieve better localization. Both methods effectively ground the distinct background element ("textured log"), with boundary precision improving at higher K.

References

[1] Lorenzo Bianchi, Fabio Carrara, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. *arXiv*, 2023. 4

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 4

[3] Falconsai. nsfw_image_detection. https://huggingface.co/Falconsai/nsfw_image_detection. Hugging Face. 3

[4] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020. 3

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 4

[6] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions, 2024. 4

[7] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 4

[8] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip, 2024. 4