Table 1: Results on Flickr30k in the *Low-Budget* (i.e., $\varepsilon = 4/255$) setting. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. The full knowledge scenario (i.e., $q_t = q_u$) is denoted by "-". For each transformation (Tr), the first column reports results with random selection ($SelRand$), while the second uses similarity-based selection ($Sel\phi$). The best results for $SelRand$ are highlighted, and those for $Sel\phi$ are underlined.

VLP	n	ASR@1								ASR@5							
		TrSyn		**TrLLM**		**TrIC-1**		**TrIC-5**		***TrSyn***		**TrLLM**		**TrIC-1**		**TrIC-5**	
CLIP_{ViT}	-	100.0								100.0							
CLIP_{ViT}	0	89.2								95.8							
CLIP_{ViT}	1	86.9	88.9	89.7	89.9	89.5	91.1	87.1	90.9	94.8	96.0	96.4	95.8	96.4	96.4	95.6	96.9
CLIP_{ViT}	5	85.2	89.0	90.5	90.0	89.8	90.6	87.2	92.9	94.4	96.0	96.5	96.3	97.1	96.7	97.9	97.7
CLIP_{ViT}	10	84.5	89.3	90.5	90.4	88.5	90.1	88.5	93.9	93.5	95.7	96.7	96.6	96.3	96.7	98.1	98.2
CLIP_{ViT}	15	85.7	89.1	**91.0**	91.0	89.1	88.8	89.2	**94.0**	94.6	96.0	96.8	96.8	96.4	96.4	**98.2**	**98.4**
CLIP_{CNN}	-	99.8								99.8							
CLIP_{CNN}	0	89.4								96.1							
CLIP_{CNN}	1	90.0	89.8	90.6	89.7	92.4	92.9	91.7	92.2	97.0	96.2	96.5	96.4	97.8	97.7	98.3	96.9
CLIP_{CNN}	5	91.0	90.7	91.6	91.5	92.6	94.6	92.2	94.9	97.7	96.5	96.9	96.9	98.3	99.0	**99.5**	98.6
CLIP_{CNN}	10	90.6	90.7	91.5	91.1	**93.0**	93.7	92.0	**95.8**	97.6	96.5	96.9	97.0	98.6	98.9	99.3	99.2
CLIP_{CNN}	15	91.0	90.6	91.3	91.5	92.2	92.4	91.6	**95.8**	97.5	96.5	97.2	97.1	98.5	98.7	99.1	**99.5**
BLIP-2	-	85.2								98.8							
BLIP-2	0	31.0								80.7							
BLIP-2	1	23.6	31.7	31.0	30.8	22.1	**38.6**	13.8	37.0	76.1	80.0	80.5	80.3	78.0	84.8	71.4	83.0
BLIP-2	5	20.6	31.4	31.8	32.3	21.0	35.9	8.5	34.2	74.0	80.6	80.8	80.7	80.9	**86.2**	65.1	85.0
BLIP-2	10	20.8	31.3	31.8	31.7	23.6	30.3	8.0	32.2	73.7	80.1	80.6	80.7	81.9	85.0	67.0	85.5
BLIP-2	15	19.9	31.8	**31.9**	32.0	22.8	24.3	8.1	29.1	73.1	81.4	80.6	80.7	**82.5**	83.0	68.4	85.2
$\text{BLIP-2}_{\text{ITM}}$	-	38.4								93.6							
$\text{BLIP-2}_{\text{ITM}}$	0	**4.7**								71.8							
$\text{BLIP-2}_{\text{ITM}}$	1	3.3	4.6	4.3	4.5	2.5	4.7	2.0	4.6	67.5	71.4	72.5	71.4	71.2	76.0	65.4	74.9
$\text{BLIP-2}_{\text{ITM}}$	5	3.0	3.9	4.1	4.6	3.1	3.8	1.2	4.1	64.9	72.1	72.4	72.2	72.1	**76.2**	62.4	75.7
$\text{BLIP-2}_{\text{ITM}}$	10	3.0	4.3	4.6	**4.8**	2.4	3.2	0.8	3.1	64.1	72.2	72.3	72.2	73.6	75.3	63.1	**76.2**
$\text{BLIP-2}_{\text{ITM}}$	15	3.0	4.0	4.6	4.3	2.9	3.0	1.1	2.8	64.0	72.4	72.2	72.9	**74.6**	74.4	63.3	75.1

Table 2: Results on MSCOCO in the *Low-Budget* (i.e., $\varepsilon = 4/255$) setting. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. The full knowledge scenario (i.e., $q_t = q_u$) is denoted by "-". For each transformation (*Tr*), the first column reports results with random selection (*SelRand*), while the second uses similarity-based selection (*Selϕ*). The best results for *SelRand* are highlighted, and those for *Selϕ* are underlined.

VLP	n	ASR@1								ASR@5							
		TrSyn		**TrLLM**		**TrIC-1**		**TrIC-5**		***TrSyn***		**TrLLM**		**TrIC-1**		**TrIC-5**	
CLIP_{ViT}	-	100.0								100.0							
CLIP_{ViT}	0	88.8								94.5							
CLIP_{ViT}	1	87.3	88.9	90.1	89.9	89.2	91.4	89.0	91.5	93.7	94.6	94.6	94.9	94.3	95.8	94.7	95.6
CLIP_{ViT}	5	85.9	89.6	90.7	90.4	89.4	91.8	90.7	93.4	93.0	94.6	95.2	94.8	94.9	95.8	96.3	97.1
CLIP_{ViT}	10	85.3	89.9	91.5	90.9	89.7	91.4	91.7	93.9	93.3	94.6	95.6	95.6	95.0	95.7	96.9	**97.4**
CLIP_{ViT}	15	86.5	89.9	91.5	91.4	89.7	89.8	**91.8**	**94.0**	93.9	94.5	95.8	95.8	95.1	95.0	**97.6**	**97.4**
CLIP_{CNN}	-	100.0								100.0							
CLIP_{CNN}	0	89.3								94.8							
CLIP_{CNN}	1	87.4	89.0	90.3	89.9	91.2	91.4	91.0	92.5	94.8	95.0	95.5	95.4	96.5	96.0	96.5	97.1
CLIP_{CNN}	5	88.4	90.0	91.1	90.7	91.2	92.0	91.9	94.0	95.4	95.9	96.6	96.3	95.7	96.0	97.1	97.7
CLIP_{CNN}	10	88.3	90.5	91.1	91.3	91.3	92.6	91.5	**94.5**	95.4	96.1	96.3	96.5	96.2	96.3	97.4	97.7
CLIP_{CNN}	15	88.6	90.2	91.1	91.7	91.2	92.1	**92.9**	94.1	95.7	95.7	96.2	96.7	95.9	96.4	**98.3**	**97.9**
BLIP-2	-	67.2								85.4							
BLIP-2	0	19.9								48.0							
BLIP-2	1	14.5	19.9	20.7	20.6	14.6	23.9	11.0	23.5	41.9	47.5	47.5	47.7	40.6	50.8	40.2	50.9
BLIP-2	5	12.4	21.2	**22.0**	20.6	14.8	23.3	10.8	**25.0**	40.9	47.9	**49.0**	48.5	45.8	53.8	40.9	54.5
BLIP-2	10	12.6	20.1	21.9	21.2	14.9	21.1	11.7	24.0	38.9	47.6	48.8	49.1	45.8	51.5	41.9	55.4
BLIP-2	15	13.0	20.8	21.6	22.0	16.0	15.4	11.6	24.4	39.4	47.5	48.9	49.2	48.2	48.2	44.1	**55.5**
$\text{BLIP-2}_{\text{ITM}}$	-	26.4								62.2							
$\text{BLIP-2}_{\text{ITM}}$	0	2.2								37.4							
$\text{BLIP-2}_{\text{ITM}}$	1	2.1	2.4	2.5	2.5	1.4	2.4	1.5	2.4	34.7	36.5	37.4	36.7	34.5	38.4	34.9	37.7
$\text{BLIP-2}_{\text{ITM}}$	5	1.8	2.4	**2.8**	2.3	1.5	2.2	1.3	**2.6**	34.0	37.2	38.0	38.4	38.1	40.3	37.6	40.2
$\text{BLIP-2}_{\text{ITM}}$	10	1.8	2.4	2.1	2.6	1.3	1.9	1.0	2.1	33.9	36.9	38.7	38.2	38.4	41.4	38.3	41.6
$\text{BLIP-2}_{\text{ITM}}$	15	2.2	2.3	2.6	**2.6**	1.5	1.6	0.9	2.2	33.7	37.5	38.1	39.1	**40.7**	41.0	39.7	**41.9**

Table 3: Results on Flickr30k in the *High-Budget* (i.e., $\varepsilon = 8/255$) setting. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. The full knowledge scenario (i.e., $q_t = q_u$) is denoted by "-". For each transformation (*Tr*), the first column reports results with random selection (*SelRand*), while the second uses similarity-based selection (*Selϕ*). The best results for *SelRand* are highlighted, and those for *Selϕ* are underlined.

VLP	n	ASR@1								ASR@5							
		TrSyn		**TrLLM**		**TrIC-1**		**TrIC-5**		***TrSyn***		**TrLLM**		**TrIC-1**		**TrIC-5**	
CLIP_{ViT}	-	100.0								100.0							
CLIP_{ViT}	0	92.6								97.0							
CLIP_{ViT}	1	91.2	93.0	93.4	92.8	93.3	94.2	92.5	94.4	96.1	96.9	97.3	97.1	97.5	97.6	97.0	97.5
CLIP_{ViT}	5	90.5	92.9	93.6	94.1	94.2	94.7	93.4	96.3	95.3	96.7	97.5	97.3	98.2	98.2	99.0	98.5
CLIP_{ViT}	10	89.6	92.3	93.9	93.9	93.8	94.6	94.9	95.9	95.8	96.7	97.7	97.9	98.3	98.4	99.1	98.6
CLIP_{ViT}	15	91.0	92.9	94.2	94.2	94.1	93.8	**95.6**	**96.8**	96.5	97.3	97.5	97.5	97.9	97.9	**99.2**	**99.2**
CLIP_{CNN}	-	99.8								99.8							
CLIP_{CNN}	0	92.4								96.6							
CLIP_{CNN}	1	93.0	92.6	93.5	93.3	95.4	95.4	95.5	94.6	97.8	97.1	97.1	97.2	98.6	98.5	98.9	98.0
CLIP_{CNN}	5	94.8	93.1	94.5	93.5	96.4	97.1	96.4	96.9	98.3	97.4	97.7	97.6	99.3	99.5	99.8	99.4
CLIP_{CNN}	10	94.3	93.3	93.9	94.1	**96.5**	96.9	96.1	98.0	98.6	97.8	97.6	97.8	99.3	99.4	99.8	99.5
CLIP_{CNN}	15	94.5	93.4	94.1	94.3	**96.5**	96.4	95.9	**98.1**	98.5	97.2	97.8	97.8	99.4	99.3	**100.0**	**99.7**
BLIP-2	-	99.0								100.0							
BLIP-2	0	49.0								88.4							
BLIP-2	1	38.3	48.9	48.7	50.4	37.4	**61.0**	25.5	55.3	84.3	87.8	87.9	88.6	87.2	92.4	83.2	90.8
BLIP-2	5	34.0	48.2	**50.2**	50.4	38.1	58.8	16.6	57.3	82.7	88.0	88.9	88.7	90.6	**93.8**	80.7	93.4
BLIP-2	10	32.9	49.3	49.80	50.1	42.7	53.0	16.5	52.7	82.6	88.6	88.7	88.7	91.8	93.6	81.2	93.7
BLIP-2	15	32.9	49.4	**50.2**	49.6	42.9	42.3	16.9	50.5	83.4	87.9	88.9	88.6	**92.1**	92.0	84.7	92.9
$\text{BLIP-2}_{\text{ITM}}$	-	53.8								99.0							
$\text{BLIP-2}_{\text{ITM}}$	0	6.9								81.7							
$\text{BLIP-2}_{\text{ITM}}$	1	5.4	7.2	7.3	7.3	4.4	**8.5**	3.4	8.2	78.0	81.3	81.8	82.1	82.3	86.5	75.9	84.2
$\text{BLIP-2}_{\text{ITM}}$	5	4.6	7.7	6.9	7.5	4.2	7.2	1.6	6.0	74.7	82.1	82.3	82.0	81.7	**87.2**	74.0	85.1
$\text{BLIP-2}_{\text{ITM}}$	10	4.9	7.3	**7.5**	6.7	4.0	5.1	1.0	4.9	74.1	82.1	82.6	82.0	**84.3**	84.8	74.7	84.8
$\text{BLIP-2}_{\text{ITM}}$	15	4.7	7.4	7.0	7.8	3.8	4.1	1.1	4.3	74.4	81.4	82.1	82.4	83.9	83.2	75.7	83.6

Table 4: Results on MSCOCO in the *High-Budget* (i.e., $\varepsilon = 8/255$) setting. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. The full knowledge scenario (i.e., $q_t = q_u$) is denoted by "-". For each transformation (Tr), the first column reports results with random selection (*SelRand*), while the second uses similarity-based selection ($Sel\phi$). The best results for *SelRand* are highlighted, and those for $Sel\phi$ are underlined.

VLP	n	ASR@1								ASR@5							
		TrSyn		**TrLLM**		**TrIC-1**		**TrIC-5**		***TrSyn***		**TrLLM**		**TrIC-1**		**TrIC-5**	
CLIP_{ViT}	-	100.0								100.0							
CLIP_{ViT}	0	93.4								96.8							
CLIP_{ViT}	1	91.9	93.6	94.1	94.3	94.4	95.4	94.0	95.5	95.8	97.0	97.4	96.9	97.4	97.9	97.1	98.1
CLIP_{ViT}	5	91.4	94.1	94.4	94.7	94.8	95.3	95.1	96.3	96.0	96.9	97.3	97.7	98.1	97.6	98.1	98.0
CLIP_{ViT}	10	92.3	94.2	94.6	94.9	94.3	95.0	95.7	**96.8**	96.7	97.0	97.6	97.8	97.4	97.6	98.8	**98.4**
CLIP_{ViT}	15	92.4	93.9	95.2	95.2	94.5	94.5	**96.9**	96.7	96.3	97.3	97.7	97.7	97.5	97.6	**99.2**	**98.4**
CLIP_{CNN}	-	100.0								100.0							
CLIP_{CNN}	0	93.1								96.5							
CLIP_{CNN}	1	93.1	92.7	94.2	93.3	94.5	94.6	94.1	95.1	97.0	97.2	97.4	96.8	97.9	97.6	97.8	97.9
CLIP_{CNN}	5	93.8	93.6	93.9	94.0	94.7	95.0	96.6	95.8	97.4	97.2	97.2	97.1	97.5	97.6	99.0	98.6
CLIP_{CNN}	10	93.2	93.9	94.7	94.1	95.0	95.2	96.4	96.8	97.3	97.7	97.5	97.6	97.7	97.6	98.8	98.8
CLIP_{CNN}	15	93.6	94.4	94.3	94.1	95.3	95.1	**96.7**	**96.9**	97.3	97.7	97.7	97.6	97.8	97.8	**99.1**	**99.0**
BLIP-2	-	90.8								98.2							
BLIP-2	0	33.7								62.8							
BLIP-2	1	25.7	33.8	35.4	34.8	26.8	40.5	21.6	38.8	56.9	63.0	62.3	63.2	60.1	66.4	56.2	66.1
BLIP-2	5	23.7	35.4	35.6	35.2	28.4	40.8	22.3	42.7	56.0	63.6	64.0	64.0	61.0	70.2	60.2	70.9
BLIP-2	10	23.1	35.0	**35.8**	35.6	30.4	39.9	21.5	**43.0**	54.9	63.1	64.3	63.8	64.0	69.1	61.6	72.3
BLIP-2	15	22.8	35.0	35.6	35.4	32.3	32.6	24.8	42.9	54.2	64.1	64.4	64.6	**66.0**	65.8	64.4	**72.9**
$\text{BLIP-2}_{\text{ITM}}$	-	40.4								81.0							
$\text{BLIP-2}_{\text{ITM}}$	0	4.1								47.7							
$\text{BLIP-2}_{\text{ITM}}$	1	2.8	4.0	3.9	**4.6**	2.3	4.2	2.0	4.3	45.0	47.2	48.1	47.5	46.7	50.9	46.1	50.2
$\text{BLIP-2}_{\text{ITM}}$	5	2.8	4.5	4.2	4.1	2.6	3.4	1.6	4.1	44.7	48.0	48.7	47.7	49.4	52.8	50.1	52.0
$\text{BLIP-2}_{\text{ITM}}$	10	2.4	3.8	3.9	4.0	2.3	3.1	1.5	3.7	42.6	47.4	50.1	48.5	51.1	52.5	49.4	**54.7**
$\text{BLIP-2}_{\text{ITM}}$	15	2.5	3.7	**4.4**	4.4	2.0	2.2	1.5	2.9	44.9	48.2	48.9	48.8	**51.2**	51.9	50.1	54.2