CLID: Controlled-Length Image Descriptions with Limited Data Supplementary Material

1. Implementation Details

Self-generating Data

1. Existing scene graphs are taken from the *Visual Genome dataset* [34].
2. For scene-graph generation (SGG) we use the re-implementation of [70] by [22].
3. For the descriptive caption task on MS-COCO, we generate 10 captions per graph that appear in VG. For other images, we use automatic scene graph generation methods. Since the generator's results are not accurate, we generate only 2 captions for each of these images. For Flickr30k, we generate 5 captions per image, all from automatically generated scene graphs.
4. For the descriptive paragraph task, we generate 5 paragraphs per image (all appear in VG).

Training

We train the models with the architecture of LaBERT on a single Nvidia Tesla V100 32GB, with the following configuration:

1. Each input image is represented by 100 embedding vectors, which correspond to 100 object areas extracted by Faster RCNN [Ren, 2015].
2. The model contains 12 Transformer layers, 12 attention heads and a hidden size of 768.
3. Training for $100,000$ steps ($30,000$ for paragraphs).
4. The loss function is cross-entropy over randomly masked positions.
5. Batch size: 64
6. AdamW optimizer:
   - learning rate: 5e-5
   - weight decay: 1e-2
   - betas: (0.9, 0.999)
   - gradient clip: 1.0
   - 1000 warmup steps with cosine scheduler
7. For inference:
   - We use iterative refinement (rather than beam search), as originally proposed by Deng et al. [14]. Furthermore, we perform nucleus sampling with $p = 0.95$.
   - The steps for the iterative refinement are set for levels 1-7 with (10, 15, 20, 25, 30, 35, 40) respectively. For paragraphs it is for levels 1-13 with (10, 10, 15, 15, 25, 25, 40, 40, 50, 50, 60, 60, 65).
   - We do not employ EOS decay, in order to eliminate factors that are not training-based.

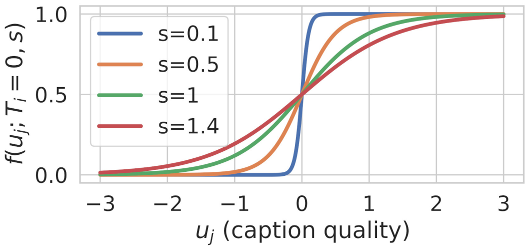


Figure 1. **Smooth step functions centered at $T_i = 0$ (Eq. 3).** This function maps quality scores to weights, which are used for sampling the data points. The smoother the function (larger s) is, the more likely it is for low-quality captions to be sampled.

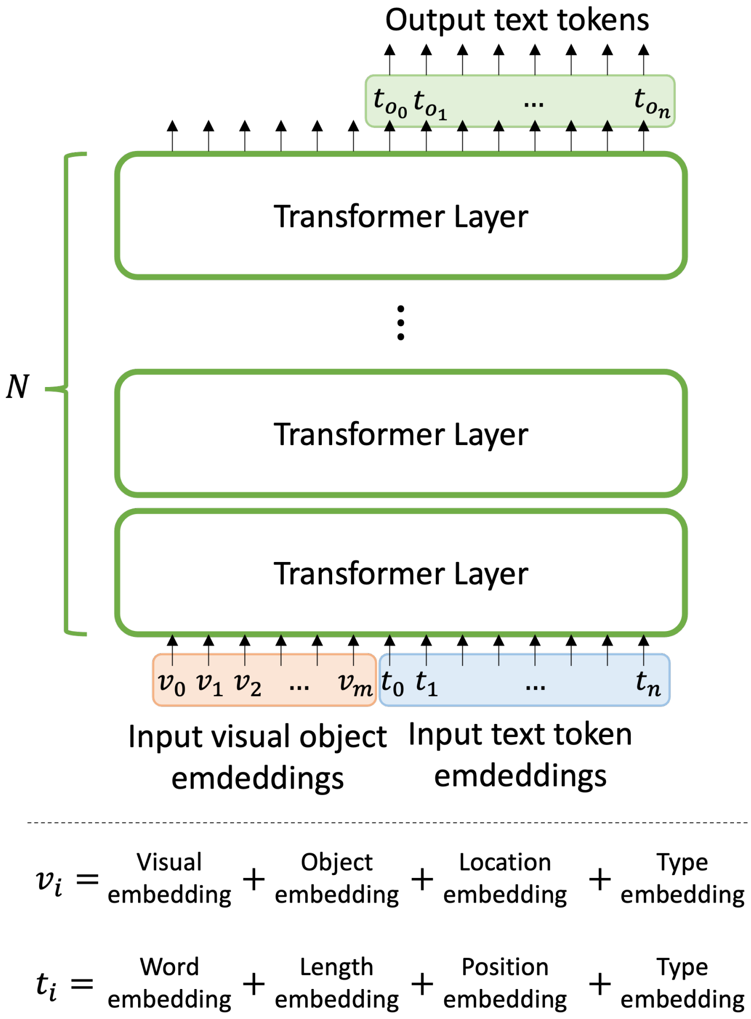

Output text tokens
t_{o_0} t_{o_1} … t_{o_n}
Transformer Layer
Transformer Layer
Transformer Layer
N
v_0 v_1 v_2 … v_m
t_0 t_1 … t_n
Input visual object emdeddings
Input text token emdeddings
v_i = Visual embedding + Object embedding + Location embedding + Type embedding
t_i = Word embedding + Length embedding + Position embedding + Type embedding

Figure 2. **Length-aware BERT architecture.** The length-aware architecture is composed of N Transformer blocks. It inputs both image regions and masked text tokens, with an additional length embedding. In our experiments, $N = 12$ and $m = 100$. n depends on the maximal description length in the batch.

2. Pseudocodes

Algorithm 1 Scene graph exploration for caption generation

```
Input scene graph $G$, object saliency vector $p$
Output image caption
while explored saliency $< T_{sal}$ do
    Randomly choose an unvisited vertex $v$ according to the probabilities of $p$
    DFS-like($G$, $p$, $v$, visited)
    Update the explored saliency of the path & the visited vertices
end while
For each explored vertex, sample up to $n_a$ attributes
Generate the captions according to the exploration; stop randomly to vary length

procedure DFS-LIKE($G$, $p$, $v$)
    Choose at most $k$ children $v'_{1:k}$, according to their probabilities in $p$
    for each chosen child $v'$ do
        if $v'$ was not visited then
            DFS-like($G$, $p$, $v'$)
        end if
    end for
    Return the visited vertices in order of their visit
end procedure
```

Algorithm 2 Training Procedure

```
Input $\tilde{D}$ – extended dataset, $D$ – trusted dataset, $s$ – smoothness value
Output A captioning model
$M_{\tilde{\theta}} \leftarrow$ train a captioning model on $\tilde{D}$
$M_{\theta} \leftarrow$ tune $M_{\tilde{\theta}}$ on $D$
Compute the quality $u_j$ of each data point $j$ using $M_{\tilde{\theta}}$, $M_{\theta}$ (Eq.1)
Initialize the captioning model
$i \leftarrow 0$
for $q$ in $[0, ..., 100]$ do
    $T_i \leftarrow$ quality value of the $q$ percentile of the self-generated data points
    Compute the weight for each data point $j$ by $f(u_j; T_i, s)$ (Eq.3)
    Randomly sample $100 - q\%$ of the self-generated data based on the weights
    Train the model on the trusted and sampled data for $\eta$ steps
    $i \leftarrow i + 1$
end for
Continue training the model on the trusted data until max-steps
```

3. Paragraph Generation - Details & Results

Data Histograms

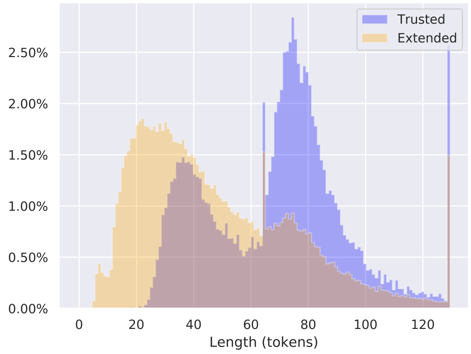


Figure 3. **Length of paragraph datasets.** The lengths of paragraphs in the trusted dataset ([33]; blue) compared to the extended dataset (orange). The spike at 129 is due to clipping. Furthermore, the extended dataset is bigger by a factor of 6. (Overlaps cause the third color.)

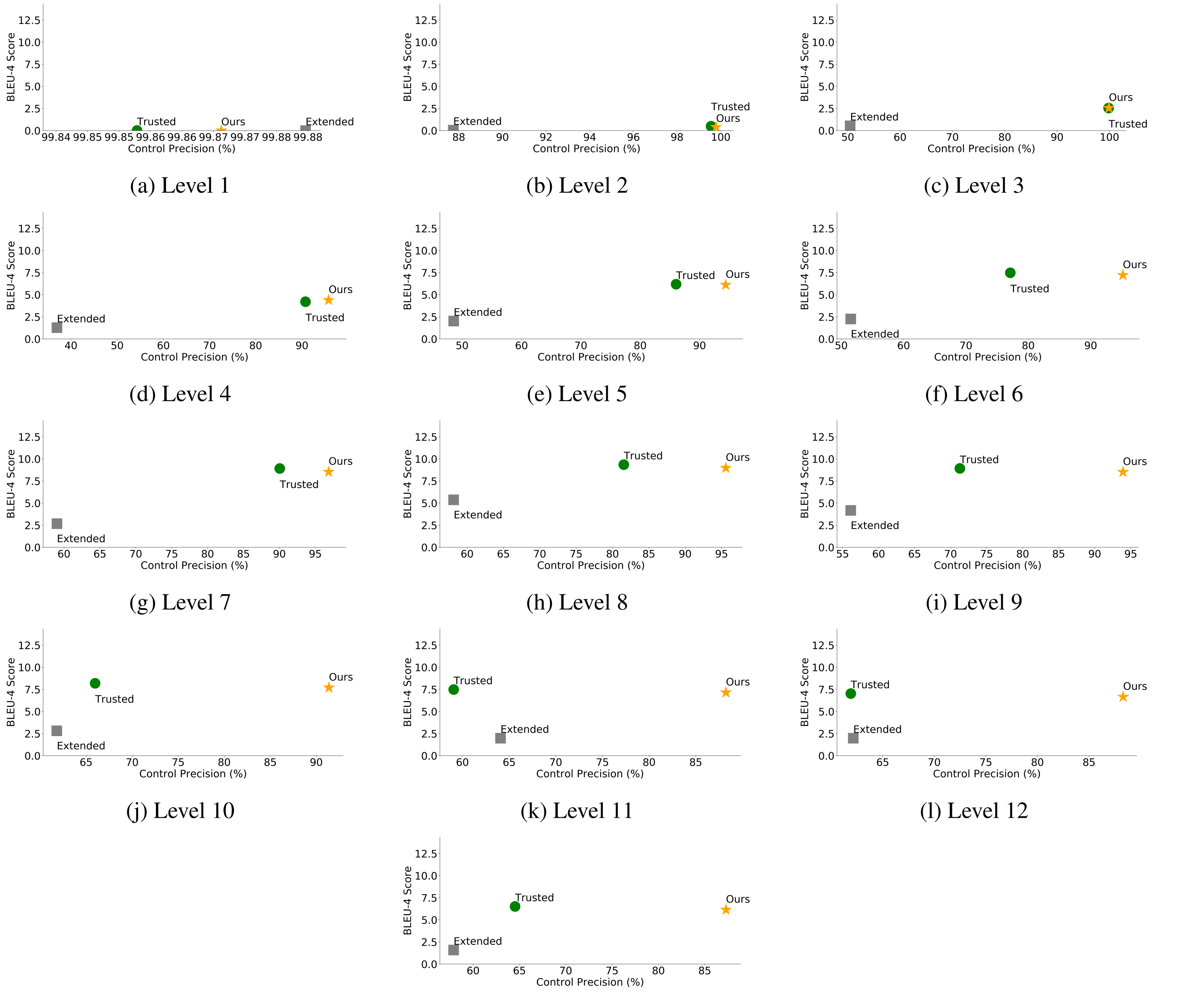
Control Precision

Level Length (tokens)	1 1-9	2 10-19	3 20-29	4 30-39	5 40-49	6 50-59	7 60-69
% in trusted dataset	0.006%	0.03%	1.84%	13.38%	11.51%	7.03%	9.56%
% in extended dataset	2.38%	9.5%	14.37%	15.29%	12.83%	10.07%	8.63%
LaBERT [14]	99.85%	99.55%	99.83%	90.77%	86.04%	77.11%	90.04%
CLID (ours)	**99.86%**	**99.78%**	**99.87%**	**95.77%**	**94.40%**	**95.17%**	**96.85%**

Level Length (tokens)	8 70-79	9 80-89	10 90-99	11 100-109	12 110-119	13 120-129	Average
% in trusted dataset	25.01%	18.1%	8.18%	3.77%	0.2%	1.32%	
% in extended dataset	9.63%	7.14%	4.29%	2.8%	1.75%	1.25%	
LaBERT [14]	81.55%	71.28%	65.97%	59.05%	61.90%	64.52%	80.57%
CLID (ours)	**95.58%**	**93.92%**	**91.37%**	**88.22%**	**88.30%**	**87.29%**	**94.34%**

Table 1. **Paragraph control precision**. For each level, the second row shows the range of tokens for this level. The next two rows show the percentage of paragraphs of each level in the training datasets. The two bottom rows compare the precision results of [14], trained on the trusted dataset of paragraphs, to those of our method. Our results outperform [14]'s both on average and for all of the levels.

Control vs. Quality (BLEU-4)



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

(h) Level 8

(i) Level 9

(j) Level 10

(k) Level 11

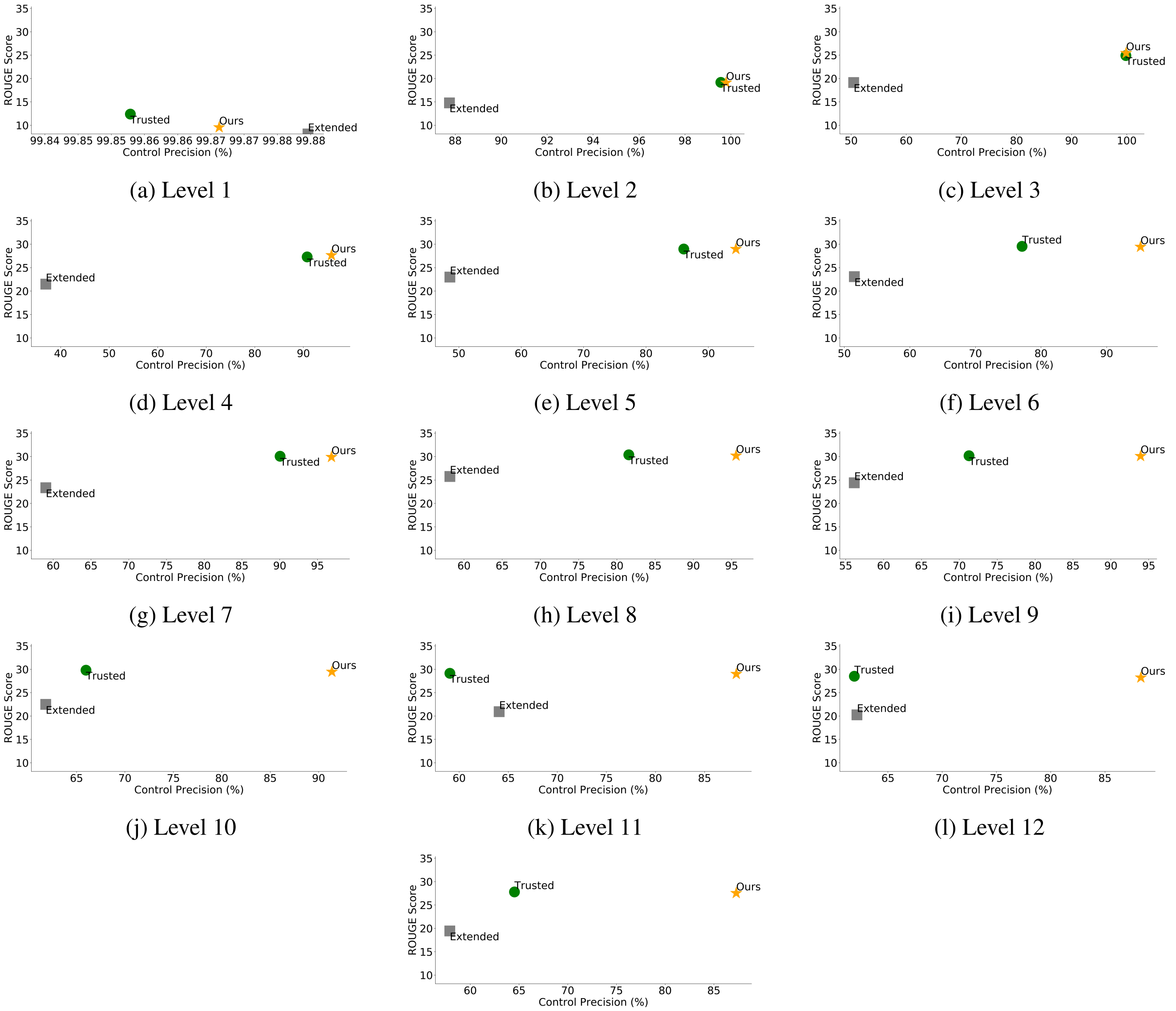
(l) Level 12

(m) Level 13

Figure 4. **Performance comparison to baselines (paragraphs).** In terms of the BLEU-4 quality measure (vertical axis), our results (orange star) are similar to [14]'s (green circle), which is trained on the trusted dataset. The quality of other solutions (gray), trained only on the extended dataset, is dramatically degraded. While comparable to [14] quality-wise, our model highly improves the control precision (horizontal axis) at all levels. In both measures, higher is better. As BLEU is an n-gram-based metric, very short paragraphs (Levels 1-2) are expected to have low scores for all methods.

Similar results with other metrics (ROUGE, METEOR & CIDEr*) can be seen in Fig. 5- 7.

* Importantly, the dataset of [33] contains only a single description per image. Hence, agreement-based metrics like CIDEr might not be suitable for this setup (results are reported regardless).



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

(h) Level 8

(i) Level 9

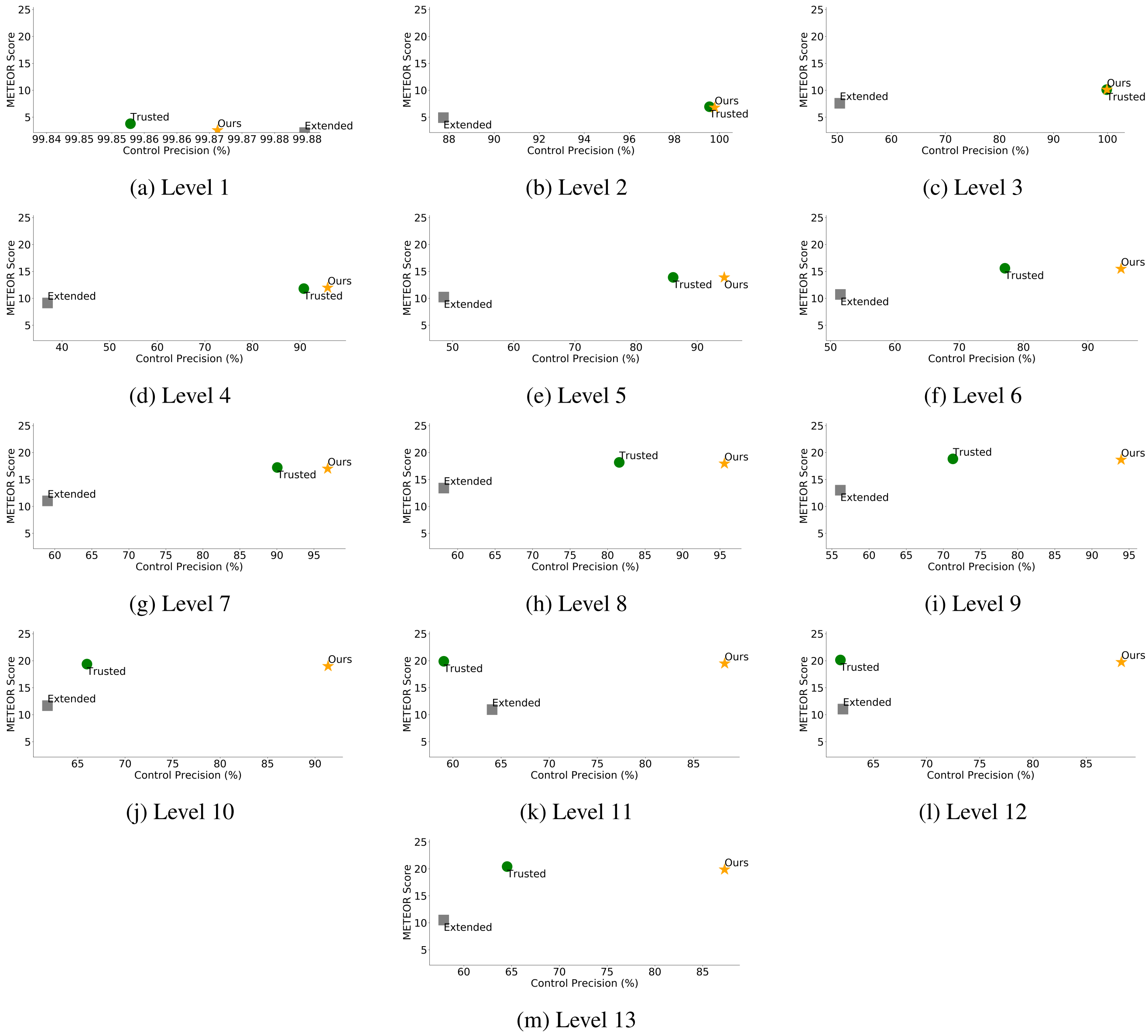
(j) Level 10

(k) Level 11

(l) Level 12

(m) Level 13

Figure 5. **Performance comparison to baselines (paragraphs).** Control precision (horizontal axis) vs. ROUGE score (vertical axis).



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

(h) Level 8

(i) Level 9

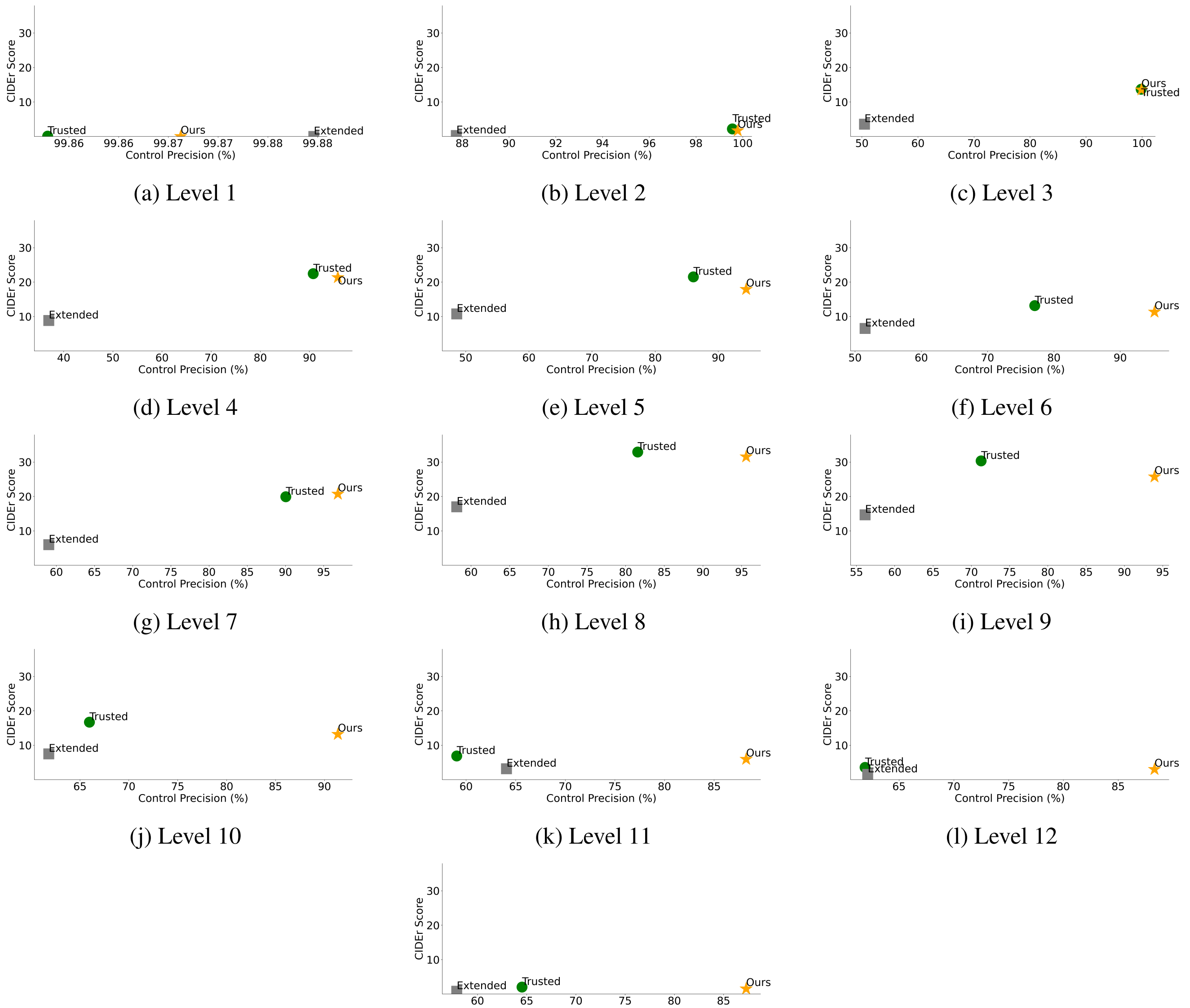
(j) Level 10

(k) Level 11

(l) Level 12

(m) Level 13

Figure 6. **Performance comparison to baselines (paragraphs).** Control precision (horizontal axis) vs. METEOR score (vertical axis).



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

(h) Level 8

(i) Level 9

(j) Level 10

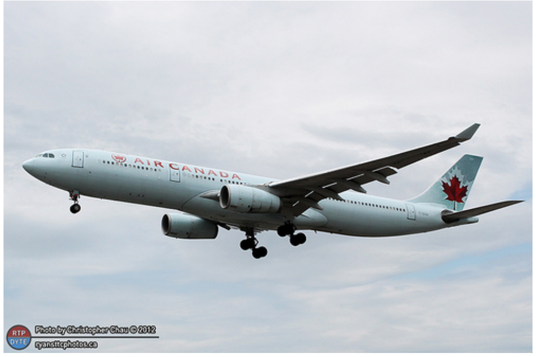
(k) Level 11

(l) Level 12

(m) Level 13

Figure 7. **Performance comparison to baselines (paragraphs).** Control precision (horizontal axis) vs. CIDEr score (vertical axis). Recall that CIDEr is a consensus-based metric which requires a set of ground-truth descriptions. However, in the dataset of [33] there is only one ground-truth description per image; hence, the metric's applicability within this specific setup might be limited.

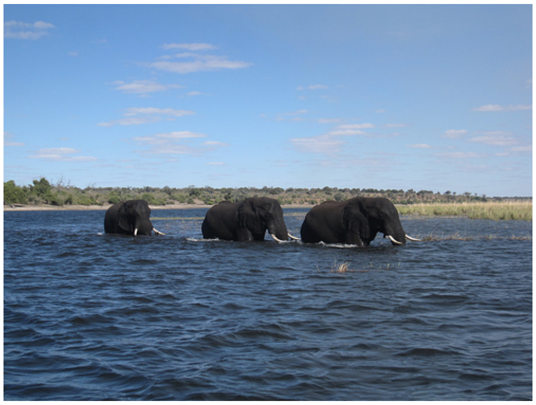
Paragraphs: Qualitative Results



GT: A plane is flying in the sky. The plane is very white in color. There is a logo on the tail. The logo is in the shape of a leaf.

Generated Paragraphs:

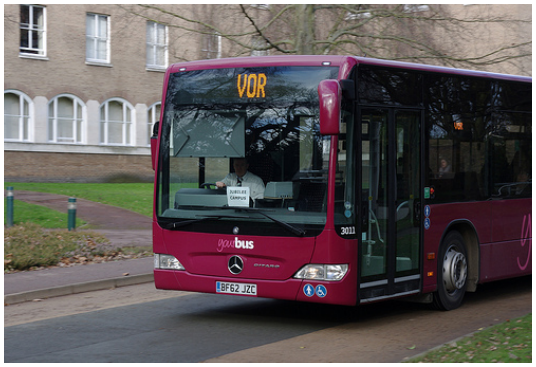
LaBERT [14]	Level 1	There is a large airplane in the sky.
	Level 4	The sky is gray and overcast. The sky is gray and white. There plane has a plane with red leaf on the tail the plane is white and has a on it.
CLID (ours)	Level 1	A plane is in the sky.
	Level 4	A plane is flying in the sky. The plane is white in color. The tail is red and white. There is a leaf on the side of it.
	(a) Length	(b) Generated Paragraph



GT: Three elephants are in the water. They are all walking in the same direction. They all have very long tusks showing that are not in the water.

Generated Paragraphs:

LaBERT [14]	Level 1	There are six elephants in the blue sky.
	Level 3	Four elephants are standing in the water. There are small white tusks in the water. The sky is blue.
CLID (ours)	Level 1	Three elephants in the water.
	Level 3	Three elephants are walking in the water. The elephants have white tusks. There are bushes. The sky is blue.
	(a) Length	(b) Generated Paragraph

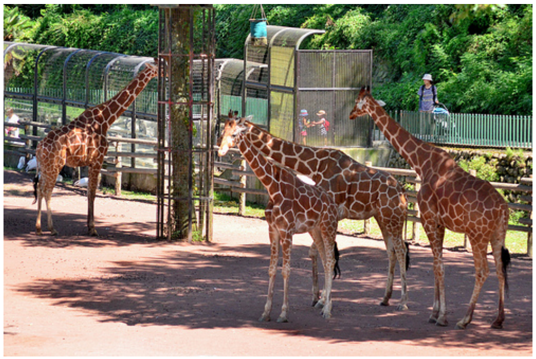


GT: A large red bus is on the street and there is a brick building behind the bus. The bus has large glass doors and the person driving the bus is wearing a white shirt. There is a tree behind the bus.

Generated Paragraphs:

LaBERT [14]	Level 2	A white and red bus is on the street. There is a building behind the bus.
	Level 4	There is a large red bus on the street. There is a white number on the side of the bus. There is a brick building behind the bus.
CLID (ours)	Level 2	A red bus is on the street. There is a building behind the bus.
	Level 4	There is a large red bus on the street. There is a man in a white shirt on the bus. There is a brick building behind the bus.

(a) Length (b) Generated Paragraph

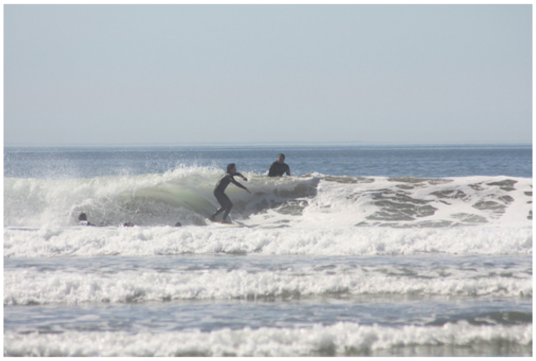


GT: Four giraffes stand in a giraffe enclosure. Each of them have broad brown spots. One giraffe is standing by a tree. The tree is surrounded by a metal cage. Three of the giraffes are standing in one shadow and the other one in another shadow. There is a tall fence in front of the giraffes.

Generated Paragraphs:

LaBERT [14]	Level 7	There are three giraffes standing around a fence. The fenceraffes are brown and tall. The fence is made of dirt. There are trees on the other side of the fence. There are many people around the giraffes. There are a few trees behind the animals.
	Level 8	There are three giraffes standing around a fenced the area. The giraffes are tall and brown and are standing in the middle of the fence. There are green trees on the inside of the fence. There are people walking on the outside of the fence. There are green trees on the background behind the giraffes.
CLID (ours)	Level 7	There are three giraffes in the dirt. There are green trees behind the fence. The giraffes are very tall. The giraffes are in the shade. There is a crowd of the giraffe. There is a crowd of the people behind the giraffes. The ground is dirt.
	Level 8	There are four giraffes on the dirt. There are green trees to the left of the giraffes. The giraffes are tall. The ground is very bright. There is a fence behind the giraffes. There are other people in the background. The sun is shining on the ground. There is a person standing behind the giraffes.
	(a) Length	(b) Generated Paragraph

* Note that people do exist in the image, although not mentioned in the GT paragraph.



GT: person rides a surfboard on top of a wave. The wave is large and filled with foam. There is another surfer behind the surfer who is riding the wave. The sky above the surfers is hazy and blue. There are other smaller waves in front of the wave the surfer is riding on. There is a wet beached covered in foamy waves along the body of water that the waves are in.

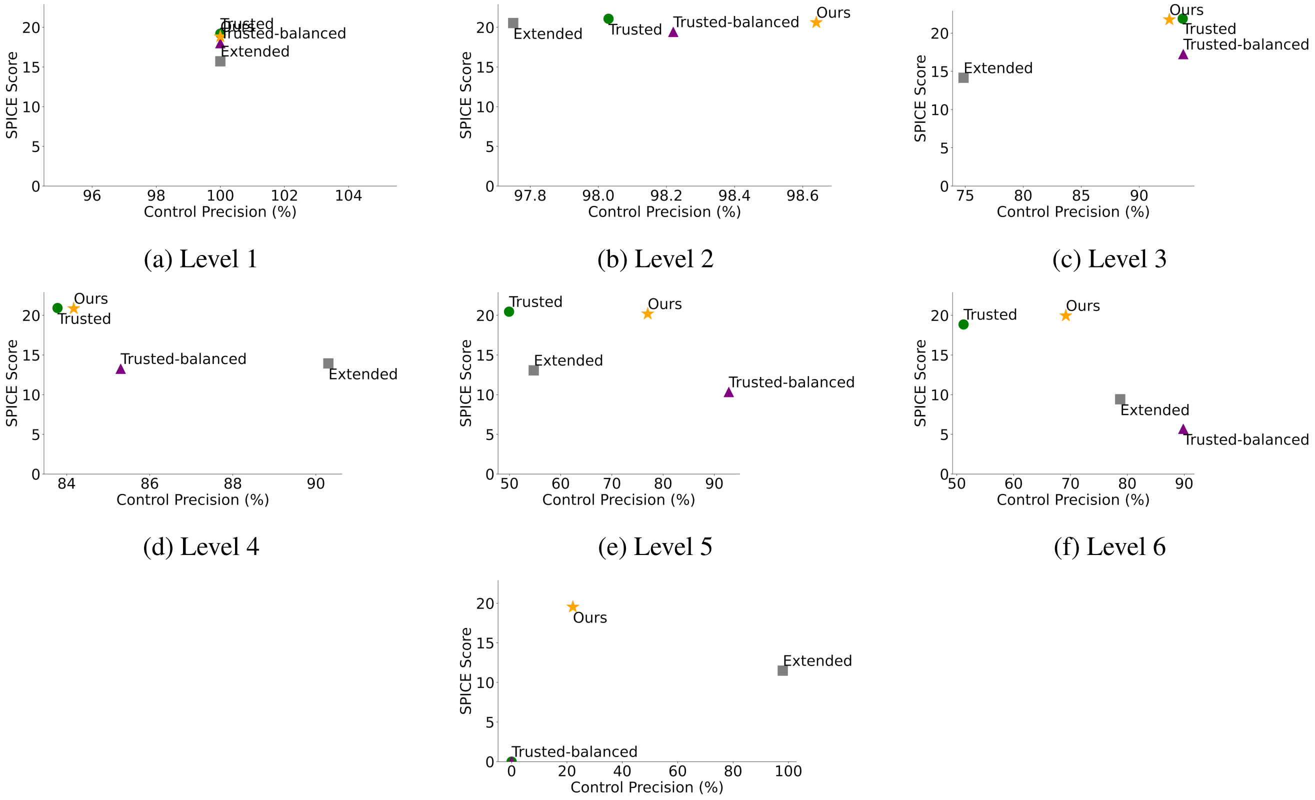
Generated Paragraphs:

LaBERT [14]	Level 1	Three people are in the ocean.
	Level 9	This is an image of a ocean. The water is very rough and the waves are moving fast. There waves are large and fast. There are three surfer in the picture. The sky is very cloudy and graycast there are many waves in the water. The surfers are in the water. The water is very violent. There is a single surfer surfing in the water. The photo is white and black.
CLID (ours)	Level 1	A person on a surfboard.
	Level 9	The image is of the ocean. There are white waves in the water. The water is all white. There are many waves in the water. There is a man surfing in the middle of the wave. The person in the water is wearing black pants. There are many ripples in the crest of the wave. There is a man on a surfboard. There is another person in the water out of the wave.
	(a) Length	(b) Generated Paragraph

4. Captioning - Additional Results (MS-COCO)

Quality vs. Control

This is the full figure for Fig. 5 in the paper.



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

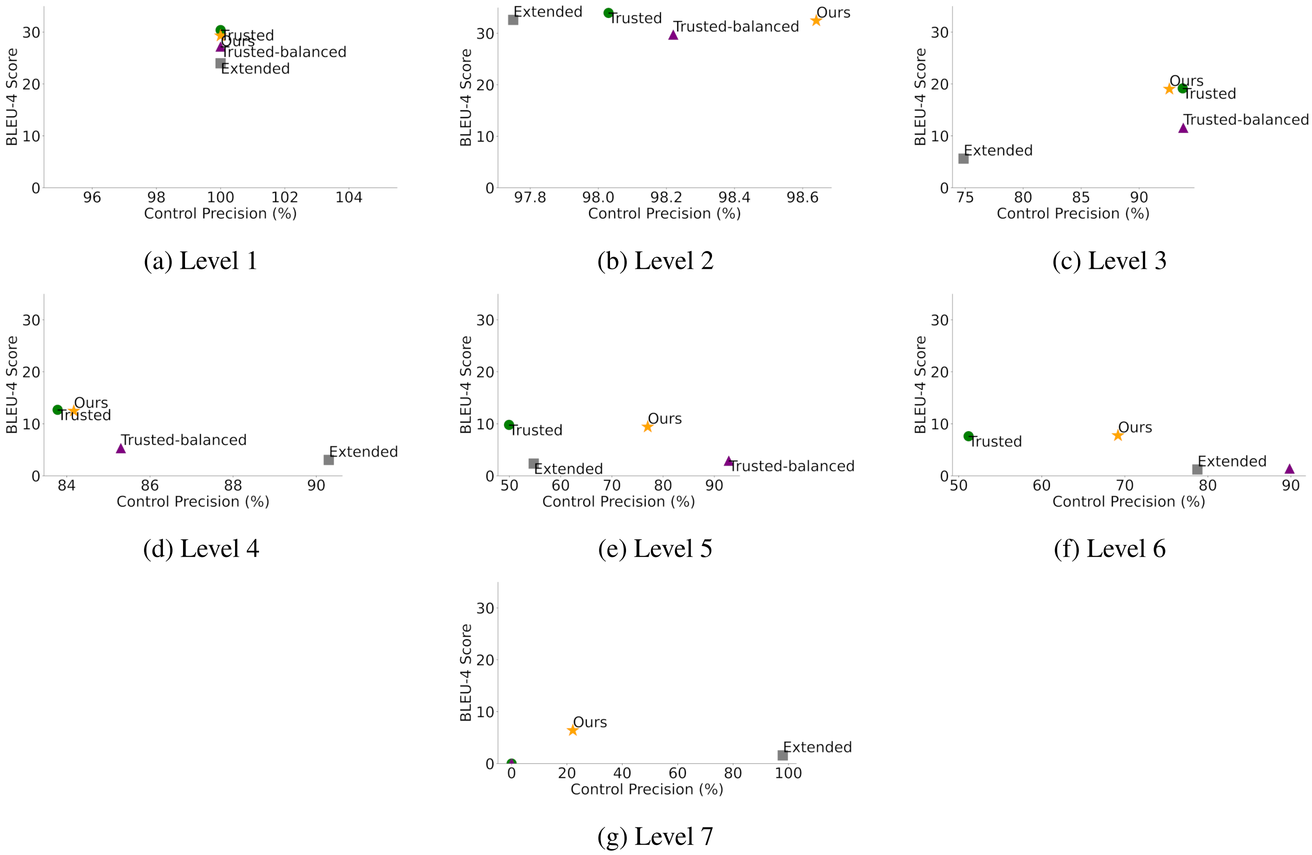
(f) Level 6

(g) Level 7

Figure 8. **Performance comparison to baselines (captions).** Control precision (horizontal axis) vs. SPICE score (vertical axis). (a)-(g) show the results for different length levels. The base model of [14] (green) and our model (orange) achieve similar quality (SPICE) results, while our model also improves control precision. The quality of other solutions (gray & purple), trained only on the extended dataset and on a balanced version of the trusted dataset (respectively), is dramatically degraded.

Similar results with other metrics (BLEU-4, ROUGE, METEOR & CIDEr) can be seen in Fig. 9- 12.

As previously discussed in [14], note that these metrics are n-gram based, making them less suitable for captions with high lengths compared to SPICE.



(a) Level 1

(b) Level 2

(c) Level 3

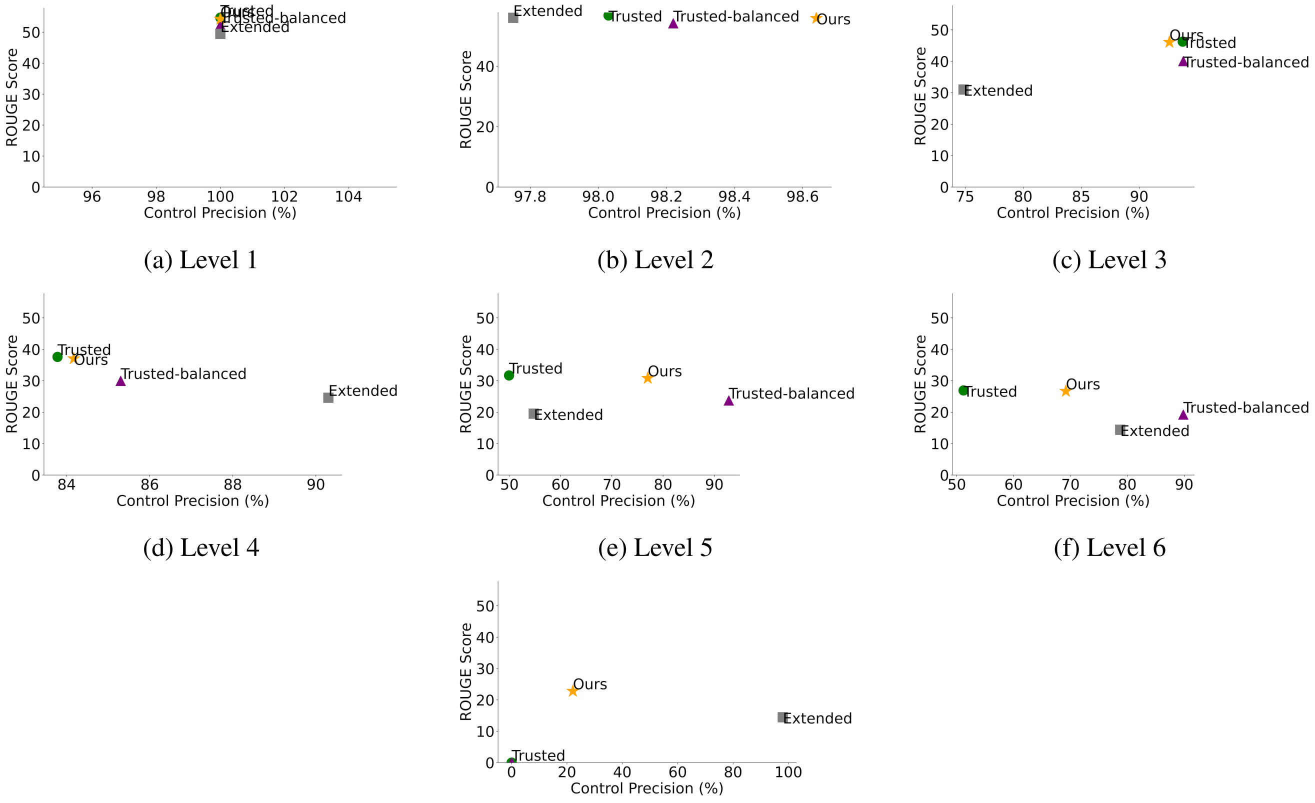
(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

Figure 9. **Performance comparison to baselines (captions).** Control precision (horizontal axis) vs. BLEU-4 score (vertical axis).



(a) Level 1

(b) Level 2

(c) Level 3

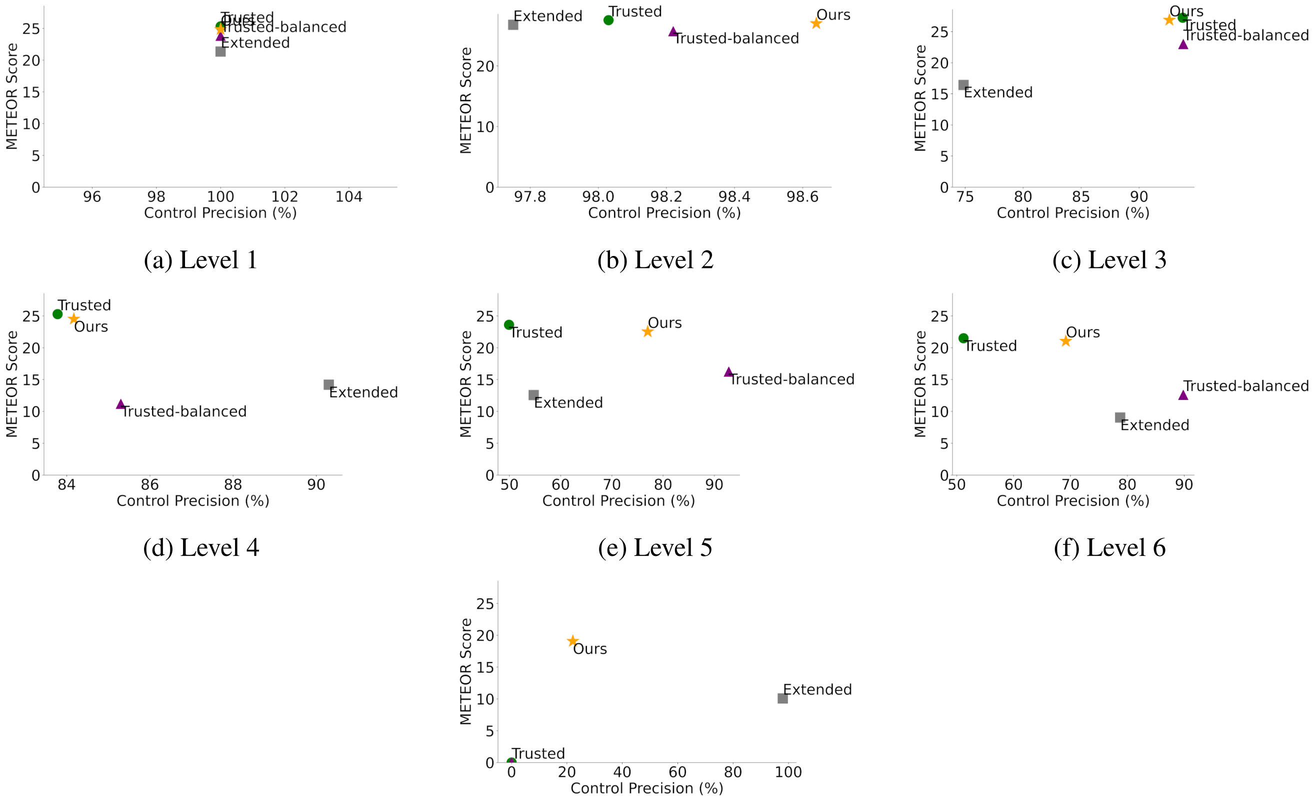
(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

Figure 10. **Performance comparison to baselines (captions).** Control precision (horizontal axis) vs. ROUGE score (vertical axis).



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

Figure 11. **Performance comparison to baselines (captions).** Control precision (horizontal axis) vs. METEOR score (vertical axis).

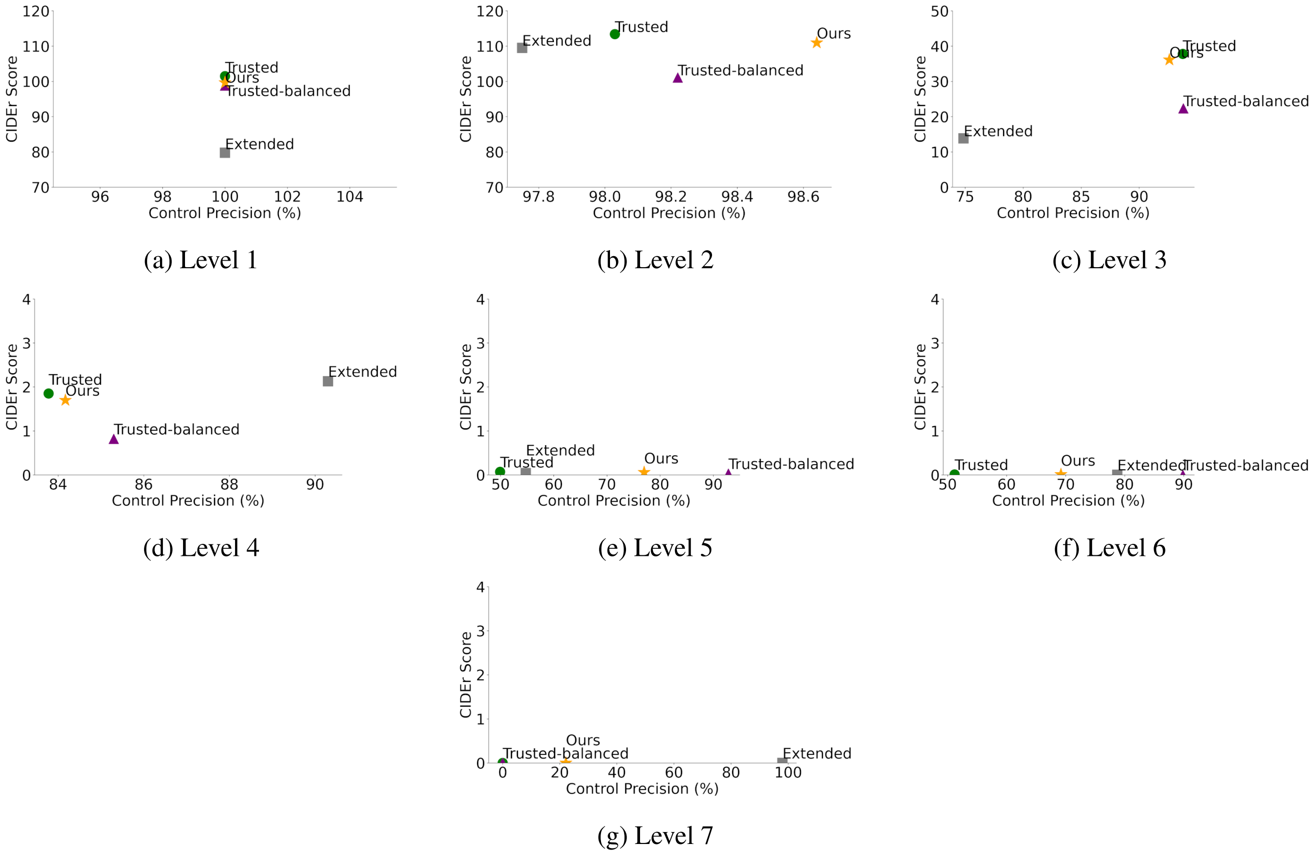
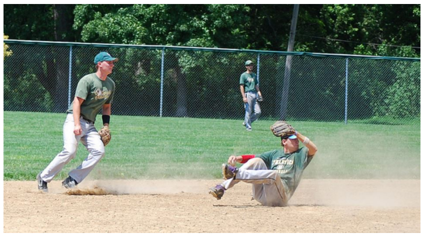


Figure 12. **Performance comparison to baselines (captions).** Control precision (horizontal axis) vs. CIDEr score (vertical axis). As ground-truth captions are short, CIDEr score is expected to drop for long captions for all methods.

Quantitative Results

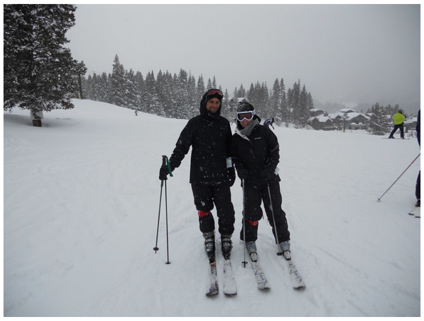
We present additional examples for outputs of our model and of [14]'s. Furthermore, we show the ground-truth captions, which are indeed short.



GT	1. baseball players fielding a hit in the infield. 2. a baseball game where a player has fallen on the ground. 3. a couple of baseball players running a play during the game. 4. baseball players catching a ball in the field. 5. an action shot of some baseball players on a field

Generated Captions:

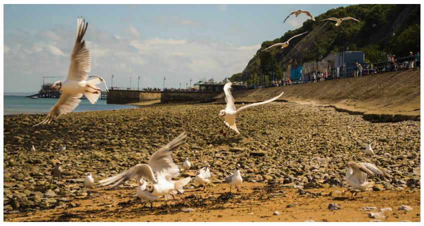
Length	LaBERT [14]	CLID (ours)
Level 1	A group of baseball players on a field.	Three men playing baseball together in a field.
Level 2	A baseball player sliding in to a base at a game.	A baseball player catching a ball on a field.
Level 3	A baseball player sliding into a base while another other player is on the other team running behind him.	A baseball player in the middle catching a ball with another player tries to the side on the field.
Level 4	A baseball player sliding into a base while another player stands behind him and another player is standing in the foreground with another player standing on the field.	A baseball player getting ready to catch a baseball ball while another player catches the ball in the otherfield behind him with the other baseman in the background.
Level 5	-	A baseball player wearing a green shirt and a cap is on the ground next to a man in a green shirt is getting ready to catch the ball with a man lying on the ground.
Level 6	A baseball game in progress with a man standing next to the fence and a player on the ground on the other side of the fence and close to the fence with a baseball player standing at the base and the other on the other side in the outfield.	A man in a green shirt wearing a baseball glove is throwing a baseball while a man wearing a green shirt is standing in the grass covered field and another man in a green shirt is on a baseball field with a fence in the background.



GT	1. two people on skis taking a photo in the snow. 2. there are two people that are standing together on the skies. 3. a man and woman take a skiing break. 4. two skiers stand next to each other outside on a snowy hill. 5. two people standing next to each other wearing skis.

Generated Captions:

Length	LaBERT [14]	CLID (ours)
Level 1	Two people on skis in the snow.	Two people are standing in the snow.
Level 2	Two men on skis standing on a snowy surface.	Two people standing together in the snow on skis.
Level 3	Two men on skis standing on top of a snow covered slope with skiers in the background.	Two snow skiers standing on a ski slope with a snow covered hill and trees in the background.
Level 4	A man on skis and a pair of skis standing on top of a snow covered pole with a group of other skiers in the background.	A pair of skiers with their skis standing on a snow covered ski slope on a snow covered slope with a forest and trees in the background.
Level 5	A man and woman are posing for a picture with a snow covered hill in the background while a man in a ski suit on is standing in the snow of a ski skier with another pair of skis on.	Two people wearing black pants and a black jacket a snow skis and a few trees in the distance in a snowy area with snow capped trees and a building with snow skis and two ski poles.
Level 6	Two people standing on a ski slope with skis in the background and skiers in the background and a snow covered hill with trees in the background and other skiers at the top of the hill in the background and other skiers in the background.	Two people on skis standing in a snow outfit standing on a side of a hill with trees in the background with their skis and a view of a building in the background and a person standing in the snow on a cloudy day.



GT	1. seagulls flying down to rock covered beach. 2. a rocky beach area with seagulls swooping down and some standing on the ground. 3. a flock birds sitting on top of a sandy beach. 4. seagulls flock to feed on a rocky, clay-colored beach. 5. a number of birds on a beach near the water.

Generated Captions:

Length	LaBERT [14]	CLID (ours)
Level 1	A formation of birds flying over a beach.	A flock of birds flying over a beach.
Level 2	A flock of seagulls flying over a sandy beach.	A flock of birds flying through the air at a beach.
Level 3	A number of seagulls flying over a beach with a pier and boat in the water in the background.	A flock of seagulls flying through the air with a large body of water in the background.
Level 4	A group of seagulls are flying over a body of water and a kite is on the ground and a boat is in the jade in the background.	A flock of seagulls flying in the air over a beach area with a body of water and a boat pulling in the background next to the ocean.
Level 5	A flock of seagulls flying over a sandy area with a body of water in the background and a kites sailing in the air with a boat in the background and a large ship in the background.	A group of birds flying on a beach with a shoreline in the background and several people standing on the sand near the water and a small boat on a beach with water and a city in the background.
Level 6	A flock of seagulls flying in the sky of a sandy area with some buildings on the ground and a kite in the air on the other side with a body of water and a ship in the background and a large body the water in the background.	A group of seagulls flying in the air on a rocky beach with a boat in the background and people standing on the beach and a bunch of birds on the water and a seagull flying on the beach and a ship in the distance.

* Note that our model manages to mention people & boats in the captions, which do exist in the image although may be hard to detect in the small image.

5. Flickr30k - Details & Results

Data Histograms

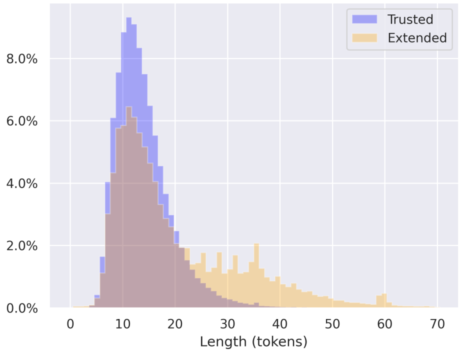


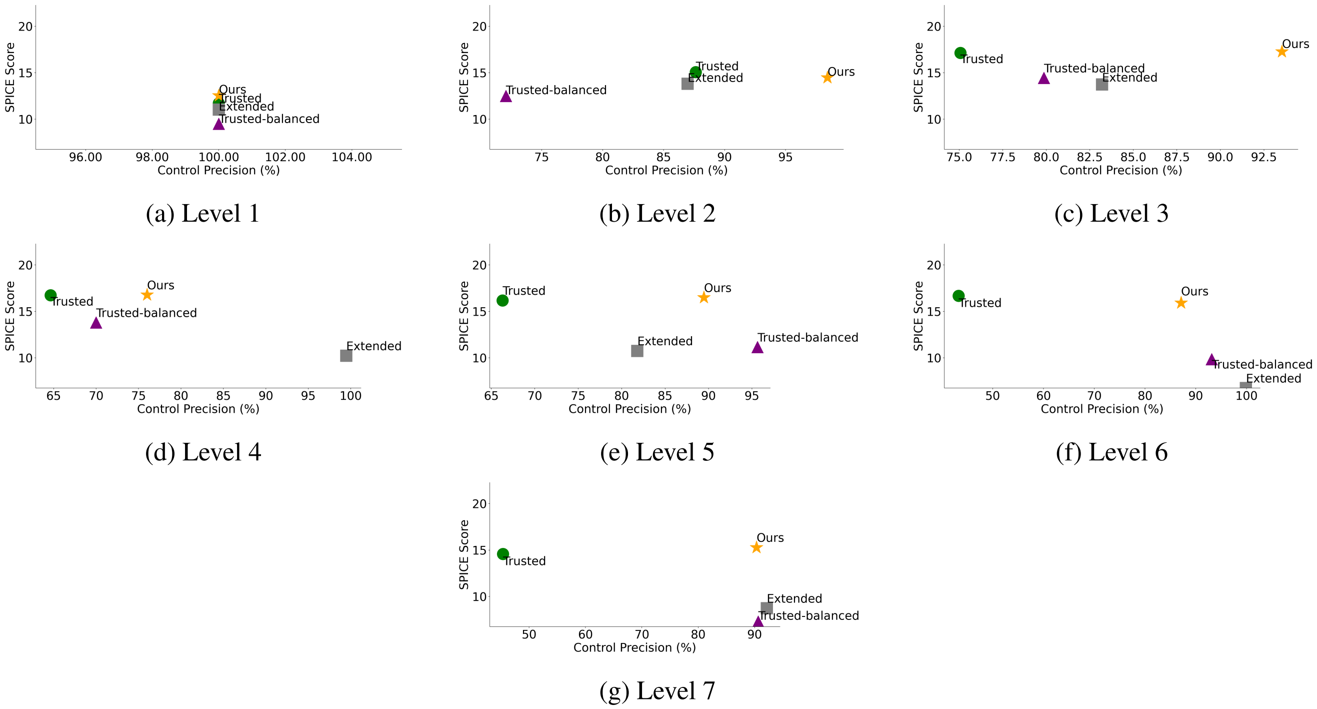
Figure 13. **Length of Flickr30k datasets.** The lengths of captions in the trusted dataset ([48]; blue) compared to the extended dataset (orange). (Overlaps cause the third color.)

Control Precision

Level Length (tokens)	1 1-9	2 10-19	3 20-29	4 30-39	5 40-49	6 50-59	7 60-69	Average
% in trusted dataset	20.1%	67.3%	11.01%	1.3%	0.2%	0.04%	∼0%	
% in extended dataset	14.5%	46.8%	15.6%	11.5%	5.6%	4.9%	0.6%	
LaBERT [14]	**100%**	87.63%	75.06%	64.63%	66.3%	43.3%	45.36%	68.89%
CLID (ours)	**100%**	**98.43%**	**93.53%**	**76.0%**	**89.46%**	**87.1%**	**90.33%**	**90.69%**

Table 2. **Captioning control precision (Flickr30k)**. For each level, the second row shows the range of tokens for this level. The next two rows show the percentage of captions of each level in the training datasets. The two bottom rows compare the precision results of [14], trained on the trusted dataset, to those of our method. Our results outperform [14]'s both on average and for all of the levels.

Control vs. Quality



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

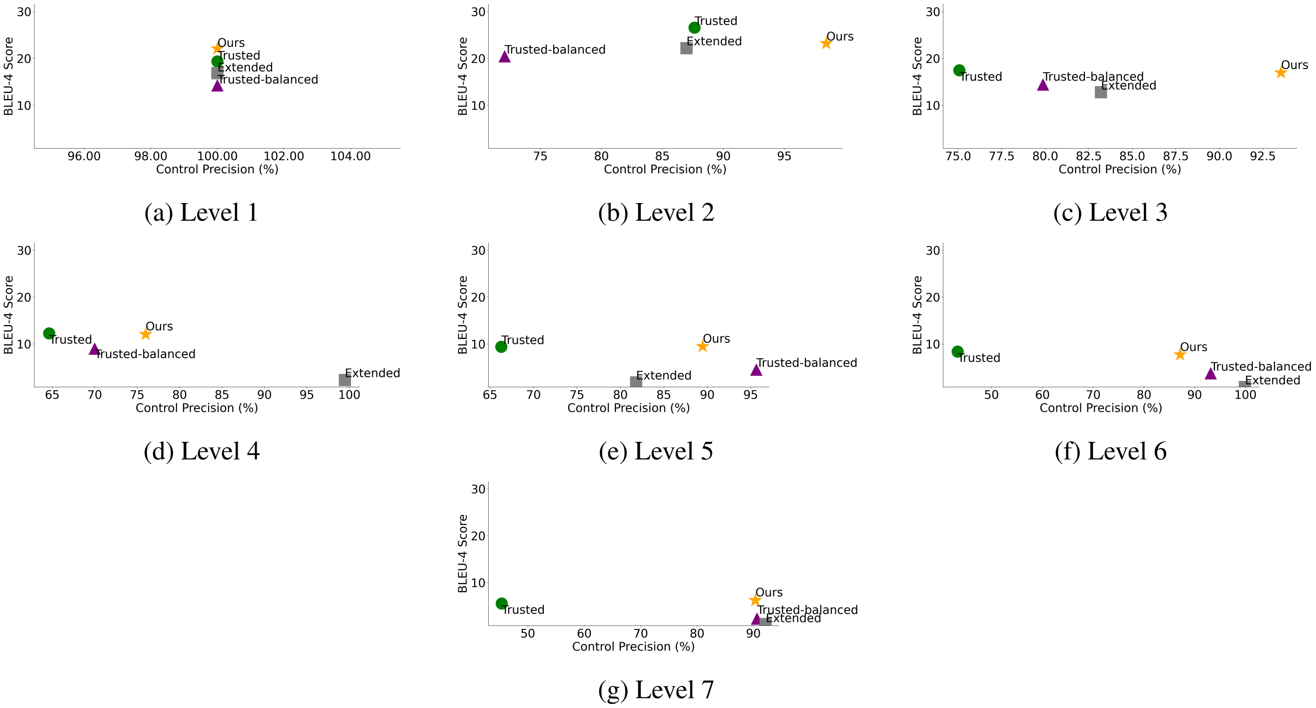
(f) Level 6

(g) Level 7

Figure 14. **Performance comparison to baselines (Flickr30k).** In terms of the SPICE quality measure (vertical axis), our results (orange star) are similar to [14]'s (green circle), which is trained on the trusted dataset. The quality of other solutions (gray & purple), trained only on the extended dataset and on a balanced version of the trusted dataset (respectively), is dramatically degraded. While comparable to [14] quality-wise, our model highly improves the control precision (horizontal axis) at all levels. In both measures, higher is better.

Similar results with other metrics (BLEU-4, ROUGE, METEOR & CIDEr) can be seen in Fig. 15- 18.

As previously discussed in [14], note that these metrics are n-gram based, making them less suitable for captions with high lengths compared to SPICE.



(a) Level 1

(b) Level 2

(c) Level 3

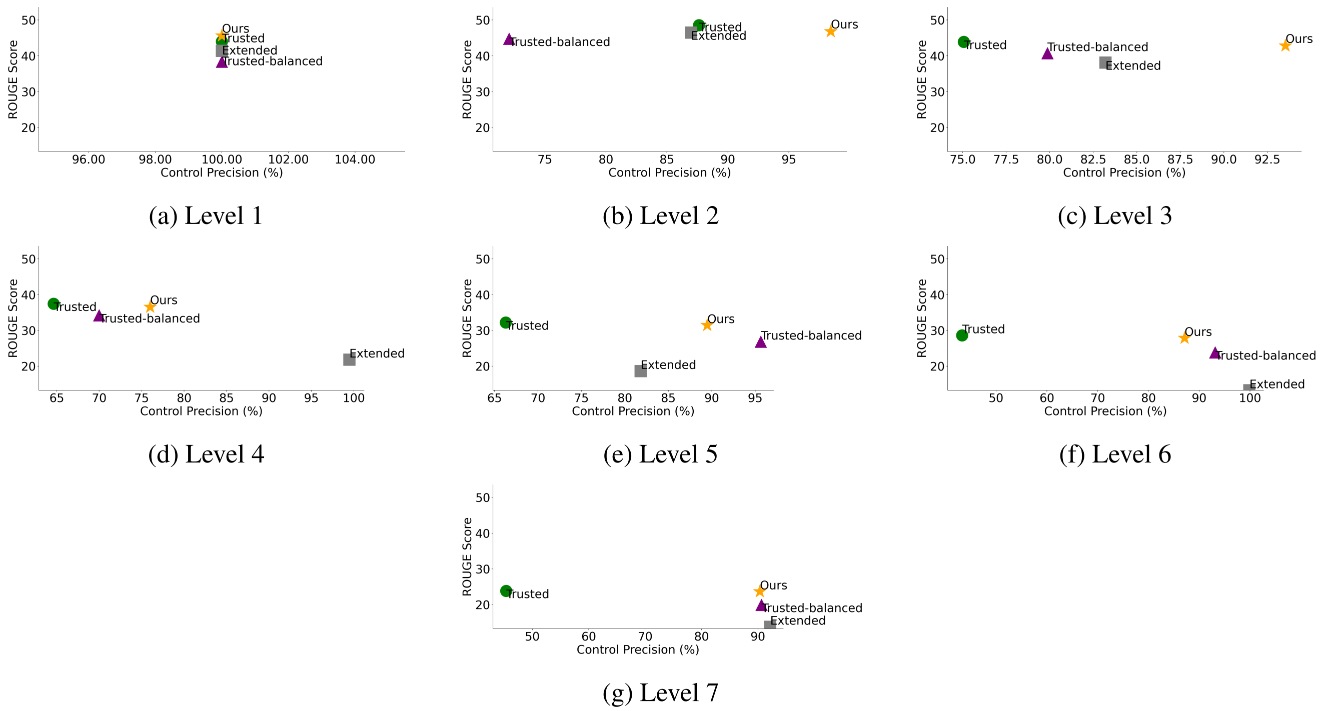
(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

Figure 15. **Performance comparison to baselines (Flickr30k).** Control precision (horizontal axis) vs. BLEU-4 score (vertical axis).



(a) Level 1

(b) Level 2

(c) Level 3

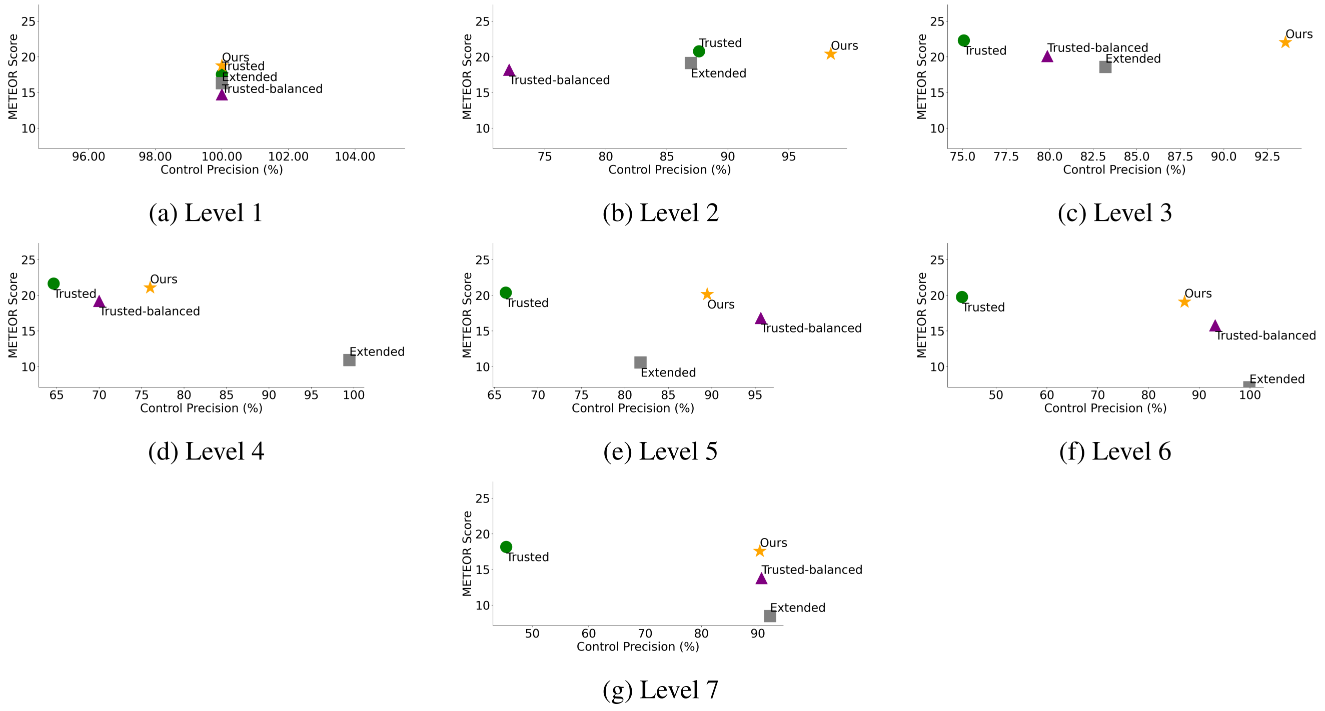
(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

Figure 16. **Performance comparison to baselines (Flickr30k).** Control precision (horizontal axis) vs. ROUGE score (vertical axis).



(a) Level 1

(b) Level 2

(c) Level 3

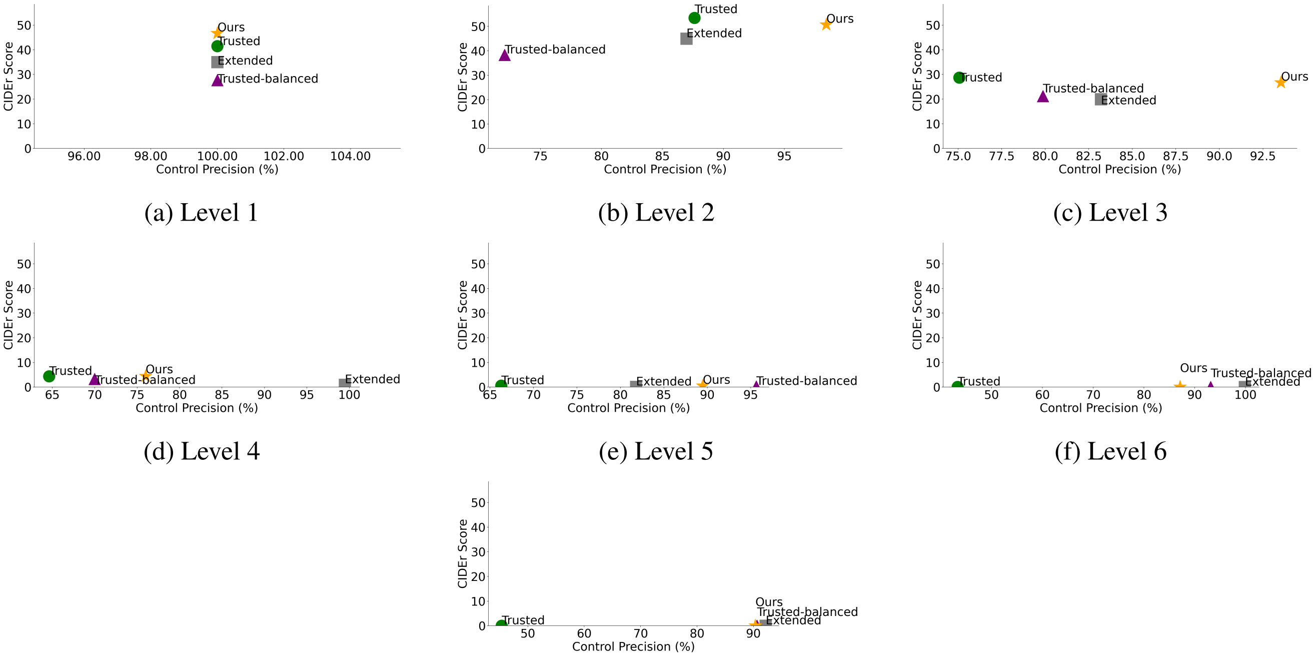
(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

Figure 17. **Performance comparison to baselines (Flickr30k).** Control precision (horizontal axis) vs. METEOR score (vertical axis).



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

(e) Level 5

(f) Level 6

(g) Level 7

Figure 18. **Performance comparison to baselines (Flickr30k).** Control precision (horizontal axis) vs. CIDEr score (vertical axis). As ground-truth captions are short, CIDEr score is expected to drop for long captions for all methods.

6. Self-Generated Dataset Examples

These description examples were extracted from the self-generated datasets (which were created from scene graphs). Recall that these descriptions, of diverse quality, are employed for training purposes and do not constitute the output of our model. The domain gap is bridged during training; for instance, the SPICE score of such descriptions, if used in place of our model's output, is lower by 30-60% (depending on the level) than that of our captions.

Single Sentence

1. A person wearing a shirt in a room and a cabinet and a bed in the room.
2. A boy wearing a brown wrap holding a red umbrella and a cattle in a field of brown cows.
3. A young boy touching lips wearing black headphones next to a table with a computer.
4. A man standing in a kitchen and a baker wearing grey suspenders with spatulas on a wall and multiple silver pans hanging and a table in the kitchen.
5. A black dishwasher and a cabinet with drawers and a black microwave in a kitchen and burners on a stove and a cutting board and utensils in a container and plastic glasses.
6. A man wearing shorts with a boy and a woman on a sidewalk with a tree behind the man.
7. A head of an elephant and a man sitting on the large animal and a giant rock in the water and a tree and an orange item.
8. A woman holding a black pan in the woman's hands in a kitchen and cooking utensils on a marble counter.
9. A man on a black wooden skateboard and a person riding white bike in a park and graffiti on a wall of a building and a person wears a helmet and a wall surrounding a skateboard park.
10. A gray runway under a plane and a red nose of the plane and a propeller and a colored tail and colored wing and green leaves behind a silver fence.

Paragraphs

1. A man wearing a red shirt flying a white black blue kite. Tracks in a sand. The man wearing tan pants. White strings attached to the black blue white kite. Grass covered background dunes.
2. Brown fries on a plate on a sandwich. A foil on the sandwich.
3. A train on a train's track. A number 3 printed on a sign next to the red train. A square window of the train. A word RENFE on the train.
4. Orange carrots between chopped light-colored mushrooms on a board. Vegetables on the board. A vegetarian nutritious meal. A chopped diced food.
5. A building next to a horse on a parking lot. A man on the brown horse. A window on the brick white building.
6. A woman sitting on a bench. A man with a red shirt on the bench. A small bush by the bench. A red bridge across a river. A tree by the water. A fence behind the bench.
7. A brown black dog on a colorful skateboard with red wheels on an asphalt. The brown dog with black spots. A harness on the dog. An item hangs from the dog's collar.
8. A bus with a number 8 on the front on a paved road. Leaves on trees on a sidewalk. An open window on the red bus. A dull wheel on the bus. A motorcycle and a yellow jeep on the road.
9. A clock on a colored design floor against a green light wall. A roman numeral on the clock. Shield wooden coat of arms carved into the antique wooden clock. An eagle on top of the large antique clock.
10. A smiling man wears a blue tie and pinstriped shirt. Bushy dark eyebrows are above brown eyes. A reflection in a reflective dark window behind the man. White blue flowers are on the blue tie. A white car in the window.