Supplementary Material for

Story Visualization by Online Text Augmentation with Context Memory

Daechul Ahn[1,§] Daneul Kim[2] Gwangmo Song[3] Seung Hwan Kim[3]
Honglak Lee[3,4] Dongyeop Kang[5] Jonghyun Choi[1,†]

[1]Yonsei University [2]GIST [3]LG AI Research [4]University of Michigan [5]University of Minnesota

{dcahn,jc}@yonsei.ac.kr flytodk98@gm.gist.ac.kr gwangmo.song@lgresearch.ai
skcruise@gmail.com honglak@eecs.umich.edu dongyeop@umn.edu

Note: We use blue color to refer to figures, tables, section numbers and citations **in the main paper** (*e.g.*, [17]). We use orange color to refer to figures, tables, section numbers **in the reference paper**. We use red or green colors to refer to figures, tables, section numbers and citations in this supplementary material.

1. Detailed Comparison with VP-CSV [2]

Our proposed method (CMOTA) focuses on encoding general context (including the character semantic retention), while the VP-CSV heavily focuses on maintaining the character semantic information over time. Specifically, the VP-CSV is specialized in generating 'characters' in each image using a two-stage approach, *i.e.*, one) character token planning for generating characters and two) visual token completion. Particularly, for the 'character token planning', VP-CSV trains an auxiliary classifier model to classify the characters from given images, which is used for extracting character regions using Grad-CAM [12]. In the first stage (*i.e.*, character generating stage), the model generates a specific *character region* given the input sentence while the non-character regions are masked-out. Here, they utilize the pre-trained character classifier model to obtain character region's information (refer to the paper [2] for more information). In the second stage (*i.e.*, visual completion stage), the model completes the image created in the first stage. This two-stage process is beneficial in generating accurate characters that matches with each sentence in a story paragraph, but it shows marginal improvement in the image quality (*i.e.*, relative improvement of FID -1.05 between vanilla and full version of VP-CSV) Tab. 1 of [2].

In contrast, we observe relatively larger improvement in image quality with our CMOTA (*i.e.*, relative improvement of FID -11.75 between vanilla and full version of CMOTA) along with the higher global semantic matching scores (*i.e.*, R-precision and BLEU score) as shown in Tab.[2]. In other words, state-of-the-art VP-CSV [2] performs on par with our CMOTA in character-related metrics (*i.e.*, Char.F1, Frm. Acc.), while our CMOTA outperforms in all other metrics (*i.e.*, FID, BLEU, R-precision). VP-CSV shows higher performance in Char.F1 and Frm. Acc. compared to our CMOTA due to the character-centric module, but our method generates high quality image sequence that maintains global semantic matching with story paragraph compared to VP-CSV.

2. More Discussions on Relevant Literature of Text-to-Image Generation

Again, text-to-image generation can be considered as a sub-problem of the story visualization task. Most literature focus on enhancing the semantic relevance of the generated image for the input text description and on resolution improvements. MC-GAN [7] models both background and foreground information to generate photo-realistic foreground objects for a background. StackGAN [15] uses a two-stage process to enhance the resolution of the image conditioned on an input text description. Subsequent works focus on architectural enhancements over StackGAN [3, 14, 16, 17]; adding attention networks for improved semantic relevance [14], extending the two-stage process [3, 16], or adding memory networks to improve the resolution of generated images and others [17].

Recently, text-based image synthesis has been greatly improved with the help of a vast amount of training data with a hyper-scale model. DALL-E [32] and CogView [5] concurrently propose an auto-regressive transformer to model the text and image as a single data stream. Recent studies use a diffusion model for this task. Its benefits include no need of adversarial learning and better scalability compared to GAN's, making diffusion models attractive in the literature [2]. As the research matures, hyper-scale diffusion models [31,34] generate state-of-the-art quality images in zero-shot fashion. But due to its persisting high computational cost, LDM [11] generates image in latent space, thereby decreasing the computational complexity. More recently, Make-A-Scene [7] to generate images that follow human's prior with a simple sketch is proposed.

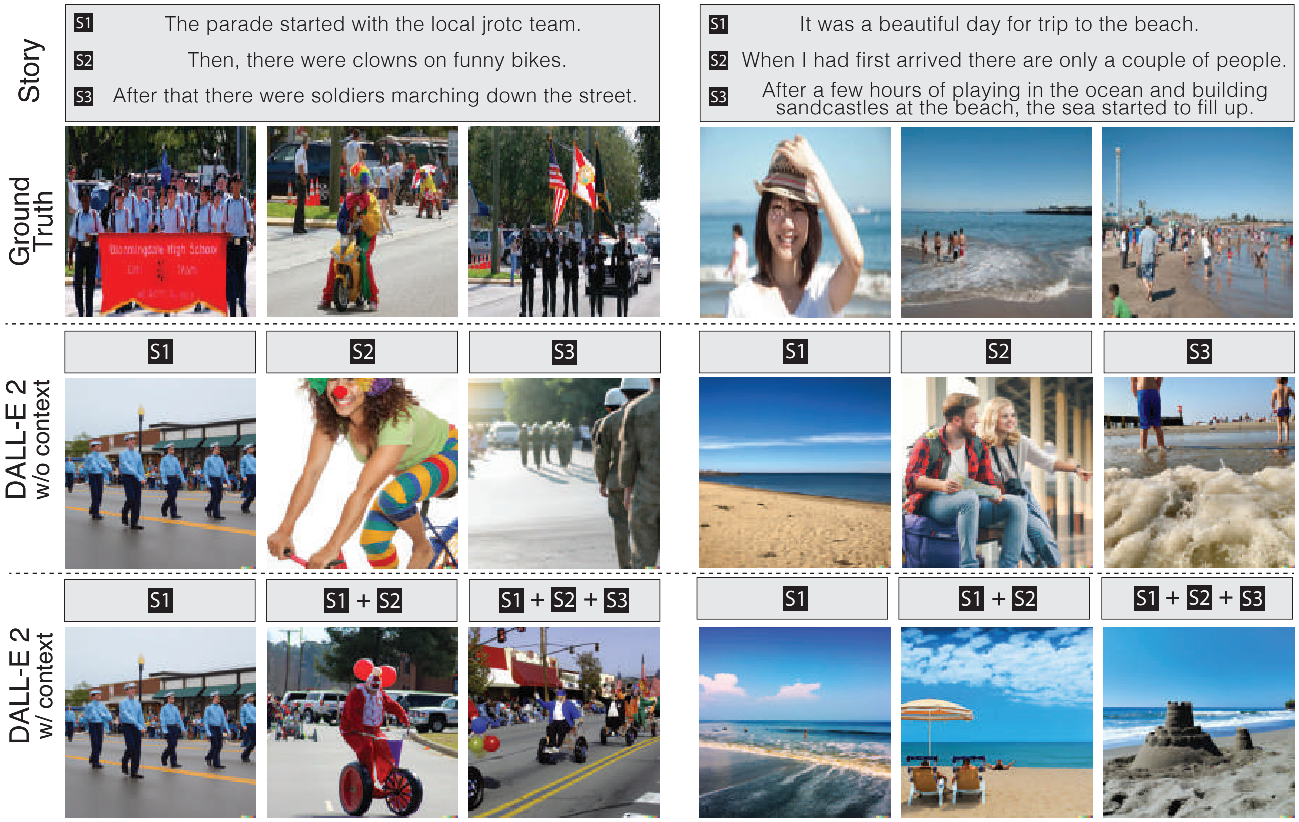

Story
S1 The parade started with the local jrotc team.
S2 Then, there were clowns on funny bikes.
S3 After that there were soldiers marching down the street.
S1 It was a beautiful day for trip to the beach.
S2 When I had first arrived there are only a couple of people.
S3 After a few hours of playing in the ocean and building sandcastles at the beach, the sea started to fill up.
Ground Truth
DALL-E 2 w/o context
S1
S2
S3
S1
S2
S3
DALL-E 2 w/ context
S1
S1 + S2
S1 + S2 + S3
S1
S1 + S2
S1 + S2 + S3

Figure 1. **A preliminary study to generate images from a text using a large scale pre-trained text-to-image model DALL-E2 [31].** The 'context' in the second and last row refers to *historical contexts*, which we provide as concatenated previous sentences.

Although text-to-image generation models work well with high-quality visualization abilities, it lacks an understanding of context in abstract, metaphoric sentences, *i.e*. *story*. In addition, naively using state-of-the-art text-to-image generation models is computationally prohibited. For example, diffusion-based models [31,34] have hyper-scale model size, (*e.g*., Imagen [34] parameter count of 2-B, DALL-E2 [31] parameter count of 3.5-B) making it non-trivial for applying it in a wide range of inference scenarios that may not have the sufficient computing resource. Here, we consider relatively light architectures as our base model for computational efficiency.

3. A Preliminary Study Using the DALL-E 2 (a popular large image generation model) for the Story Visualization

As the large-scale pre-trained model can be trivially used as a story visualizer, we conduct a preliminary study using the pre-trained text-to-image generation model, *i.e*., DALL-E 2[1] [31], for generating image sequence on the real-world story visualization benchmark dataset [5][2]. First, we use it as a single text-to-image generation task, *i.e*., by using each sentence in a story paragraph, we produce images. As shown in the second row of Fig. 1, *i.e*., DALL-E2 *w/o context*, it generates semantically well-aligned images for each sentence. But we observe drastic changes of background in second image of both examples in Fig. 1, showing inconsistent sequence of images compared to ground truth. In the second row of the first example (left), first image shows 'parade', but second image only shows 'clown' not related to 'parade' or 'jrotc team' of the first image. Moreover, in second row of the second example (right), first image shows 'beach', but second image only shows 'a couple' not related to 'beach' of the first image.

To address the problem of abrupt scene changes, we feed the DALL-E2 with historical contexts by concatenating the past sentences as input. Third row in Fig. 1, *i.e*., 'DALL-E2 w/ context', shows the result that is more temporally cohesive compared to the second row in Fig. 1. In the first example (left), second image in the third row shows 'clown' on the road (*i.e*., visually relevant to 'parade' and the previously generated image), thereby showing that the contextual information is better encoded for generating temporally coherent images. Also, in the second example (right), a picture that a couple sitting there is illustrated, which is temporally coherent. These results imply the effectiveness of the context information for generating temporally coherent images. But, the third image in the third row in the first example (left) shows that is less relevant to 'soldiers' which is in need to be generated in a third image, rather than showing parade on the funny bikes.

[1] https://labs.openai.com/

[2] In this experiment, we use a real-world benchmark dataset because the pre-trained generation model was trained for the real images not cartoons [31].

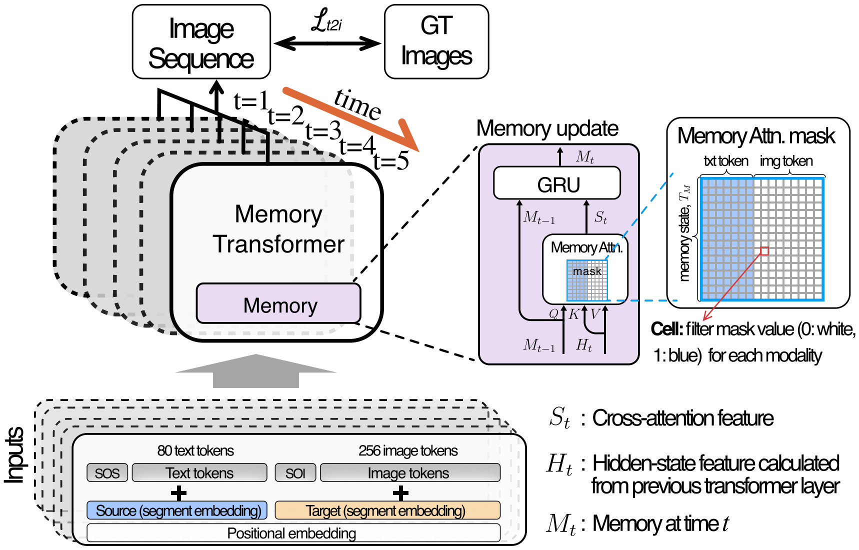


Figure 2. **Detailed input configuration for the proposed transformer.** As an input, we concatenate text and image tokens with two embeddings, *i.e*., 'SOS' (start of sentence token) and 'SOI' (start of image token), and then add positional and segment embedding. As we described in the main paper, we use memory-attention (memory attn.) mask to selectively determine which information to be propagated as a memory represented as a blue-colored box. Here, we illustrate using the 'text' information as memory content.

We believe that the reason the generated image does not semantically matches with the paired sentence is because the information spreads over multiple sentences thus requiring the generative model to discover the attentive words in the large set of words in the paragraph.

We observe that even the large model does not effectively encode the long term contexts when generating images. To this end, we propose to use memory transformer which adaptively utilizes historical contexts in a paragraph, with attentively weighted memory (Sec. 3.1). As these large-scale models are based on transformer architecture, our method may improve the large model's story visualization performance. But the computational complexity prevents us from using the large-scale models as our base model. As a promising future work, we eagerly want to incorporate our new memory module to the large models.

4. Memory-Attention Mask for Selective Modality Propagation

The memory update scheme from M_{t-1} to M_t is as follows. (1) l-th intermediate layer of transformer is modified for receiving previous memory from previous state, and (2) current state, H_t, and previous memory, M_{t-1} is passed to the memory updater shown in Fig. 2 purple box. To obtain a holistic understanding of current and memory state, we apply cross attention between the current and the past memory state by using the hidden state as a key/value and the past memory state as a query, respectively. Then, we choose which modality of input (*e.g*., text, image or text-image) to be propagated as memory content by using memory-attention mask, as mentioned in Sec. 3.1, which is illustrated in the cyan colored region of Figures 3 and 2. By applying the memory-attention mask onto attention score matrix, which is calculated from query (M_{t-1}) and key (H_t), we can choose which information to be propagated as a memory.

Here, we empirically investigate which modality (*i.e*., text or image or both) of historical contexts needs to be propagated into the future memory for generating temporally coherent image sequences. Table 1 shows comparative results using different modality as a memory. Interestingly, when we use the 'text' as a memory (third row in Tab. 1), it performs the best. In contrast, when we use the 'image' as a memory, we observe degradation in overall performance (first and second row in Tab. 1). We believe that because the story visualization task requires generating image frames arbitrarily distant in time (*i.e*., so-called 'key-frames') corresponding to different sentences, the 'image' information as a memory could be a strong constraint to the distant future, thereby hindering generation process at current time and degrading the overall performance.

5. A Discussion for Memory Connection Scheme

The conventional memory module [4,15] connects all the levels' intermediate layers as shown in Fig. 4-(a). But, it is not immediately clear how to connect the levels of memory modules in the multi-level transformer architecture for better contextual encoding. Inspired by prior studies, *i.e*., knowledge distillation [1, 4, 10], which mentions that structured and abstract representation can be extracted from deeper (high-level) layer [8, 9, 13], we propose to apply a partial same high-level connection path as shown in Figures 3-(a) and 4-(b). Although a partial-same level connection is a subset of the all-same level connection, we believe that rather many of connection paths would make it difficult to convey the necessary information as a memory.

To empirically validate our design for memory propagation, we compare with various design choices as shown in Fig. 3; (1) partial

Propagated modality in a memory	FID↓	Char. F1↑	Frm. Acc.↑	BLEU-2/3↑	R-Prec.↑
Text-Image	61.49	47.62	21.04	3.92 / 1.63	6.12
Image	63.25	46.56	18.36	3.72 / 1.53	5.77
Text	**59.05**	**49.72**	**21.79**	**4.41 / 1.77**	**6.28**

Table 1. **Story visualization performance by differently masked modality.** When we use 'text' as a propagated memory, we perform the best. In contrast, the 'image' as a memory degrade the performance. We conjecture that the SV task need to generate image frames arbitrarily distant in time, the 'image' information could be a strong constraint for the generating image at the current time.

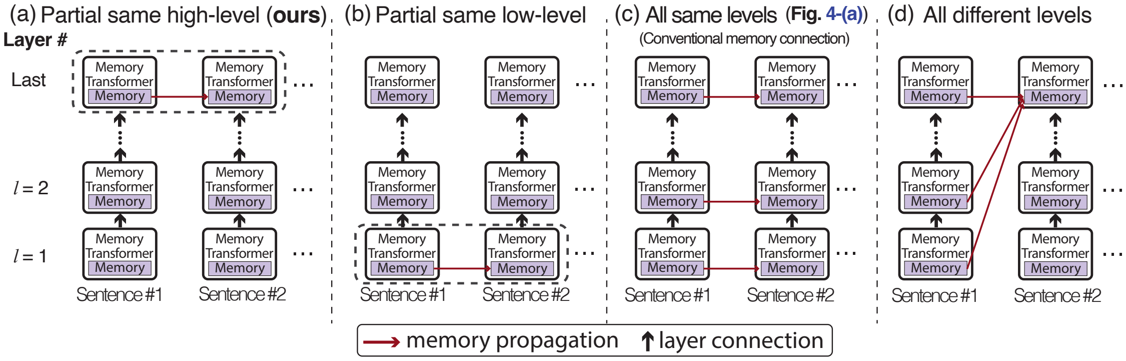


Figure 3. **Comparison to other possible memory connection schemes.** For propagating historical contexts using memory transformer, we investigate various memory connection path configurations between adjacent transformers, inspired by knowledge distillation [8, 9, 13]. Same as the main paper, we use red arrow and black arrow to indicate the direction of memory propagation and layers' feature propagation, respectively. (a) Connection between partial same high-level intermediate layers, (b) Connection between partial same low-level intermediate layers (c) Connection between all same level intermediate layers which connect all layers (Conventional memory connection [15] in Fig. 4-(a) of the main paper), (d) Connection between all different levels intermediate layers known as knowledge review that is similar to the human's learning strategy [10].

connection on same level with high-level feature (2) partial connection on same level with low-level feature, (3) all same level connection (conventional memory module) [15] and (4) all different level connection [10]. We can observe the effectiveness of the partial same level connection with high-level feature propagation in terms of performance and computational efficiency, thereby determining it as default design. Same as the main paper, we use red arrow and black arrow to indicate the direction of memory propagation and layers' feature propagation, respectively.

Connetcion Type	# Param.	FID↓	Char. F1↑	Frm. Acc.↑	BLEU-2/3↑	R-Prec.↑
w/o memory	93.7M	63.88	45.48	18.44	4.18 / 1.69	5.67
Partial same high-level (Fig. 3-(a))	95.8M	**59.05**	**49.72**	**21.79**	**4.41 / 1.77**	**6.28**
Partial same low-level (Fig. 3-(b))	95.8M	62.78	47.09	20.08	4.18 / 1.71	6.18
All same level (Fig. 3-(c))	118M	61.23	47.21	19.21	4.21 / 1.70	6.08
All diff. level (Fig. 3-(d))	115M	63.63	46.98	19.22	4.19 / 1.71	6.12

Table 2. **Story visualization performance for the considered memory connection schemes illustrated in the Fig. 3.** On Pororo-SV Test set. Fig. 3-(a) and (b) shows connection between partial same level intermediate layers (high or low-level connection), Fig. 3-(c) shows connection between all same level intermediate layers which connect all layers, Fig. 3-(c) shows connection between all different levels' intermediate layers.

6. Detailed Training Procedure

Here we explain the procedure of the proposed bi-directional training in detail with the proposed online text augmentation. The bi-directional generation, *i.e.*, text-to-image and image-to-text generation, is known to be effective in encoding multi-modal information [14,30,33]. To exploit the benefit, we use bi-directional training scheme to generate sequential data, *i.e.*, story paragraph to image sequence and image sequence to story paragraph generation simultaneously, with the aid of context memory (Sec. 3.1).

Thanks to this bi-directional generation, we can naturally integrate the procedure of generating pseudo-texts in an *online-manner* to the process of learning image-to-text and the text-to-image generation model as depicted in Fig. 5. However, at initial stage of training, online text-augmentation produces inappropriate pseudo-texts as the image-to-text generation model is not trained well yet. This could be harmful to the training of text-to-image generation. To address this issue, we propose a method to filter-out the inappropriate pseudo-texts by comparing the co-occurrence of *character's name* (*e.g.*, Pororo, Eddy, Poby, etc) between gold-label (*i.e.*, ground-truth caption) and generated pseudo-texts. For example, if the generated pseudo-text does not contain more than certain ratio of the character's name corresponding to the gold label, we reject it and train the model with ground-truth caption, as described in Alg. 1. Here, we set the character-occurrence threshold for filtering the generated pseudo-text as $p = 0.5$, depicted in Alg. 1. With the filtering, performance improves; FID (↓): 54.67 → 52.13, Char.F1 (↑): 48.96 → 53.25, Frm.Acc. (↑): 21.42 → 24.72 in 64×64 resolution Pororo-SV dataset.

Algorithm 1: Online Text Augmentation with Bi-directional Training

```
Given : Ground truth story with story paragraph and image sequence ($T_{gt-seq}$, $I_{gt-seq}$) in dataset $D_{story}$, Sequential index of
        image/text in story $j$, Story length $L$, j-th text from story paragraph $T_{j,gt}$, j-th text from generated pseudo-texts $T_{j,ps}$,
        Generated image sequence $I_{gen-seq}$, Generated text sequence $T_{gen-seq}$, Generated pseudo-text sequence $T_{ps-seq}$,
        Character name detector $Detr$, Characters in the description $\mathrm{Char.}$, Total epoch number $K_{epoch}$, CMOTA parameters $\theta$,
        Threshold of character occurrence $p$, Cross-entropy loss $CE$
Function $\mathrm{Character-Occurrence}(T_{gt-seq}, T_{ps-seq})$:
    for each ($T_{j,gt}, T_{j,ps}$) in ($T_{gt-seq}, T_{ps-seq}$) do
        $\mathrm{Char.}_{gt} \leftarrow Detr(\mathrm{T}_{j,gt})$;            ▷ Characters in the ground-truth text
        $\mathrm{Char.}_{ps} \leftarrow Detr(\mathrm{T}_{j,ps})$;            ▷ Characters in the pseudo-text
        if $|\mathrm{Char.}_{gt} \cap \mathrm{Char.}_{ps}| \, / \, |\mathrm{Char.}_{gt}| \leq p$ then
            return $False$;
    return $True$;
end Function
Initialize $\theta$;
$k \leftarrow 0$;
while $k < K_{epoch}$ do
    for each ($T_{gt-seq}, I_{gt-seq}$) in $D_{story}$ do
        // Bi-directional Training
        $T_{gen-seq} \leftarrow \mathrm{CMOTA}_{i2t}(I_{gt-seq}; \theta)$;            ▷ Image sequence to paragraph generation
        $I_{gen-seq} \leftarrow \mathrm{CMOTA}_{t2i}(T_{gt-seq}; \theta)$;            ▷ Paragraph to image sequence generation
        $\{\mathcal{L}_{j,t2i,\theta}\}_{j=1}^{L} = CE(I_{gen-seq}, I_{gt-seq})$;            ▷ Top equation of Eq. 1
        $\{\mathcal{L}_{j,i2t,\theta}\}_{j=1}^{L} = CE(T_{gen-seq}, T_{gt-seq})$;            ▷ Second row's equation of Eq. 1
        // Online Text Augmentation
        $T_{ps-seq} \leftarrow \mathrm{CMOTA}_{i2t}(I_{gt-seq}; \theta)$;            ▷ Image sequence to pseudo-texts generation
        if $\mathrm{Character-Occurrence}(T_{gt-seq}, T_{ps-seq})$ then
            $I_{gen-seq} \leftarrow \mathrm{CMOTA}_{t2i}(T_{ps-seq}; \theta)$;            ▷ Pseudo-texts to image sequence generation
            $\{\mathcal{L}_{j,pt2i,\theta}\}_{j=1}^{L} = CE(I_{gen-seq}, I_{gt-seq})$;            ▷ Top equation of Eq. 3
        else
            $\{\mathcal{L}_{j,pt2i,\theta}\}_{j=1}^{L} = 0$
        $\{\mathcal{L}_{j,\theta}\}_{j=1}^{L} = \{\mathcal{L}_{j,t2i,\theta}\}_{j=1}^{L} + \lambda_1\{\mathcal{L}_{j,i2t,\theta}\}_{j=1}^{L} + \lambda_2\{\mathcal{L}_{j,pt2i,\theta}\}_{j=1}^{L}$            ▷ Bottom equation of Eq. 3
        $\mathcal{L}_{\theta} \leftarrow \sum_{j=1}^{L} L_{j,\theta}$
        $\theta \leftarrow \mathrm{Optimizer}(\nabla_{\theta}\mathcal{L}_{\theta})$            ▷ Update model parameters $\theta$
    $k \leftarrow k + 1$;
end
```

7. Experimental Details

7.1. Details of Datasets

Following previous works [2,17,22,23,39], we use **Pororo-SV** dataset proposed in [17], which is a modified version of [6] for story visualization task. Each story sample consists of 5 images as a sequence with corresponding 5 descriptions. As mentioned in previous works [22], there is a lot of data overlap between training and test samples in the original dataset split of Pororo-SV dataset [17,39]. To be more challenging, we follow the dataset split proposed in [22], which contains 10191/2334/2208 samples in training, validation and test splits, respectively. In this version, there is no data overlap between training and test split.

Furthermore, following the prior work [23], we conduct story visualization task with Flintstones dataset (**Flintstones-SV**) which was originally exploited in the text-to-video synthesis task. To construct story visualization dataset with it, five images for story sequence are sampled from short video clip (*i.e.*, 75 frames) and paired with language descriptions. To be consistent with prior work [22,23], we use dataset split proposed in [22].

7.2. Details of Evaluation Metrics

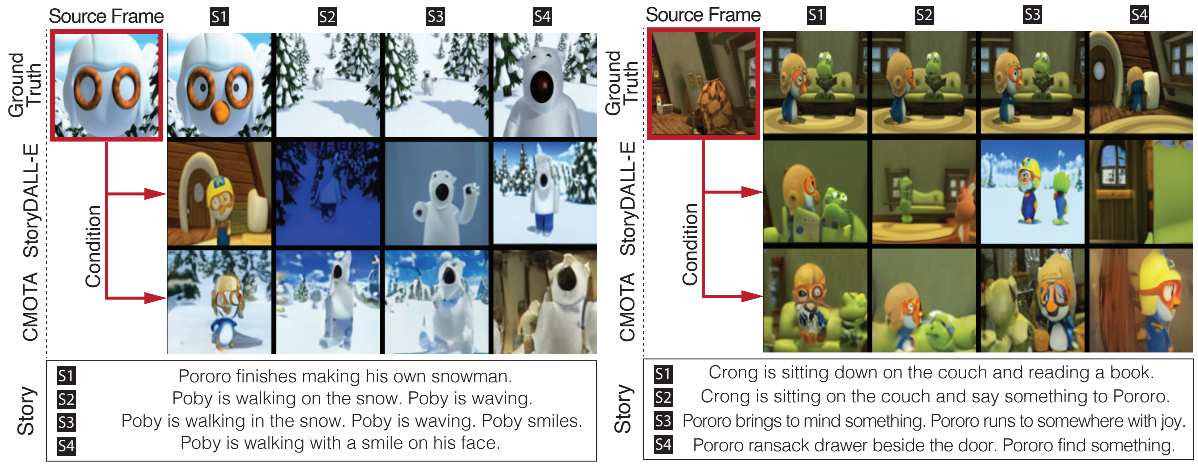
Due to the task complexity and its generative nature of the story visualization, we use various evaluation metrics for multi-faceted quantitative analysis. The analysis includes the visual quality of generated images, coherency in the generated image sequence and semantic matching between descriptions and generated images. We describe each evaluation method in detail as follows.

- **Fréchet Inception Distance (FID)**: Assessing the quality of generated image by calculating the distance of the distribution between generated and real images, which are used to train the generator as done in prior works [22,39].
- **Character Classification (Char. F1, Frm. Acc.)**: Assessing the presence of character in generated image sequence. Using pre-trained Inception-v3 with a multi-label classification loss to identify characters in the generated image. In particular, we report micro-averaged F-score of character classification (Char. F1) and exact matching using frame accuracy (Frame Acc.) as done in prior work [17,22,23].
- **Video Captioning Accuracy (BLEU-2/3)**: Assessing the global semantic matching between generated image sequence and captions. We report the BLEU2/3 (B-2/3) scores of captions predicted using generated images with pre-trained video captioner to fairly compare with prior works [2,22,23].
- **R-precision (R-Prec.)**: Assessing the global semantic matching quality between text paragraph and images in the story visualization task. We report retrieval-based metric R-precision following the prior work of [23] by quantifying the semantic aignment between the input text and generated image. With R relevant text as query, the top R-ranked retrieval results of a system are examined. With r relevant, R-precision would be r/R. Encodings for image retrieval tasks are based on Deep Attention-based Multimodal Similarity Model (DAMSM) and we train a new version Hierarchical DAMSM (H-DAMSM) following [22].
- **Human Evaluation**: Conducting human evaluation with the criterion listed in prior works [2,17,22,23,39].

8. A More Discussion on the Comparison with a Large-Scale SV Model (StoryDALL-E [24])

Despite the unfairness of model size and pretraining data, it is interesting to compare with the large-scale model StoryDALL-E (finetune) [24]. As the StoryDALL-E is proposed in the story continuation set-up, which uses first image as a condition to visualize the story, we evaluate our method in the same task set-up. We present the quantitative comparison in Tab. 5 and the qualitative comparison in Figures 4 and 5.

In Fig. 4, we observe that our CMOTA shows more temporally coherent image sequence compared to StoryDALL-E [24] on the both examples. In the first example (left) of Fig. 4, the second row shows 'Pororo' in 'house' in first image, and sudden background change occurs in the second image. Moreover, the second example (right) in the second row also shows 'Pororo' and 'Crong' outside in the third image as there's no background information in the paired-sentence. However, our CMOTA generates temporally coherent images in the both examples of Fig. 4 as we use a novel memory architecture to adaptively understand the historical contexts (Sec. 3.1).

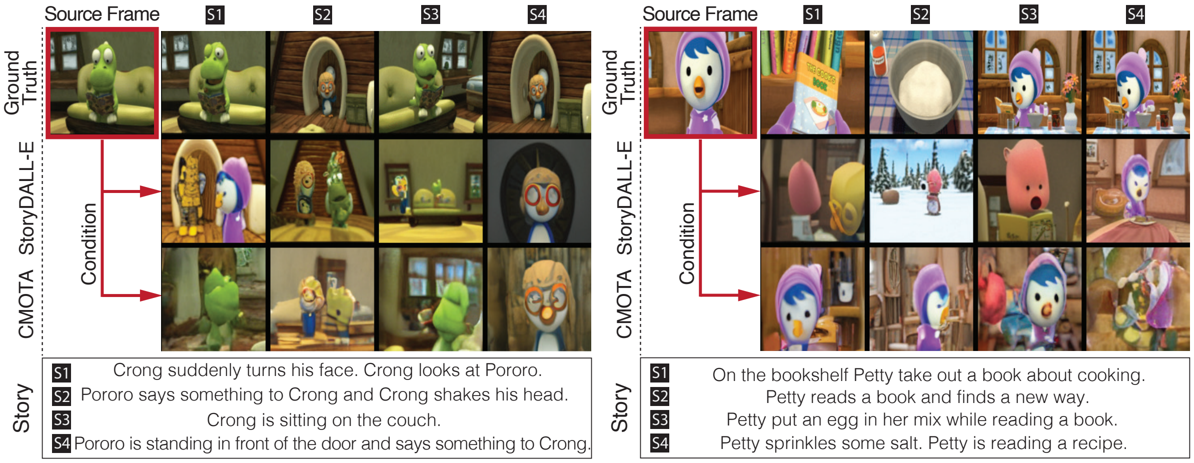


S1 Pororo finishes making his own snowman.
S2 Poby is walking on the snow. Poby is waving.
S3 Poby is walking in the snow. Poby is waving. Poby smiles.
S4 Poby is walking with a smile on his face.

S1 Crong is sitting down on the couch and reading a book.
S2 Crong is sitting on the couch and say something to Pororo.
S3 Pororo brings to mind something. Pororo runs to somewhere with joy.
S4 Pororo ransack drawer beside the door. Pororo find something.

Figure 4. **Qualitative comparison with StoryDALL-E (finetune) in the story continuation set-up [24].** Compared to the StoryDALL-E, our CMOTA generates *temporally* more coherent image sequence though our model does not include any special module for the story continuation task.

In Fig. 5, we observe that our CMOTA generates more semantically relevant image sequence with given story paragraph compared to the StoryDALL-E [24]. Second row in the first example (left) of Fig. 5 shows 'Petty' in the first image, not matching with the first sentence



S1 Crong suddenly turns his face. Crong looks at Pororo.
S2 Pororo says something to Crong and Crong shakes his head.
S3 Crong is sitting on the couch.
S4 Pororo is standing in front of the door and says something to Crong.

S1 On the bookshelf Petty take out a book about cooking.
S2 Petty reads a book and finds a new way.
S3 Petty put an egg in her mix while reading a book.
S4 Petty sprinkles some salt. Petty is reading a recipe.

Figure 5. **Qualitative comparison with StoryDALL-E (finetune) in the story continuation set-up [24].** Compared to the StoryDALL-E, our CMOTA generates *semantically* more relevant image sequence though our model does not include any special module for the story continuation task.

in a story paragraph. Also, the second row in the second example (right) shows 'Pororo' and 'Loopy' in the first and third images, even though the sentences only mention 'Petty'. However, our CMOTA shows semantically relevant images throughout the story paragraph.

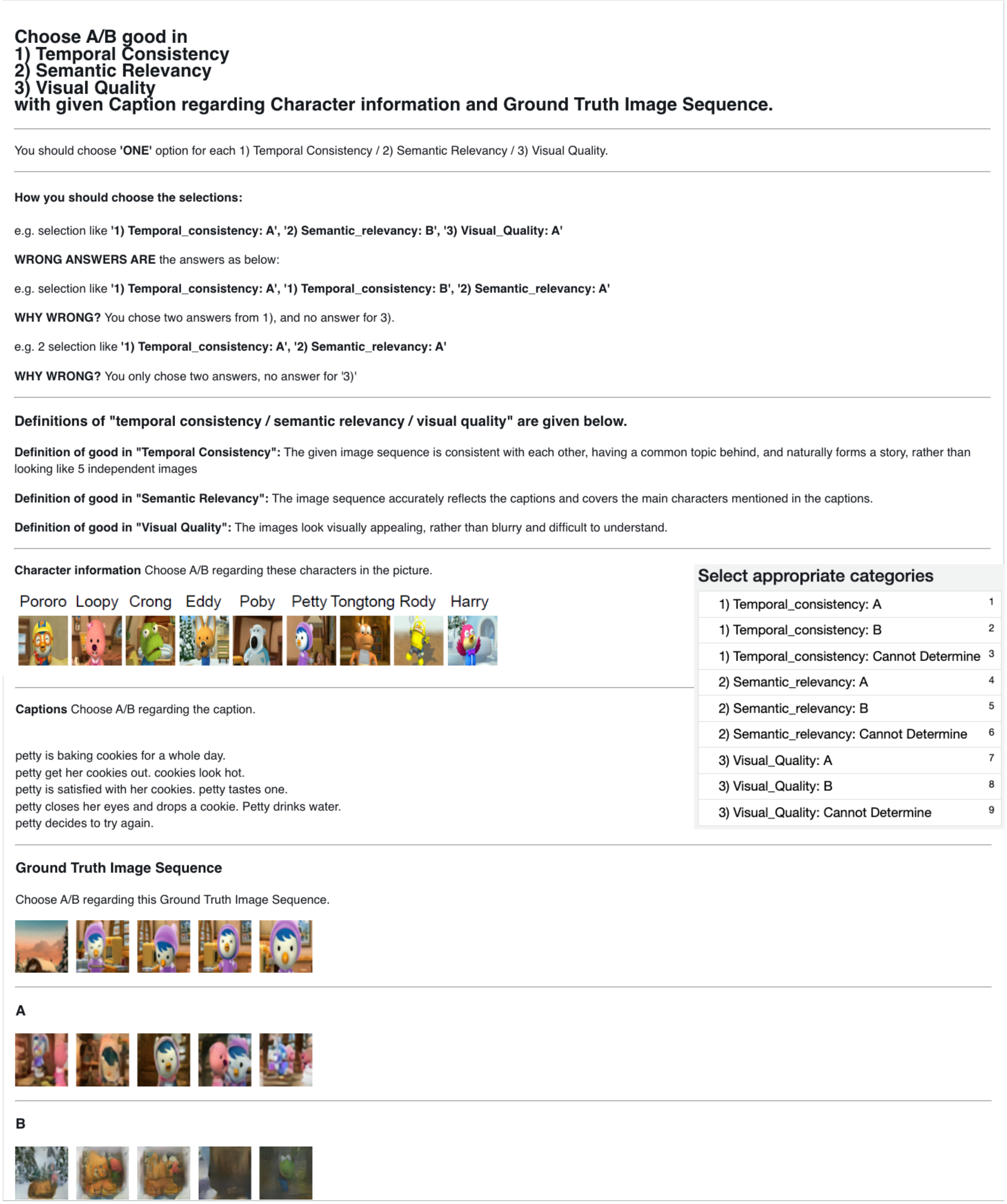
Note that the qualitative result is aligned with the quantitative result in Tab. 5 in the Char. F1 score and Frm. Acc.; our CMOTA preserves the character semantics over the multiple sentences better than the StoryDALL-E by the novel memory architecture.

9. Details about Human Preference Study

We conduct human evaluations by comparing ours (CMOTA) with VLC-StoryGAN [22] as it shows the best performance among all reproducible[3] prior arts we have compared. Specifically, we compare them on three criterions; *i.e.*, visual quality, temporal consistency and semantic relevancy, following prior works of StoryGAN [17], CP-CSV [39], DuCo-StoryGAN [23] and VLC-StoryGAN [22]. We recruit 100 annotators using Amazon Mechanical Turk platform. We ask them to blindly determine their preference for image sequences generated by VLC-StoryGAN [22] and CMOTA in Pororo-SV test split. They are asked to decide which image sequence is better in three perspectives. Here, we also consider 'Tie' that means 'can not determine' by their preference. Fig. 6 shows a screenshot of our annotation task. As shown in Tab. 6, we observe that the our model (CMOTA) shows better preference with a significant gap in all three criteria.

In the following pages, we present more qualitative results for further analyses of the proposed method.

[3] Note that the VP-CSV exhibits better performance but there is a reproducibility issue as the authors' implementation is not publicly available at the time of this submission.



Choose A/B good in
1) Temporal Consistency
2) Semantic Relevancy
3) Visual Quality
with given Caption regarding Character information and Ground Truth Image Sequence.

You should choose **'ONE'** option for each 1) Temporal Consistency / 2) Semantic Relevancy / 3) Visual Quality.

How you should choose the selections:

e.g. selection like **'1) Temporal_consistency: A', '2) Semantic_relevancy: B', '3) Visual_Quality: A'**

WRONG ANSWERS ARE the answers as below:

e.g. selection like **'1) Temporal_consistency: A', '1) Temporal_consistency: B', '2) Semantic_relevancy: A'**

WHY WRONG? You chose two answers from 1), and no answer for 3).

e.g. 2 selection like **'1) Temporal_consistency: A', '2) Semantic_relevancy: A'**

WHY WRONG? You only chose two answers, no answer for '3)'

Definitions of "temporal consistency / semantic relevancy / visual quality" are given below.

Definition of good in "Temporal Consistency": The given image sequence is consistent with each other, having a common topic behind, and naturally forms a story, rather than looking like 5 independent images

Definition of good in "Semantic Relevancy": The image sequence accurately reflects the captions and covers the main characters mentioned in the captions.

Definition of good in "Visual Quality": The images look visually appealing, rather than blurry and difficult to understand.

Character information Choose A/B regarding these characters in the picture.

Select appropriate categories

1) Temporal_consistency: A 1
1) Temporal_consistency: B 2
1) Temporal_consistency: Cannot Determine 3
2) Semantic_relevancy: A 4
2) Semantic_relevancy: B 5
2) Semantic_relevancy: Cannot Determine 6
3) Visual_Quality: A 7
3) Visual_Quality: B 8
3) Visual_Quality: Cannot Determine 9

Captions Choose A/B regarding the caption.

petty is baking cookies for a whole day.
petty get her cookies out. cookies look hot.
petty is satisfied with her cookies. petty tastes one.
petty closes her eyes and drops a cookie. Petty drinks water.
petty decides to try again.

Ground Truth Image Sequence

Choose A/B regarding this Ground Truth Image Sequence.

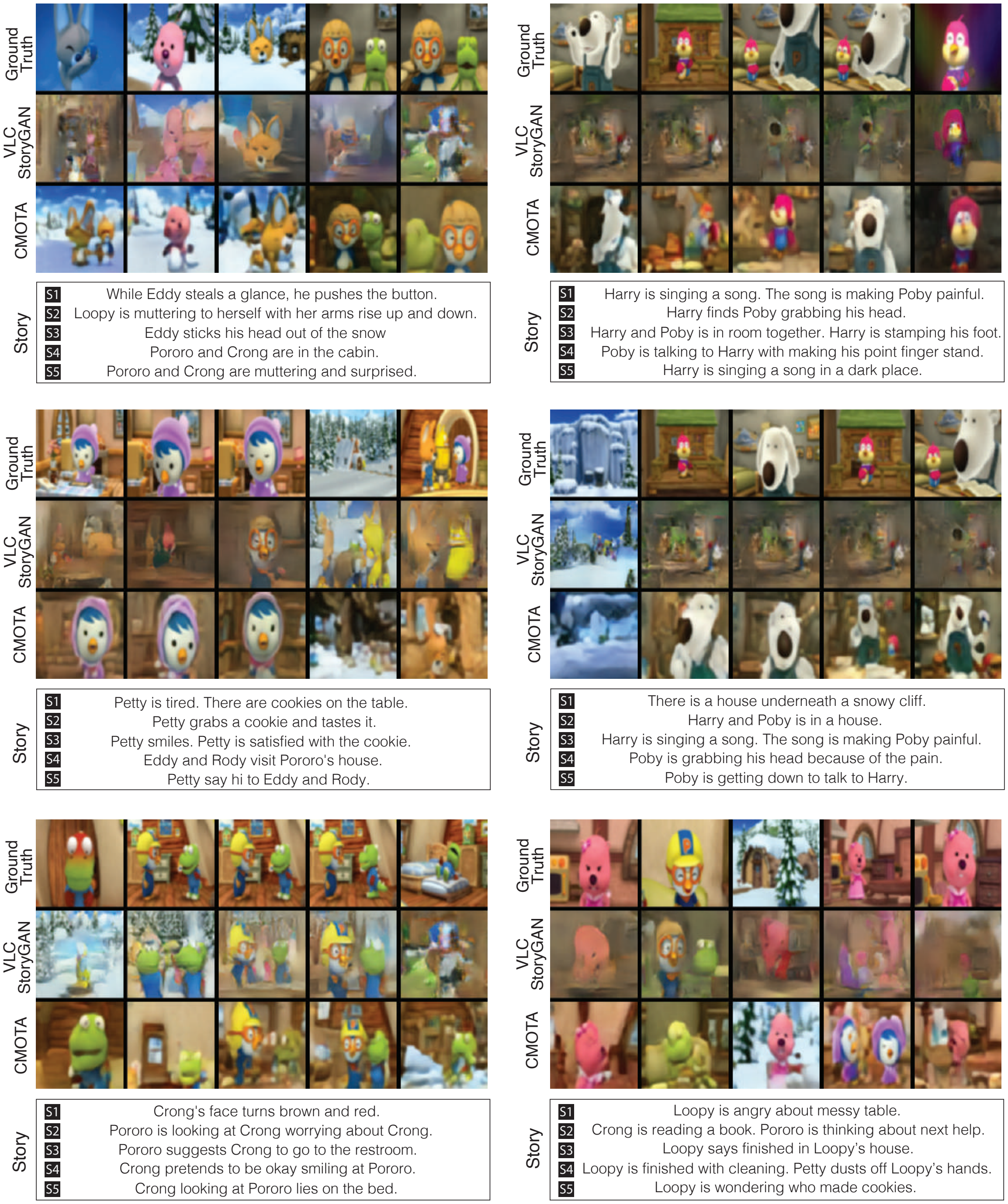
A

B

Figure 6. **The Mechanical-Turk evaluation page used in our human preference study.** We evaluate the results in three metrics; (1) temporal consistency, (2) semantic relevancy and (3) visual quality.

10. More Qualitative Results

We showcase more qualitative results compared to the prior art (*i.e.*, VLC-StoryGAN [22]) for Pororo-SV dataset on test split, as shown in Figures 7 and 8. We also showcase more qualitative results compared to the prior art (*i.e.*, DuCo-StoryGAN [23]) for Flintstones-SV dataset on test split, as shown in Figures 9 and 10.



S1 While Eddy steals a glance, he pushes the button.
S2 Loopy is muttering to herself with her arms rise up and down.
S3 Eddy sticks his head out of the snow
S4 Pororo and Crong are in the cabin.
S5 Pororo and Crong are muttering and surprised.

S1 Harry is singing a song. The song is making Poby painful.
S2 Harry finds Poby grabbing his head.
S3 Harry and Poby is in room together. Harry is stamping his foot.
S4 Poby is talking to Harry with making his point finger stand.
S5 Harry is singing a song in a dark place.

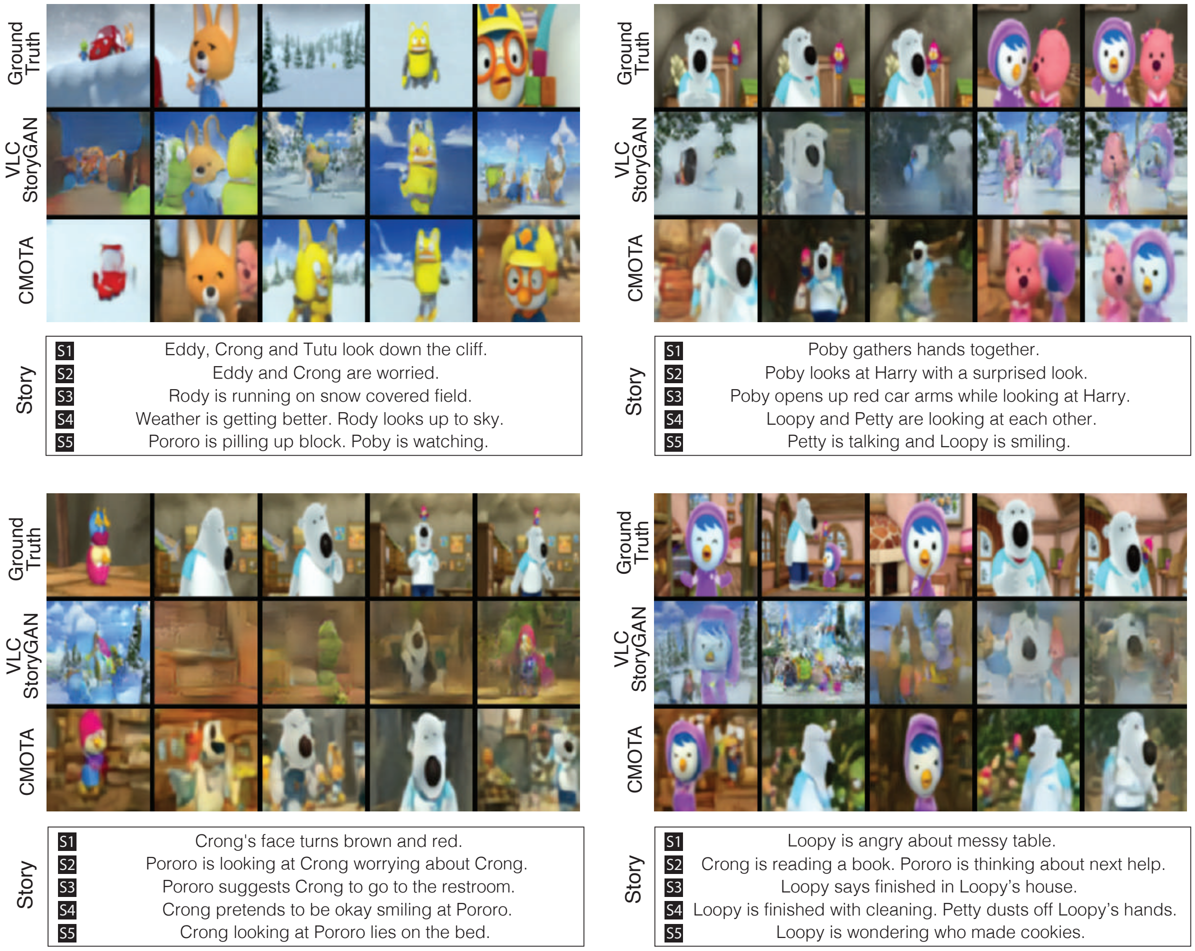
S1 Petty is tired. There are cookies on the table.
S2 Petty grabs a cookie and tastes it.
S3 Petty smiles. Petty is satisfied with the cookie.
S4 Eddy and Rody visit Pororo's house.
S5 Petty say hi to Eddy and Rody.

S1 There is a house underneath a snowy cliff.
S2 Harry and Poby is in a house.
S3 Harry is singing a song. The song is making Poby painful.
S4 Poby is grabbing his head because of the pain.
S5 Poby is getting down to talk to Harry.

S1 Crong's face turns brown and red.
S2 Pororo is looking at Crong worrying about Crong.
S3 Pororo suggests Crong to go to the restroom.
S4 Crong pretends to be okay smiling at Pororo.
S5 Crong looking at Pororo lies on the bed.

S1 Loopy is angry about messy table.
S2 Crong is reading a book. Pororo is thinking about next help.
S3 Loopy says finished in Loopy's house.
S4 Loopy is finished with cleaning. Petty dusts off Loopy's hands.
S5 Loopy is wondering who made cookies.

Figure 7. **Additional Qualitative Results on Pororo-SV Dataset.** Comparing our CMOTA's qualitative results with prior state-of-the-art model [22] and ground-truth. Our CMOTA on the third row shows a more semantically relevant image sequence compared to VLC-StoryGAN [22]. Compared to ground-truth, our CMOTA shows semantically well-aligned images with backgrounds, but [22] shows semantically less aligned images.



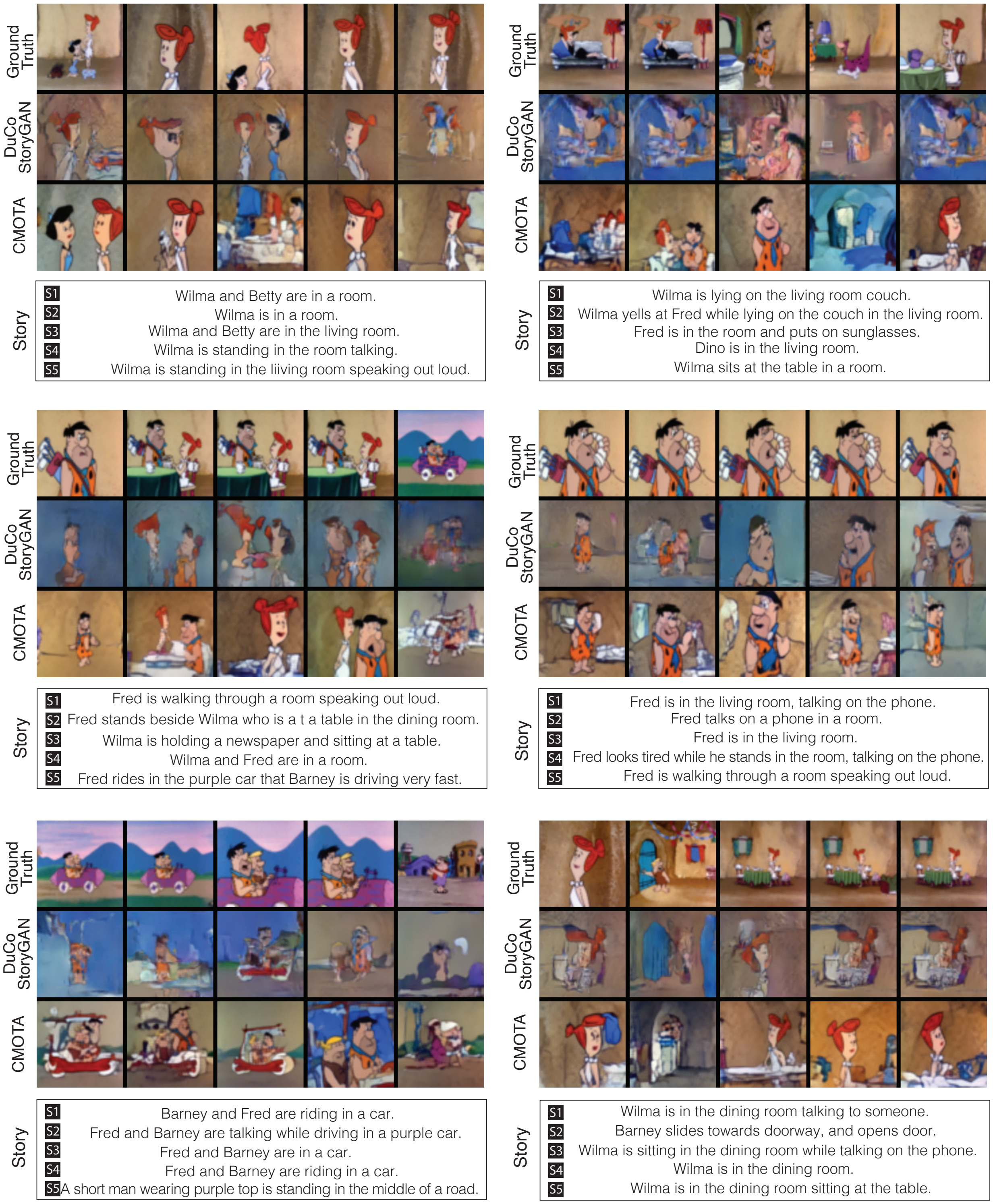
S1 Eddy, Crong and Tutu look down the cliff.
S2 Eddy and Crong are worried.
S3 Rody is running on snow covered field.
S4 Weather is getting better. Rody looks up to sky.
S5 Pororo is pilling up block. Poby is watching.

S1 Poby gathers hands together.
S2 Poby looks at Harry with a surprised look.
S3 Poby opens up red car arms while looking at Harry.
S4 Loopy and Petty are looking at each other.
S5 Petty is talking and Loopy is smiling.

S1 Crong's face turns brown and red.
S2 Pororo is looking at Crong worrying about Crong.
S3 Pororo suggests Crong to go to the restroom.
S4 Crong pretends to be okay smiling at Pororo.
S5 Crong looking at Pororo lies on the bed.

S1 Loopy is angry about messy table.
S2 Crong is reading a book. Pororo is thinking about next help.
S3 Loopy says finished in Loopy's house.
S4 Loopy is finished with cleaning. Petty dusts off Loopy's hands.
S5 Loopy is wondering who made cookies.

Figure 8. **Additional Qualitative Results on Pororo-SV dataset (Cont'd).** Comparing our CMOTA's qualitative results with prior state-of-the-art model [22] and ground-truth. Our CMOTA on the third row shows a more semantically relevant image sequence compared to VLC-StoryGAN [22]. Compared to ground-truth, our CMOTA shows semantically well-aligned images with backgrounds, but [22] shows semantically less aligned images.



S1 Wilma and Betty are in a room.
S2 Wilma is in a room.
S3 Wilma and Betty are in the living room.
S4 Wilma is standing in the room talking.
S5 Wilma is standing in the liiving room speaking out loud.

S1 Wilma is lying on the living room couch.
S2 Wilma yells at Fred while lying on the couch in the living room.
S3 Fred is in the room and puts on sunglasses.
S4 Dino is in the living room.
S5 Wilma sits at the table in a room.

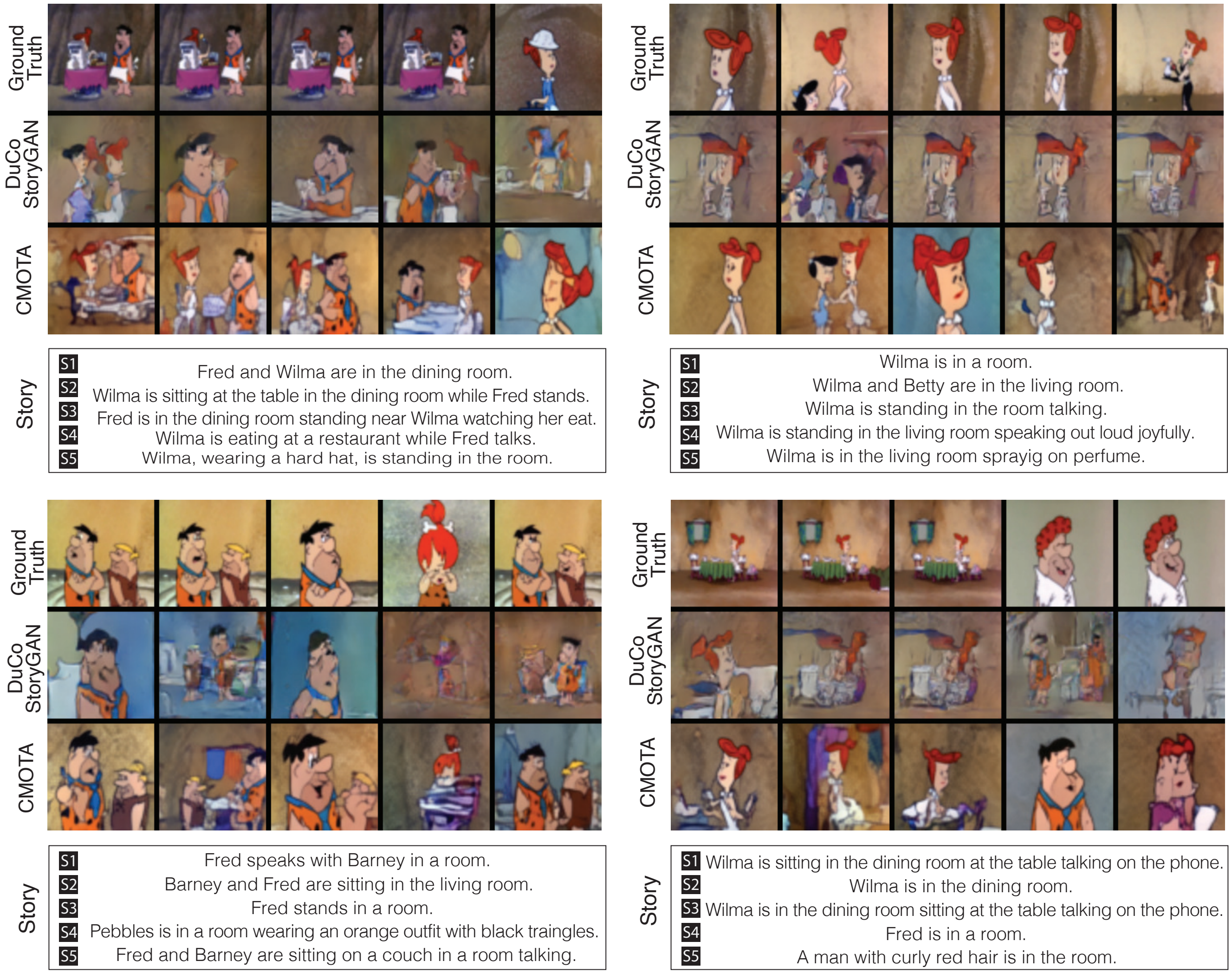
S1 Fred is walking through a room speaking out loud.
S2 Fred stands beside Wilma who is a t a table in the dining room.
S3 Wilma is holding a newspaper and sitting at a table.
S4 Wilma and Fred are in a room.
S5 Fred rides in the purple car that Barney is driving very fast.

S1 Fred is in the living room, talking on the phone.
S2 Fred talks on a phone in a room.
S3 Fred is in the living room.
S4 Fred looks tired while he stands in the room, talking on the phone.
S5 Fred is walking through a room speaking out loud.

S1 Barney and Fred are riding in a car.
S2 Fred and Barney are talking while driving in a purple car.
S3 Fred and Barney are in a car.
S4 Fred and Barney are riding in a car.
S5A short man wearing purple top is standing in the middle of a road.

S1 Wilma is in the dining room talking to someone.
S2 Barney slides towards doorway, and opens door.
S3 Wilma is sitting in the dining room while talking on the phone.
S4 Wilma is in the dining room.
S5 Wilma is in the dining room sitting at the table.

Figure 9. **Additional Qualitative Results on Flintstones-SV dataset.** Comparing our CMOTA's qualitative results with prior state-of-the-art model [23] and ground-truth. Our CMOTA on the third row shows a more semantically relevant image sequence compared to DuCo-StoryGAN [23] in the second row. Compared to ground-truth, our CMOTA shows semantically well-aligned images with backgrounds, but [22] shows semantically less aligned images.



S1 Fred and Wilma are in the dining room.
S2 Wilma is sitting at the table in the dining room while Fred stands.
S3 Fred is in the dining room standing near Wilma watching her eat.
S4 Wilma is eating at a restaurant while Fred talks.
S5 Wilma, wearing a hard hat, is standing in the room.

S1 Wilma is in a room.
S2 Wilma and Betty are in the living room.
S3 Wilma is standing in the room talking.
S4 Wilma is standing in the living room speaking out loud joyfully.
S5 Wilma is in the living room sprayig on perfume.

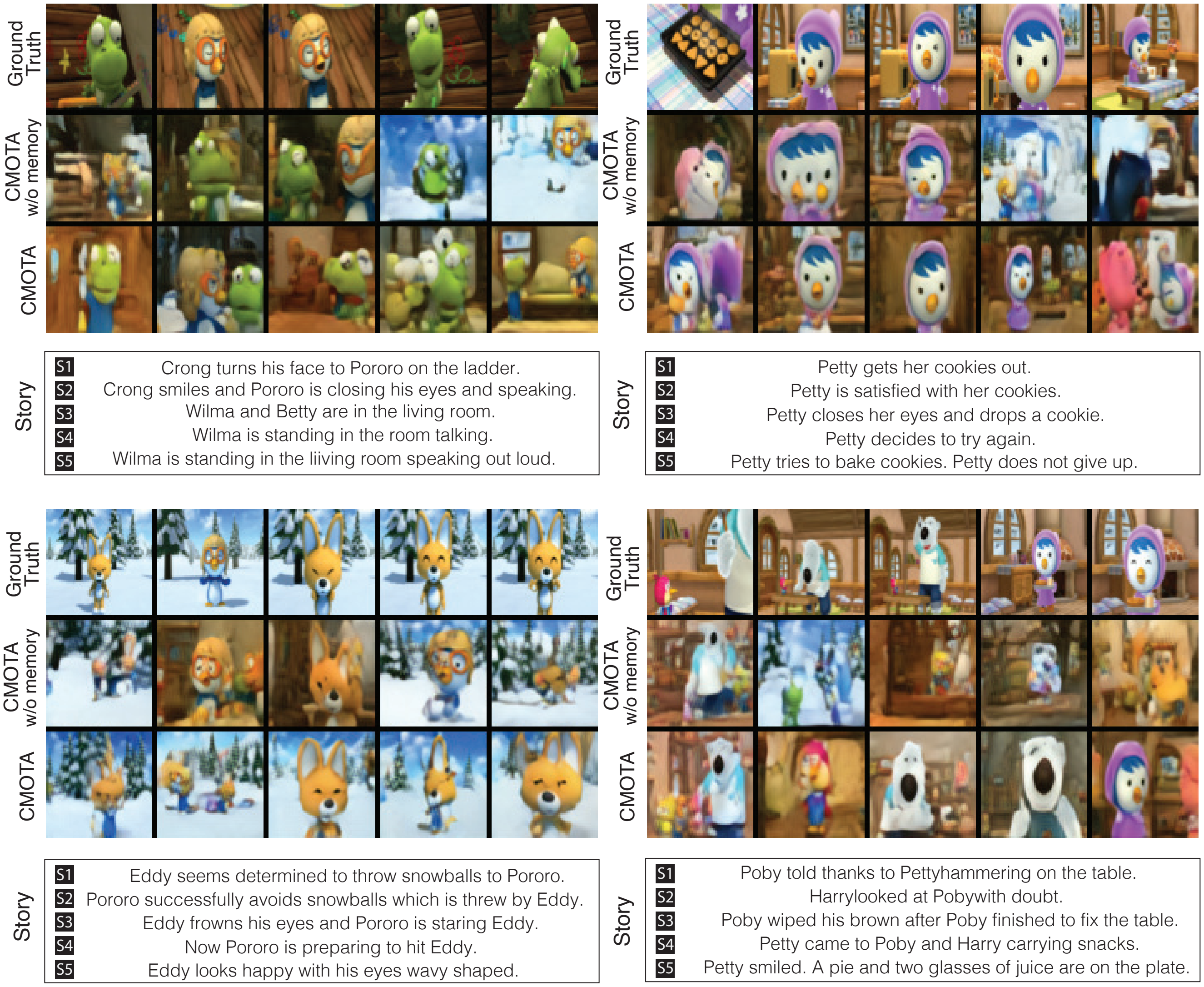
S1 Fred speaks with Barney in a room.
S2 Barney and Fred are sitting in the living room.
S3 Fred stands in a room.
S4 Pebbles is in a room wearing an orange outfit with black traingles.
S5 Fred and Barney are sitting on a couch in a room talking.

S1 Wilma is sitting in the dining room at the table talking on the phone.
S2 Wilma is in the dining room.
S3 Wilma is in the dining room sitting at the table talking on the phone.
S4 Fred is in a room.
S5 A man with curly red hair is in the room.

Figure 10. **Additional Qualitative Results on FlintstonesSV dataset (Cont'd).** Comparing our CMOTA's qualitative results with prior state-of-the-art model [23] and ground-truth. Our CMOTA on the third row shows a more semantically relevant image sequence compared to DuCo-StoryGAN [23] in the second row. Compared to ground-truth, our CMOTA shows semantically well-aligned images with backgrounds, but [23] shows semantically less aligned images.

10.1. Qualitative Results for the Efficacy of Memory Module

Figure 11 shows the efficacy of the memory module, *i.e.*, with a memory module, the proposed method (CMOTA) generates more temporally coherent images compared to the CMT-w/o-Memory, evaluated on Pororo-SV dataset.



S1 Crong turns his face to Pororo on the ladder.
S2 Crong smiles and Pororo is closing his eyes and speaking.
S3 Wilma and Betty are in the living room.
S4 Wilma is standing in the room talking.
S5 Wilma is standing in the liiving room speaking out loud.

S1 Petty gets her cookies out.
S2 Petty is satisfied with her cookies.
S3 Petty closes her eyes and drops a cookie.
S4 Petty decides to try again.
S5 Petty tries to bake cookies. Petty does not give up.

S1 Eddy seems determined to throw snowballs to Pororo.
S2 Pororo successfully avoids snowballs which is threw by Eddy.
S3 Eddy frowns his eyes and Pororo is staring Eddy.
S4 Now Pororo is preparing to hit Eddy.
S5 Eddy looks happy with his eyes wavy shaped.

S1 Poby told thanks to Pettyhammering on the table.
S2 Harrylooked at Pobywith doubt.
S3 Poby wiped his brown after Poby finished to fix the table.
S4 Petty came to Poby and Harry carrying snacks.
S5 Petty smiled. A pie and two glasses of juice are on the plate.

Figure 11. **Examples showing the efficacy of memory module for better context encoding on Pororo-SV test split dataset.** Comparing our CMOTA's qualitative results with our CMOTA without a memory module, we observe the background of scenes and the characters are more properly generated using the proposed memory.

10.2. Examples of the Generated Pseudo-Texts during the Training

As the CMOTA is a single architecture for generating text/image sequence given another modality input, *i.e*., text-to-image generation and image-to-text generation, we can immediately use CMOTA to generate multiple sentences from an image sequence, which is a task of *visual storytelling* [5]. We apply this feature to generate pseudo-texts for our cyclic data augmentation.

As we argued in the main paper (L418), we showcase the generated pseudo-texts by our method on the Pororo-SV in Fig. 6 and Figures 12 and 13. As shown in the figures, the online generation incurs richer linguistic diversity as the iteration progresses.

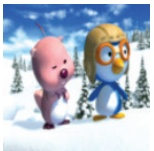
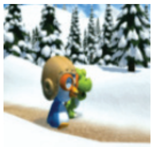
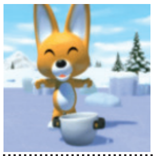
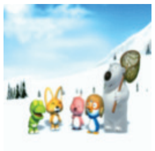
Image				
Generated Pseudo-Texts at each training iteration Training Progress	There is a tree behind Loopy. Pororo is holding a snowball in his hand.	Pororo and Crong are on the road.	Eddy opens his hands with fish toward the snow.	Poby says that Poby is lucky and waves Crong hands.
	Loopy is tilting her head.	Pororo and Crong raise their right hands and say hi.	Eddy is explaining to his friends that Eddy is going to catch all the fishes	Crong, Pororo and Loopy are standing in front of Poby.
	Loopy puts her hands on her sides. Pororo says sorry.	Pororo and Crong are walking together on the road.	Other friend sare surprised by what Eddy did, doing fishing. The sky is blue.	Eddy, Crong, Loopy, Pororo and Poby laughs all together.
	Loopy puts her hands on her sides. Pororo nods.	Pororo and Crong look in front of them and sees snow covered trees.	Eddy prepares his bowl of fishes.	Eddy, Crong, Loopy and Pororo are surprised to see Poby.
Ground Truth (Gold Label)	Loopy and Pororo are standing on the snow.	Pororo and Crong are walking together on the road.	Eddy is explaining to his friends that Eddy is going to catch all the fishes	Eddy, Crong, Loopy, Pororo and Poby laughs all together.

Figure 12. **Generated pseudo-texts by the proposed method in Pororo-SV dataset during training.** As training progresses, various pseudo-texts are generated that match with the given image on the Pororo-SV dataset.

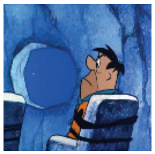
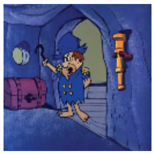
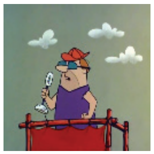
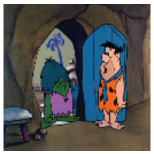
Image				
Generated Pseudo-Texts at each training iteration Training Progress	Fred is strapped in a chair in the living room at night.	A man in Napoleon costume is standing in the doorway of a room.	A man with sunglasses is in a car talking while standing.	A man with green clothes is talking in a room.
	Fred is talking as he is sitting tied at a chair in a room next to a round window.	Fred is wearing a costume with a blue shirt and cowboy hat.	The race announcer with purple, red cap and sunglasses standing in box holding microphone.	The green thing with mustache is standing in the doorway.
	Fred looks angry and is sitting in a room.	Fred is standing in the doorway.	An announcer man in purple shirt and an orange hat is outside announcing something.	The green thing with mustache is standing in the doorway.
	Fred is in a room tied up.	Fred is in the doorway.	Barney is outside.	Fred is standing in the doorway.
Ground Truth (Gold Label)	Fred sits in a chair in the room.	Fred is standing in the doorway.	An announcer man in purple shirt and an orange hat is outside announcing something.	Fred is outside at night time.

Figure 13. **Generated pseudo-texts by the proposed method in Flintstones-SV dataset during training.** As training progresses, various pseudo-texts are generated that match with the given image on the Flintstones-SV dataset.

References

[1] Samira Ebrahimi Kahou Antoine Chassang Carlo Gatta Yoshua Bengio Adriana Romero, Nicolas Ballas. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 3

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, 2021. 1

[3] Lianli Gao, Daiyuan Chen, Jingkuan Song, Xing Xu, Dongxiang Zhang, and Heng Tao Shen. Perceptual pyramid adversarial networks for text-to-image synthesis. 2019. 1

[4] Jeff Dean Geoffrey Hinton, Oriol Vinyals. Distilling the knowledge in a neural network, 2014. 3

[5] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *ACL*, 2016. 2, 14

[6] Kyung-Min Kim, Min-Oh Heo, Seongho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. *ArXiv*, abs/1707.00836, 2017. 5

[7] Hyojin Park, Youngjoon Yoo, and Nojun Kwak. Mc-gan: Multi-conditional generative adversarial network for image synthesis. In *BMVC*, 2018. 1

[8] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 3, 4

[9] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018. 3, 4

[10] Hengshuang Zhao Pengguang Chen, Shu Liu and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 3, 4

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[12] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 1

[13] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 3, 4

[14] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CVPR*, 2018. 1

[15] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *ICCV*, 2017. 1

[16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[17] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *CVPR*, 2019. 1