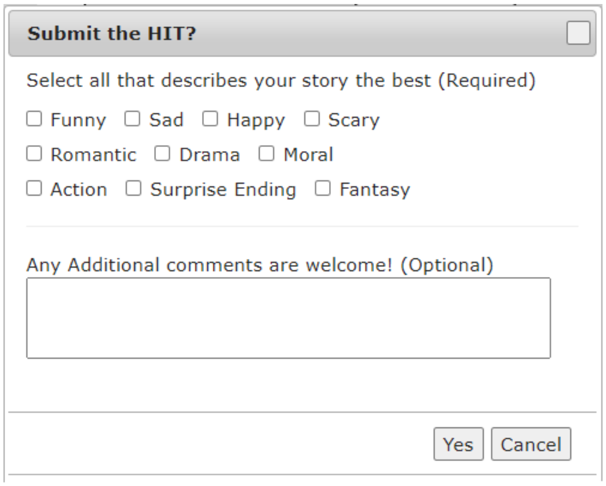
AESOP: Abstract Encoding of Stories, Objects, and Pictures

Hareesh Ravi[1] Kushal Kafle[2] Scott Cohen[2] Jonathan Brandt[2]
Mubbasir Kapadia[1]
[1]Rutgers University, [2]Adobe Research
[1]{hr268, mk1353}@cs.rutgers.edu, [2]{kkafle, scohen, jbrandt}@adobe.com

1. Data Acquisition Setup

The web interface used for data collection is shown in Fig. 3 and 4. Fig. 3 shows the top half that has the instructions given to workers and some examples of 'good' stories to inspire them. The examples are chosen randomly from a set of stories each time a worker starts a HIT. The instructions were minimal with focus on coherence, character names, non–offensive stories and creativity. The choice to not condition the story creation process on any kind of additional input results in creative and diverse stories as explained in Sec. 4 in the paper. The bottom half of the web interface is shown in Fig. 4 and displays preview of the story being created along with the title. The current panel being edited is shown below the story preview. It contains an empty canvas to the left and all clip–art objects under four categories in the right. Workers can change *scene type* to change the background, change orientation (flip) of each object and its z (ref. Sec 5.1 in main paper) value by using the slide par below the canvas. An example of 'sub types' is shown in Fig. 4 where the TV has four different types as can be seen below the canvas. The text corresponding to the current panel is written in the space provided at the bottom. Once the current visual and text parts of the story are completed, workers can continue on to the next panel. They have an option to copy the previous scene and start from that instead of from scratch for successive visual panels. The genre is asked when workers submit the story as shown in Fig. 1. They can also provide additional comments regarding the work. We received highly positive comments from workers some of which are given below:

- *This is actually my real experience!*
- *This HIT took me by surprise! Its been quite some time since I've written a story*
- *True story, happened to me*
- *I am a birder, this actually happened to me!*
- *These types of HITs are quite enjoyable. It's been a while since I've seen any similar. Hopefully there will be more available to complete, because they are unique and truly fun to work with. Thank you.*
- *Doing these hits gives a sense of satisfaction.*
- *It is a really enjoyable experience to do these hits. One of my favourite hits!*
- *This was interesting - I would love to see what others came up with*
- *Nice to do these hits after long time!*
- *Very interesting hit to do! It makes us creative.. Thanks for the opportunity.*



Submit the HIT?

Select all that describes your story the best (Required)

□ Funny □ Sad □ Happy □ Scary
□ Romantic □ Drama □ Moral
□ Action □ Surprise Ending □ Fantasy

Any Additional comments are welcome! (Optional)

Figure 1. Once the stories are written, workers are asked to select the genres for the story they wrote from a list of predefined genres. They are also asked to provide additional comments if available.

All the backgrounds and clip–art objects present in the dataset (excluding types) are shown according to their categories in Fig. 5. Objects are scaled to the same size for viewing and might blur small objects like *butterfly* and *bee*. The actual sizes of the objects and their maximum and minimum depends on what the object is and can be changed within a fixed scale using the web interface while creating

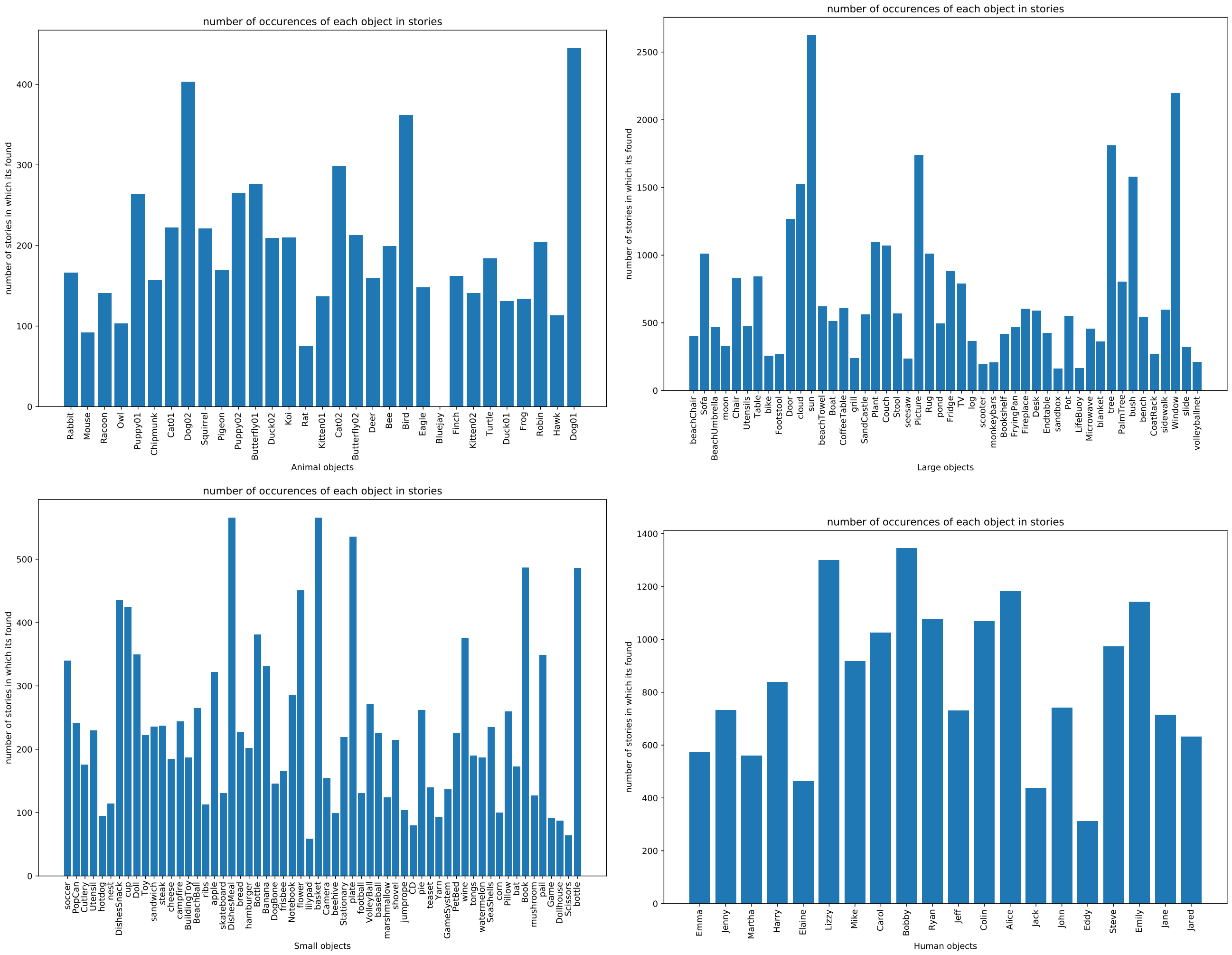


Figure 2. Distribution of objects in visual panels across the entire dataset for *Animals* (Top left), *Large objects* (Top right), *Small objects* (Bottom left) and *humans* (Bottom right).

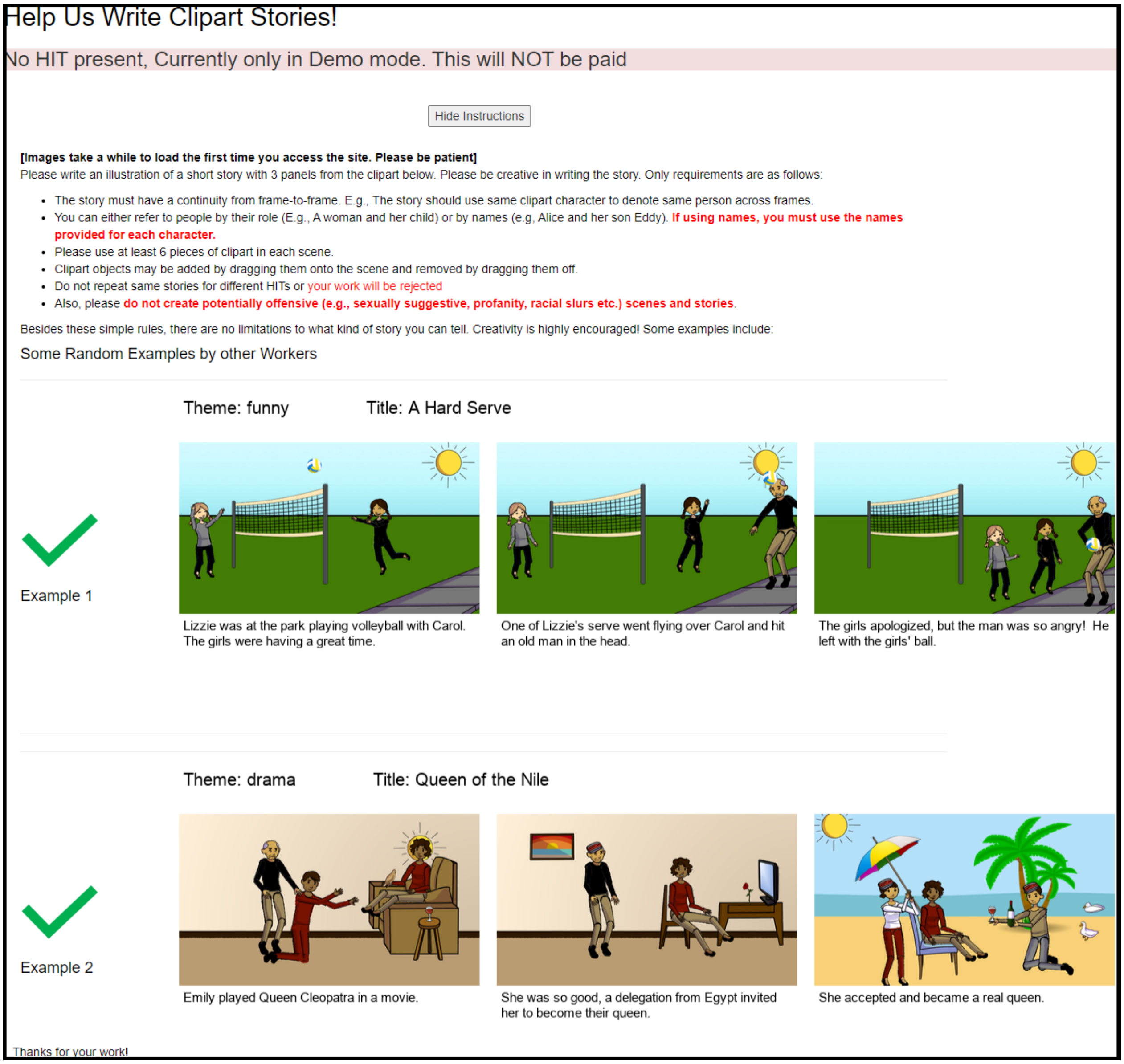
the story (see e.g. in Fig. 4). The clip–art objects that were not present in [8] are highlighted in blue including the two new backgrounds. In total we have 159 unique objects and 291 total instances of all objects that includes subtypes. Types of objects may indicate different positions, colors, perspectives etc. For example, *Fridge* and *Microwave* can be open or closed making two types, while *Window* has three colors and three shapes giving nine types in total.

1.1. Dataset Quality

We took several precautions to ensure better quality. Firstly, Annotators were required to pass an English proficiency test to access the web interface. Next, the stories completed under suspicious parameters (e.g., completed very quickly, fewer than 50 words, etc.) as well as a random sample of stories were manually reviewed using rubrics similar to the human evaluation dimensions (See main paper Sec 7). If a submission did not meet basic quality checks, we removed the workers ability to work on further stories and removed all prior submissions by that worker. All prior submissions is still accepted and paid, but they do not appear in our final dataset). The quality of the final dataset is also evidenced by comparison with existing datasets in terms of coherence and diversity (See main paper Sec 3.2).

1.2. Dataset Applicability

We believe that using clipart helps us focus more on joint understanding of narratives, text and visual component demanded by AESOP without getting bogged down by other active research areas like object detection, etc. There is also some proof [5], that knowledge from abstract images can transfer to real world.



Help Us Write Clipart Stories!

No HIT present, Currently only in Demo mode. This will NOT be paid

[Images take a while to load the first time you access the site. Please be patient]

Please write an illustration of a short story with 3 panels from the clipart below. Please be creative in writing the story. Only requirements are as follows:

- The story must have a continuity from frame-to-frame. E.g., The story should use same clipart character to denote same person across frames.
- You can either refer to people by their role (E.g., A woman and her child) or by names (e.g, Alice and her son Eddy). **If using names, you must use the names provided for each character.**
- Please use at least 6 pieces of clipart in each scene.
- Clipart objects may be added by dragging them onto the scene and removed by dragging them off.
- Do not repeat same stories for different HITs or your work will be rejected
- Also, please **do not create potentially offensive (e.g., sexually suggestive, profanity, racial slurs etc.) scenes and stories**.

Besides these simple rules, there are no limitations to what kind of story you can tell. Creativity is highly encouraged! Some examples include:

Some Random Examples by other Workers

Theme: funny Title: A Hard Serve

✓ Example 1

Lizzie was at the park playing volleyball with Carol. The girls were having a great time.

One of Lizzie's serve went flying over Carol and hit an old man in the head.

The girls apologized, but the man was so angry! He left with the girls' ball.

Theme: drama Title: Queen of the Nile

✓ Example 2

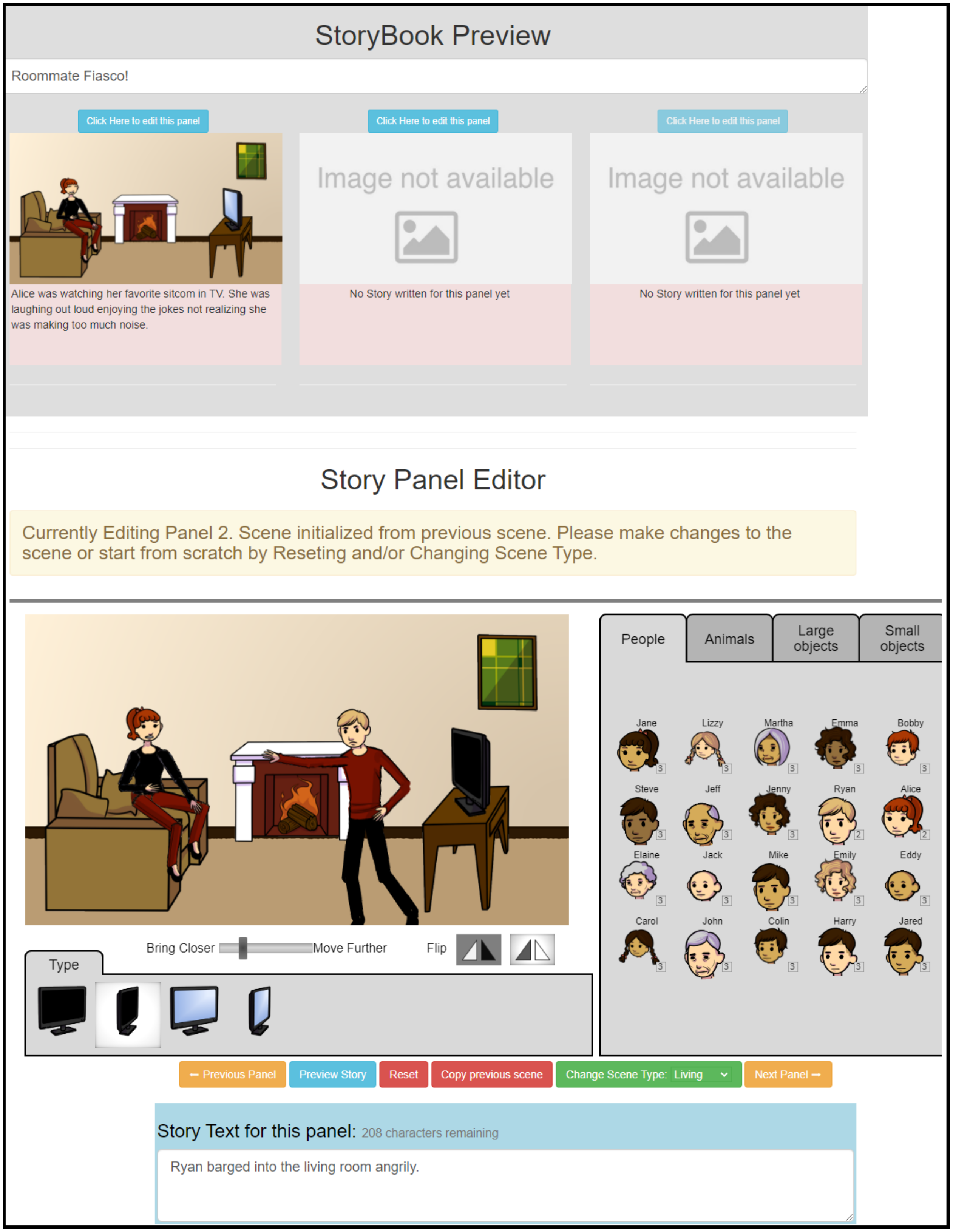
Emily played Queen Cleopatra in a movie.

She was so good, a delegation from Egypt invited her to become their queen.

She accepted and became a real queen.

Thanks for your work!

Figure 3. Top half of the data collection interface showing the instructions to workers followed by some examples of good stories. Important and necessary instructions are highlighted in red. Otherwise, constraints are limited and the workers are asked to be creative.



StoryBook Preview

Roommate Fiasco!

Alice was watching her favorite sitcom in TV. She was laughing out loud enjoying the jokes not realizing she was making too much noise.

No Story written for this panel yet

No Story written for this panel yet

Story Panel Editor

Currently Editing Panel 2. Scene initialized from previous scene. Please make changes to the scene or start from scratch by Reseting and/or Changing Scene Type.

Change Scene Type: Living

Story Text for this panel: 208 characters remaining

Ryan barged into the living room angrily.

Figure 4. Remaining part of the data collection interface showing a preview of the story in its current form (followed by canvas to create visual panel and then space to provide story text. To the right of the canvas are all the clip–arts split into categories that can be dragged to the canvas.

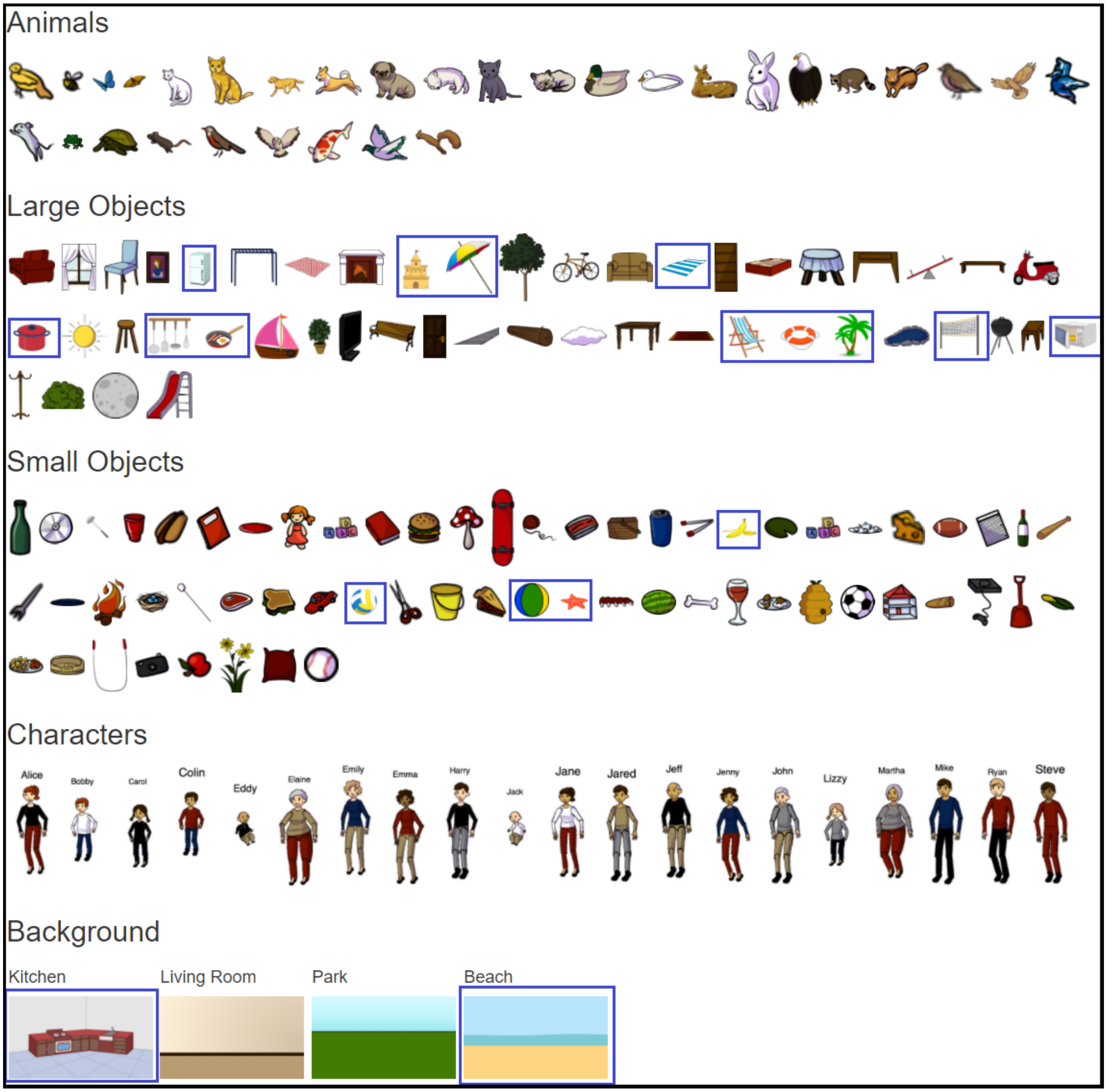


Figure 5. All the clip–art objects present in the dataset along with the four backgrounds are shown for reference. The objects are all scaled to a uniform size for display (This causes blurring of small objects like *bee*). The actual sizes depend on what the object is and is different than what is shown here.

Statistic	Per story	Per Panel
Avg # unique objects	10	8
Avg # unique humans	2.38	2.04
Avg time to create visual (in seconds)	950	–
Avg # words	84	28

Table 1. Object and Word statistics of AESOP dataset.

2. Stories in AESOP

The design principles explained in Sec. 3 in the main paper has resulted in diverse and creative stories in AESOP. More examples from the dataset are given in Fig. 6–9. For example, there are stories based on *fantasy* genre such as the third story in Fig. 6 about a dystopian future or the fourth story in Fig. 7 about a magic door. There are stories with moral such as second and fourth stories in Fig. 6 and second story in Fig. 7, or stories that emulate day–to–day activities (first story in Fig. 6 or third in Fig. 8).

There are also stories where objects are used in situations that do not usually define the characteristics of that object. For example, we have a significant number of stories based on COVID–19 where *seashells* are used as masks, *starfish* on a person's shirt for indicating they are police, *stapler* as a gun, *pond* as portal, *CD* as wheels and so on. Some examples of such stories are shown in Fig. 9. We expect that reasoning about such visuals require modelling the visual appearance such as regular pixel image features obtained from pre–trained CNNs. As part of future work, we would like to explore adding pixel information along with the abstract representation used in the models to overcome this limitation.

2.1. Dataset Statistics

Distributions of objects across stories in the dataset according to their categories is given in Fig. 2. Objects like *sun*, *cloud*, *tree*, *bush* are some of the most common and found in most of the park or beach scenes whereas *Window* is the most common object that concerns indoor scenes. Most common animal is *dog* while *Bluejay* is not used in any of the stories. Similarly, *Basket*, *DishesMeal* that has varieties of food on plates are the most common small objects. In humans *Bobby* and *Lizzy* of *kid* age group are the most common while *Jack* and *Eddy* that are toddlers are the least common.

More object and word statistics are provided in Tab. 1.

3. AESOP Model Details

We provide more details for the proposed model (defined in Sec. 5 in the main paper), below along with details on the experimental setup.

3.1. Story Encoder

The story encoder consists of a visual, text and a cross–modal encoder. The visual and textual encoders are separate Bidirectional GRUs [1], that encode modality specific coherence in the story. While the text encoder learns plausible story lines, the visual encoder learns plausible visual sequences as given in (1).

$$\begin{aligned} h_{v_i} &= Enc_{vis}(f(v_i), h_{v_{i-1}}, h_{v_{i+1}}) \\ h_{w_i} &= Enc_{text}(g(w_i), h_{w_{i-1}}, h_{w_{i+1}}) \\ h'_{v_i} &= \phi_{text}([h_{v_i}; g(word(o_i))], [h_{w_i}; g(w_i)]) \\ h'_{w_i} &= \phi_{vis}([h_{w_i}; g(w_i)], [h_{v_i}; g(word(o_i))]) \end{aligned} \tag{1}$$

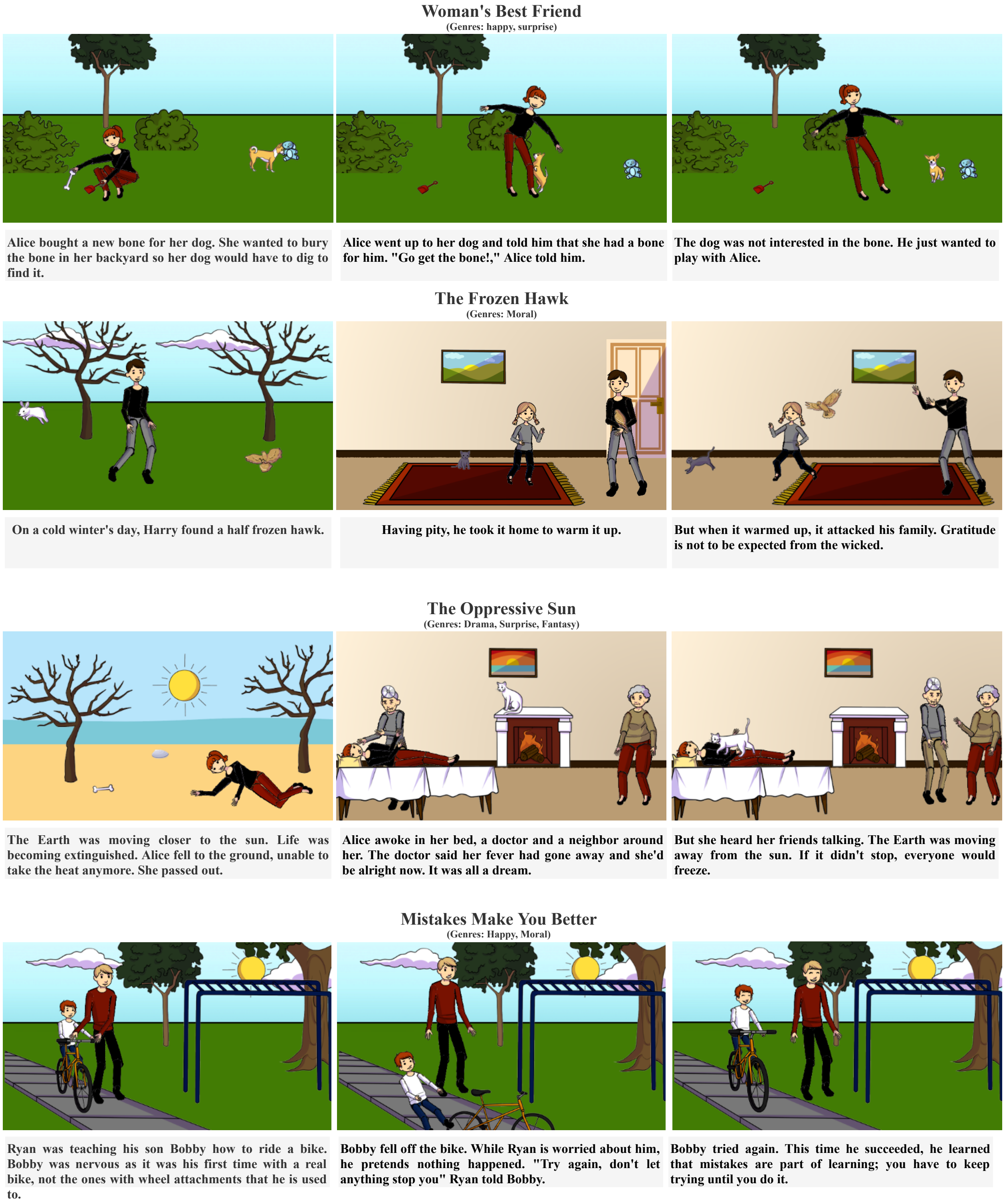
where Enc_{vis} and Enc_{text} are the BiGRUs, h'_{v_i} and h'_{w_i} are the final object and word representations. We initialize word embedding layer $g(\cdot)$ with pre–trained glove embeddings [3]. The function $\phi(\cdot)$ is the dot product attention layer similar to [6]. We concatenate the word embeddings associated with each object and word along with their respective learned representations before the attention layer. We do not provide separate names for each sub–type of the objects but use the general name to semantically ground the objects in text to avoid dependency on explicit object labels.

3.2. Visual Panel Decoder

For visual panel decoder, we use two GRUs one to track the sequence of objects and another to track the state of the visual panel. The hidden state of both the GRUs are initialized with outputs of Enc_{vis} and Enc_{text} as $[h_{v0}; h_{v_n}] + [h_{w_1}; h_{w_n}]$. At each time step, we first predict what is the next object based on the objects added to the scene so far as given in (3).

$$\begin{aligned} o_i^{state} &= Dec_{obj}(what(v_{i-1}), o_{i-1}^{state}) \\ o_i^{vis} &= \phi_{obj}^{vis}(o_i^{state}, [h'_{v_0}, ..., h'_{v_n}]) \\ o_i^{text} &= \phi_{obj}^{text}(o_i^{state}, [h'_{w_0}, ..., h'_{w_n}]) \\ o_i &= MLP(o_i^{state}, o_i^{vis}, o_i^{text}) \end{aligned} \tag{2}$$

where Dec_{obj} is a GRU that tracks the objects, $g(o_{i-1})$ is the word embedding corresponding to the previous object, o_i^{state} is the current state of the object GRU. ϕ_{obj}^{vis} and ϕ_{obj}^{text} are linear attention layers similar to [1]. ϕ_{obj}^{vis} attends to input visual panels to model visual coherence. Since story text is abstract, lot of the objects in the visual panel do not have explicit mentions in the text. Hence, to maintain coherence, it is imperative to have independent attention over other visual panels. ϕ_{obj}^{text} is required to attend to relevant text that is not visualized yet. Then, o_i^{vis} represents suggestions based on other visual panels while o_i^{text} represents suggestions based on text. Finally, the object is predicted by combining the current object GRU state and the visual and textual object suggestions. We treat the object prediction as



Woman's Best Friend

(Genres: happy, surprise)

Alice bought a new bone for her dog. She wanted to bury the bone in her backyard so her dog would have to dig to find it.

Alice went up to her dog and told him that she had a bone for him. "Go get the bone!," Alice told him.

The dog was not interested in the bone. He just wanted to play with Alice.

The Frozen Hawk

(Genres: Moral)

On a cold winter's day, Harry found a half frozen hawk.

Having pity, he took it home to warm it up.

But when it warmed up, it attacked his family. Gratitude is not to be expected from the wicked.

The Oppressive Sun

(Genres: Drama, Surprise, Fantasy)

The Earth was moving closer to the sun. Life was becoming extinguished. Alice fell to the ground, unable to take the heat anymore. She passed out.

Alice awoke in her bed, a doctor and a neighbor around her. The doctor said her fever had gone away and she'd be alright now. It was all a dream.

But she heard her friends talking. The Earth was moving away from the sun. If it didn't stop, everyone would freeze.

Mistakes Make You Better

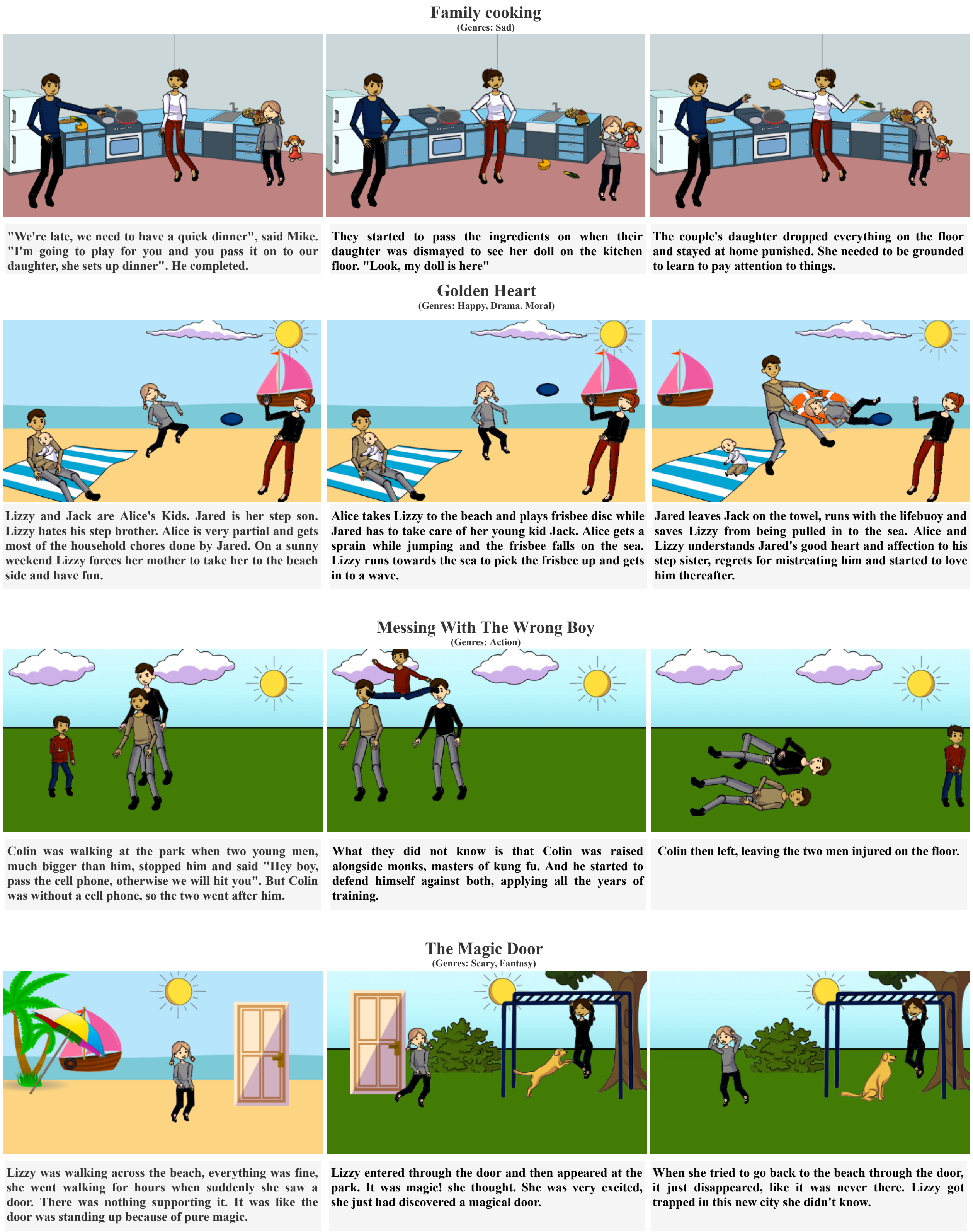
(Genres: Happy, Moral)

Ryan was teaching his son Bobby how to ride a bike. Bobby was nervous as it was his first time with a real bike, not the ones with wheel attachments that he is used to.

Bobby fell off the bike. While Ryan is worried about him, he pretends nothing happened. "Try again, don't let anything stop you" Ryan told Bobby.

Bobby tried again. This time he succeeded, he learned that mistakes are part of learning; you have to keep trying until you do it.

Figure 6. Example stories from AESOP dataset with title and genres. The changes in location, pose and expression of objects align with the events in the story. Also events in the story have clear causality and coherence with a diverse set of backgrounds, poses, scenes and story arcs.



Family cooking

(Genres: Sad)

"We're late, we need to have a quick dinner", said Mike. "I'm going to play for you and you pass it on to our daughter, she sets up dinner". He completed.

They started to pass the ingredients on when their daughter was dismayed to see her doll on the kitchen floor. "Look, my doll is here"

The couple's daughter dropped everything on the floor and stayed at home punished. She needed to be grounded to learn to pay attention to things.

Golden Heart

(Genres: Happy, Drama. Moral)

Lizzy and Jack are Alice's Kids. Jared is her step son. Lizzy hates his step brother. Alice is very partial and gets most of the household chores done by Jared. On a sunny weekend Lizzy forces her mother to take her to the beach side and have fun.

Alice takes Lizzy to the beach and plays frisbee disc while Jared has to take care of her young kid Jack. Alice gets a sprain while jumping and the frisbee falls on the sea. Lizzy runs towards the sea to pick the frisbee up and gets in to a wave.

Jared leaves Jack on the towel, runs with the lifebuoy and saves Lizzy from being pulled in to the sea. Alice and Lizzy understands Jared's good heart and affection to his step sister, regrets for mistreating him and started to love him thereafter.

Messing With The Wrong Boy

(Genres: Action)

Colin was walking at the park when two young men, much bigger than him, stopped him and said "Hey boy, pass the cell phone, otherwise we will hit you". But Colin was without a cell phone, so the two went after him.

What they did not know is that Colin was raised alongside monks, masters of kung fu. And he started to defend himself against both, applying all the years of training.

Colin then left, leaving the two men injured on the floor.

The Magic Door

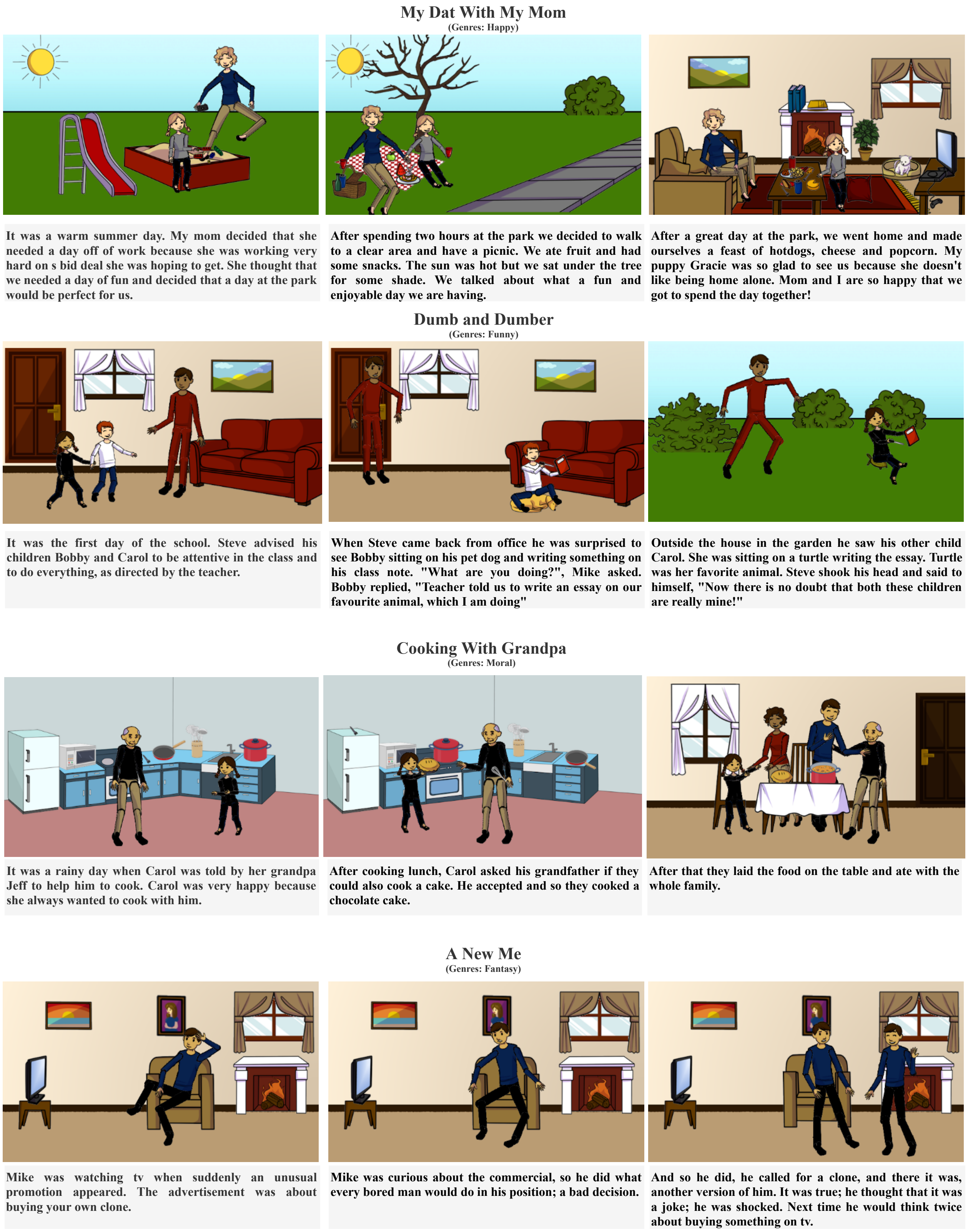
(Genres: Scary, Fantasy)

Lizzy was walking across the beach, everything was fine, she went walking for hours when suddenly she saw a door. There was nothing supporting it. It was like the door was standing up because of pure magic.

Lizzy entered through the door and then appeared at the park. It was magic! she thought. She was very excited, she just had discovered a magical door.

When she tried to go back to the beach through the door, it just disappeared, like it was never there. Lizzy got trapped in this new city she didn't know.

Figure 7. Example stories from AESOP dataset with title and genres. The short and interesting titles and the genres indicating the entire emotional arc in the story are useful for emotional perception and controlled storytelling tasks. They could also be provided as additional input to the model to condition the generation of missing panel or even entire stories.



My Dat With My Mom

(Genres: Happy)

It was a warm summer day. My mom decided that she needed a day off of work because she was working very hard on s bid deal she was hoping to get. She thought that we needed a day of fun and decided that a day at the park would be perfect for us.

After spending two hours at the park we decided to walk to a clear area and have a picnic. We ate fruit and had some snacks. The sun was hot but we sat under the tree for some shade. We talked about what a fun and enjoyable day we are having.

After a great day at the park, we went home and made ourselves a feast of hotdogs, cheese and popcorn. My puppy Gracie was so glad to see us because she doesn't like being home alone. Mom and I are so happy that we got to spend the day together!

Dumb and Dumber

(Genres: Funny)

It was the first day of the school. Steve advised his children Bobby and Carol to be attentive in the class and to do everything, as directed by the teacher.

When Steve came back from office he was surprised to see Bobby sitting on his pet dog and writing something on his class note. "What are you doing?", Mike asked. Bobby replied, "Teacher told us to write an essay on our favourite animal, which I am doing"

Outside the house in the garden he saw his other child Carol. She was sitting on a turtle writing the essay. Turtle was her favorite animal. Steve shook his head and said to himself, "Now there is no doubt that both these children are really mine!"

Cooking With Grandpa

(Genres: Moral)

It was a rainy day when Carol was told by her grandpa Jeff to help him to cook. Carol was very happy because she always wanted to cook with him.

After cooking lunch, Carol asked his grandfather if they could also cook a cake. He accepted and so they cooked a chocolate cake.

After that they laid the food on the table and ate with the whole family.

A New Me

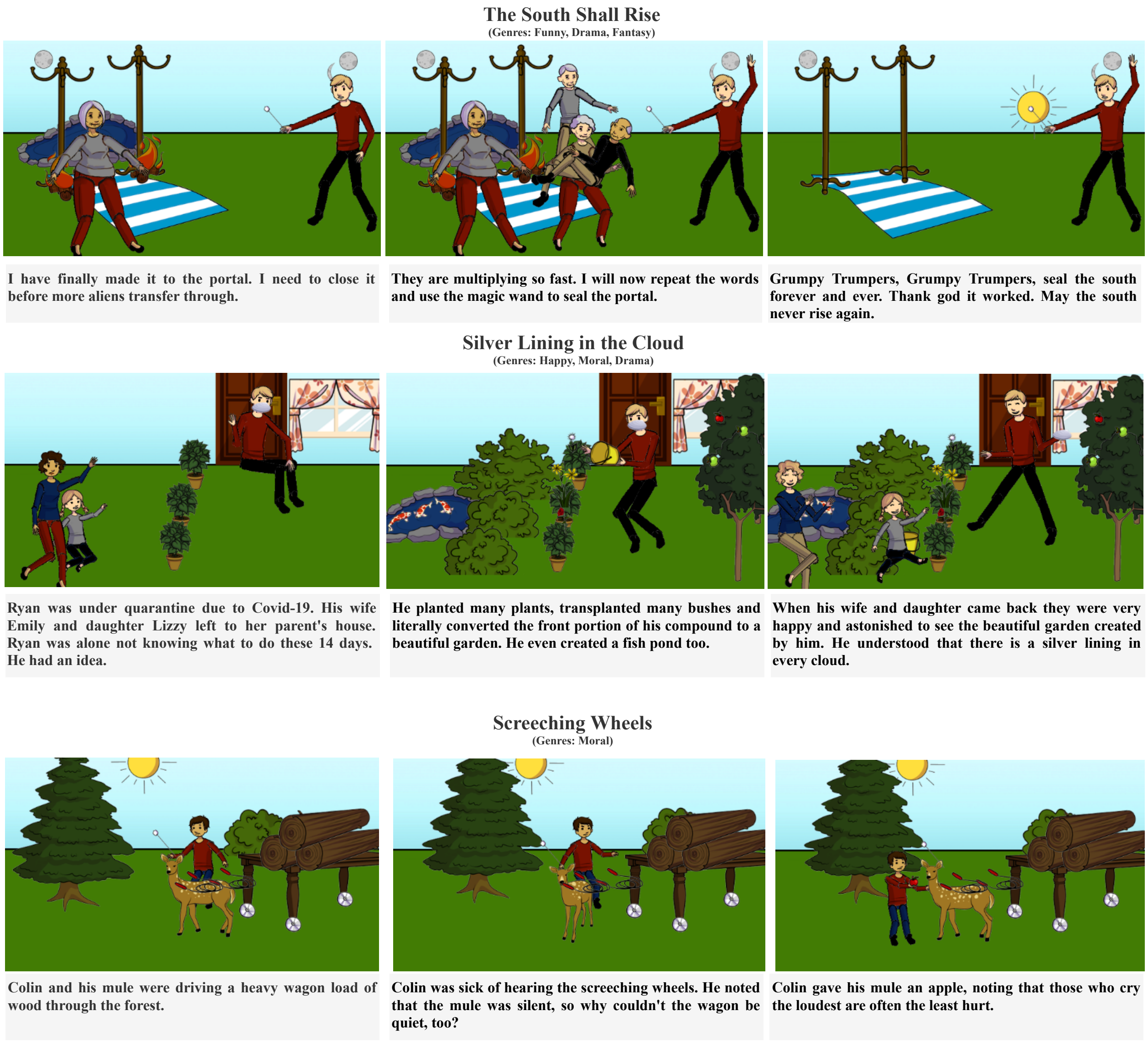
(Genres: Fantasy)

Mike was watching tv when suddenly an unusual promotion appeared. The advertisement was about buying your own clone.

Mike was curious about the commercial, so he did what every bored man would do in his position; a bad decision.

And so he did, he called for a clone, and there it was, another version of him. It was true; he thought that it was a joke; he was shocked. Next time he would think twice about buying something on tv.

Figure 8. More examples from AESOP dataset. we have stories that talk about *clones*, normal day of cooking, outing with mom and even a funny story with the title ‘Dumb and Dumber’.



The South Shall Rise

(Genres: Funny, Drama, Fantasy)

I have finally made it to the portal. I need to close it before more aliens transfer through.

They are multiplying so fast. I will now repeat the words and use the magic wand to seal the portal.

Grumpy Trumpers, Grumpy Trumpers, seal the south forever and ever. Thank god it worked. May the south never rise again.

Silver Lining in the Cloud

(Genres: Happy, Moral, Drama)

Ryan was under quarantine due to Covid-19. His wife Emily and daughter Lizzy left to her parent's house. Ryan was alone not knowing what to do these 14 days. He had an idea.

He planted many plants, transplanted many bushes and literally converted the front portion of his compound to a beautiful garden. He even created a fish pond too.

When his wife and daughter came back they were very happy and astonished to see the beautiful garden created by him. He understood that there is a silver lining in every cloud.

Screeching Wheels

(Genres: Moral)

Colin and his mule were driving a heavy wagon load of wood through the forest.

Colin was sick of hearing the screeching wheels. He noted that the mule was silent, so why couldn't the wagon be quiet, too?

Colin gave his mule an apple, noting that those who cry the loudest are often the least hurt.

Figure 9. More examples from AESOP dataset where objects are used creatively for purposes that does not usually define the object such as CD for wheels, seashell for masks and so on.

a classification over the entire object vocabulary. Next, we decode the attributes of the predicted object as follows.

$$\begin{aligned}
V_i^{state} =& Dec_{attr}(what(v_{i-1}) + where(v_{i-1}) + how(v_{i-1}), \\
& V_{i-1}^{state}) \\
& V_{state} = [V_i^{state}; what(v_i)] \\
& attr_i^{vis} = \phi_{attr}^{vis}(V_i^{state}, h'_{v_j}) \\
& attr_i^{text} = \phi_{attr}^{text}(V_i^{state}, h'_{w_j}) \\
& attr_i = MLP(V_i^{state}, attr_i^{vis}, attr_i^{text})
\end{aligned} \tag{3}$$

where I_{state} is the current state of the visual panel with all the objects and their attributes so far. We combine the representation of the predicted object o_i with the scene state as query for visual and textual attention. The visual and textual attention modules ϕ_{attr}^{vis} and ϕ_{attr}^{text} suggest possible set of attributes for the predicted object based on the current scene state and input panels. Finally, the attributes are predicted as a single 33–dim vector, 4 for x_i, y_i, z_i and $flip_i$, 20 for poses and 9 for expressions. The dimensions corresponding to *where* attributes are clamped to be between 0 and 1 while *softmax* function is applied for pose and expression classification.

3.3. Training Details

Out of 7062 stories we use 5562 for training, 500 for validation and test on the remaining 1000 stories. All human evaluation experiments and human baseline models are run on a subset of the test set containing 500 stories. The hidden dimensions of the encoder and decoder are 512 and the visual and text tokens have an output dimension of 1024 (including the word embeddings). Maximum number of words considered per story is 150, 50 per text panel, while the maximum number of objects is set at 45, 15 per panel. The maximum excludes special tokens such as $\langle$MASK$\rangle$, $\langle$SEP$\rangle$, $\langle$SOS$\rangle$ and $\langle$EOS$\rangle$ separately for visual and text. There is also $\langle$SOG$\rangle$ and $\langle$EOG$\rangle$ indicating start and end of generation for the decoders. We use the Adam optimizer [2], initialized with a learning rate of $3.5e$-4 and a decay rate of 0.8 whenever scene similarity metric plateaus with patience of 8 epochs. We train all the models for 80 epochs on the training set and the epoch with the highest Scene Similarity metric on validation data is chosen as the best epoch for evaluation on the test set.

During inference, metrics are calculated over the entire test set. We choose learning rate of $3.5e$-4 was empirically chosen from {$1e$-4, 1.5e-4, 2e-4, ..., 5e-2}, hidden dimension of 512 for the networks from {128, 256, 512, 768, 1024} and batch size from {8, 16, 24, 32, 48, 64}. The word vocabulary size of the text encoder decoders is 11,158 while the object vocabulary is 291 + 10 for special tokens and backgrounds.

4. Additional Tasks

Additional to the tasks defined in Sec. 4 in the main paper that masks the last visual or text panel, we also train and test by masking the middle visual or text panels. This requires considering the context of both past and future inputs, and may be used as a surrogate task to represent nonlinear nature of story creation. The quantitative results are given in Tab. 2. As emphasizes in Sec. 8 in the main paper, these metrics are unreliable with similar values as seen for the last panel generation. We observe that the values for the metrics for middle visual panel generation are higher for all the models when compared to the scores for last panel generation (ref. Tab.3 in main paper). This is because, there is much less surprise in the middle visual panels compared to the last ones. More stories have different backgrounds and new objects in the last panel while the middle panel contains minimal changes. If stories introduce new characters in the middle panel those mostly do not leave the story in the last panel, requiring models to learn which of the available two panels (first or last) to copy from to beat the metrics. Note that pose, expression and other changes still do exist but the metrics give an overall notion of similarity to ground truth and since most objects do not change significantly, learning to replicate the other panels can lead to higher scores as can be observed from the scores of 'Repeat' baselines. Repeat baseline again outperforms on most of the metrics. Human baseline begins to close the gap on these quantitative metrics because there are less variations in the human created scenes as well for the middle panel. Note that we do collect human baselines for middle panel generation as well but do not perform user study to evaluate the performance of the models.

On the 'text' side, the results are similar to what we observed in the last panel completion. The proposed model has similar BLEU and ROUGE-L scores indicating it is able to capture the contents present in the corresponding visual panel to some extent while it struggles to form grammatically correct and coherent sentences. We observe these through examples as well (Fig. 13), and we anticipate similar results as last panel generation if we run human studies. GPT2 on the other hand scores slightly higher on the middle panel completion across all metrics. Human baselines outperform across all metrics for middle panel text generation.

We also tried a single model trained to generate any missing panel but given the bias towards minimum changes in the middle panel and the last panel for most stories, the model ends up learning to replicate the 'Repeat' baseline. This further motivates the need for a model that encodes only change of objects and attributes in the visual side. This would also ensure better alignment with text as the text has explicit mentions of only what changes throughout a story.

Model↑		**BG** ↑	**O-IOU** ↑	**Loc** ↑	**Dep** ↑	**Flip** ↑	**Pose** ↑	**Expr** ↑	**Scene**↑		**B–1**↑	**B–4** ↑	**M** ↑	**R–L** ↑	**C** ↑
Proposed	Illustrator	92.0	68.5	0.77	94.2	90.8	32.6	43.2	4.3	Writer	25.6	2.5	8.8	22.5	17.6
Unimodal		91.5	71.7	0.77	94.1	90.9	31.9	44.0	4.6		8.2	0.36	6.2	11.1	5.1
One-to-one		69.0	18.5	0.40	42.5	32.4	8.1	7.72	1.4		22.6	0.95	7.0	12.7	8.1
Pixel		55.4	15.9	0.28	21.4	12.1	4.8	6.12	1.3		11.2	0.85	7.9	18.2	16.1
Human		95.8	79.5	0.90	73.5	71.1	25.1	31.3	4.4		28.4	4.9	12.5	28.3	20.1
Repeat		91.4	79.4	0.94	95.3	92.4	41.4	52.5	5.3		–	–	–	–	–

Table 2. Results of all models on Assistant Illustrator and Assistant Writer modes when the *middle* panel is masked. For Assistant Illustrator, we provide accuracy over entire test set for prediction of **BG** (background) **Dep** (z value), **Flip**, **Pose** and **Expr** (Expression). **Loc** is the location similarity while **O-IOU** is the intersection over union between predicted and ground truth set of objects. Metrics for object attributes are calculated only if the predicted object is present in ground truth. **Scene** is the scene similarity metric. For Assistant Writer mode, **B–1** indicates BLEU–1, **B–4** is BLEU–4, **M** is METEOR, **R–L** is ROUGE–L and **C** is CIDEr.

5. Additional Results and Examples

The results of user study for the Assistant Writer experiment described in the main paper for generation of the last panel's text is given in Tab. 3.

Experiment	**Meaningful**	**Relevant**	**Coherent**	**Overall**
Human	96.2	91.6	96.2	96.4
Proposed	2.4	2.0	2.6	2.6
Both	1.4	6.4	1.2	1.0
Unimodal	80.6	59.0	64.8	70.2
Proposed	10.8	18.8	17.6	15.0
Both	8.6	22.2	17.6	14.8

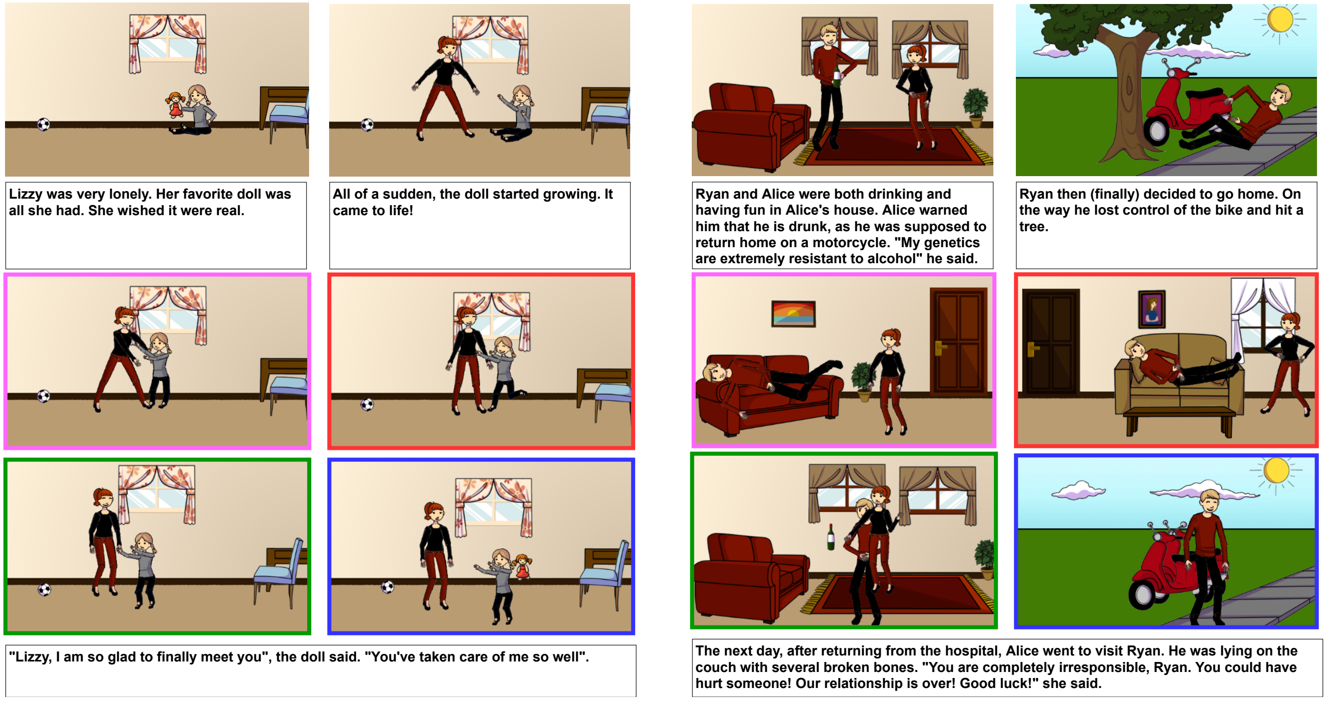
Table 3. Results of user study comparing models pairwise along three dimensions for Assistant Writer task. Values are given in % and *overall* indicates the overall preference between the two shown models. *Both* indicated equal preference.

Fig. 10 and 11 show more examples of last visual panel generation by the proposed model compared with human, ground truth and unimodal baselines where the proposed model performed relatively better than baselines as rated in the user study. For example, in Fig. 11 (left) the proposed and human baselines got same relevance and meaningfulness score but human baseline won on coherence giving the overall score to the human baseline. In the proposed vs unimodal comparison experiment, the proposed model won on all fronts for this story. We can see that the model is able to learn cross–modal relevance and visual coherence. In all the examples the first two panels are given at the top and the last panel for all the baselines are given below them with colored borders indicating the model type. Each figure also gives a brief analysis on the generations. In Fig. 12 we show some failure cases of the proposed model. In most of the failure cases, the model is unable to figure out the changes and ends up replicating the previous panel or tries to change but is unable to completely create a new unseen scene. Based on these results, we highlighted next steps for better models in Sec. 9 in the main paper. Additional examples for the Assistant Writer task for the last panel are given in Fig. 13 and 14.

5.1. Significance Numbers for Human Experiments

We calculate confidence intervals for all our pairwise human evaluation experiments using Wilson's [7] method. We ignore all samples that were undecided and only consider samples that were given a definite vote.

In the Assistant Illustrator experiments provided in Table 3 in the main paper, our **Proposed** model was found to be statistically significant compared to **Unimodal**. 43.71% preferred Unimodal and 56.29% preferred Proposed with a 95% confidence interval of ± 5.45.
Human performance obtained 91.9% votes while the **Proposed** model obtained 8.1% votes in another experiment. Human performance was statistically significant with a 95% confidence interval of ± 2.46.
Similarly, **Human** baseline obtained 94.3% votes when compared with **Unimodal** that obtained 5.6% with a statistical significance with 95% confidence interval of ± 2.15.
We can also observe from these numbers that **Proposed** is better than the **Repeat** baseline, in contrast to the numbers shown in Table 2 in the main paper.
We also observe similar statistical significance in the Assistant writer experiments as well.



Lizzy was very lonely. Her favorite doll was all she had. She wished it were real.

All of a sudden, the doll started growing. It came to life!

"Lizzy, I am so glad to finally meet you", the doll said. "You've taken care of me so well".

Ryan and Alice were both drinking and having fun in Alice's house. Alice warned him that he is drunk, as he was supposed to return home on a motorcycle. "My genetics are extremely resistant to alcohol" he said.

Ryan then (finally) decided to go home. On the way he lost control of the bike and hit a tree.

The next day, after returning from the hospital, Alice went to visit Ryan. He was lying on the couch with several broken bones. "You are completely irresponsible, Ryan. You could have hurt someone! Our relationship is over! Good luck!" she said.

Figure 10. Examples of Assistant Illustrator for last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model successfully generated relevant and coherent visual panels.

Analysis:

Left: The generated visual panel removes the doll, retains the woman and changes her expression correctly. Note that the human baseline and ground truth do not show clustered poses and hence looks more realistic while the generated visual panel has predicted the closest pose from the 20 possible poses (during training ground truth and input poses are clustered but general poses are shown here to retain the realism in data). Unimodal as expected does not know to remove the doll or change expressions indicating that the proposed model takes text into account.

Right: The story starts at Alice's house but ends at Ryan's place. The ground truth story does not have a change of scene in the third panel. The human baseline however, correctly captures the change of scene. In the proposed model's generation, the scene change from park to a house, Alice's presence and her angry expression are all captured perfectly when compared with ground truth. However, the model misses that Ryan is lying on the couch.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*, 2015. 6

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 11

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 6

[4] Gorjan Radevski, Guillem Collell, Marie-Francine Moens, and Tinne Tuytelaars. Decoding language spatial relations to 2D spatial arrangements. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4549–4560, Online, Nov. 2020. Association for Computational Linguistics. 15

[5] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6710–6719, 2019. 2

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6

[7] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. 12

[8] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016. 2

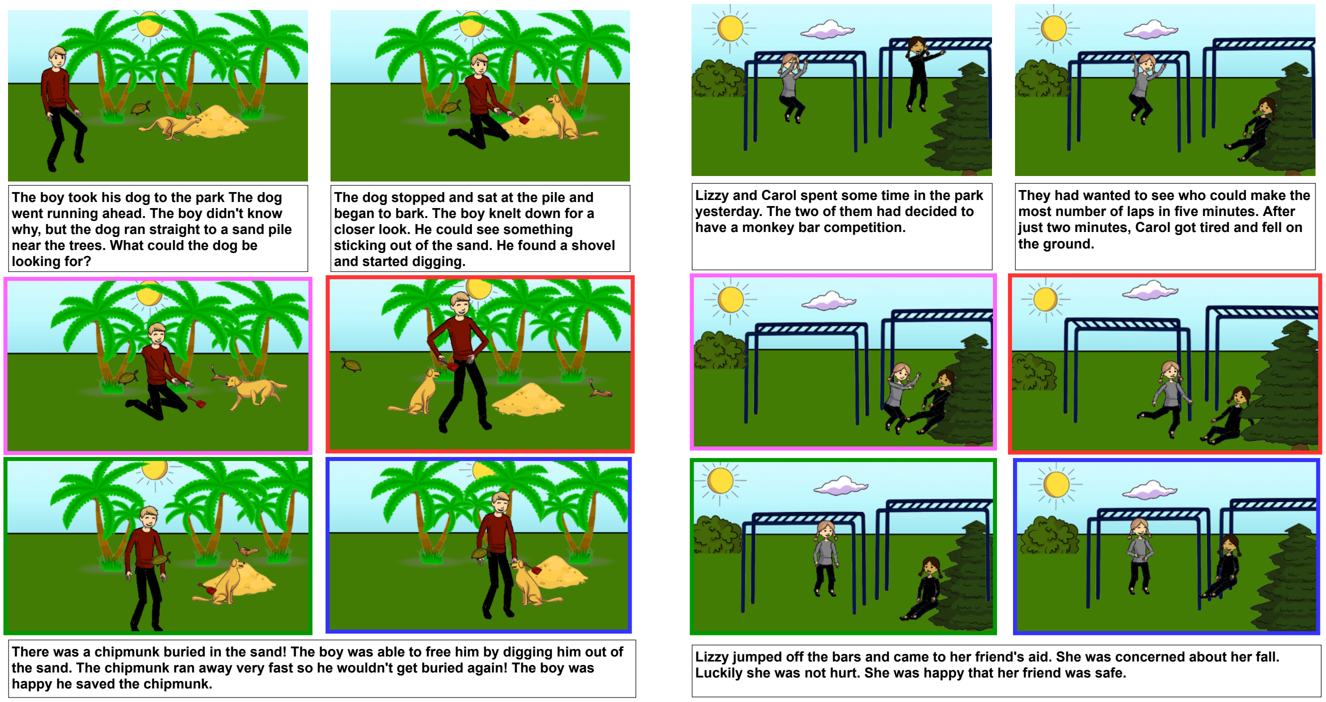
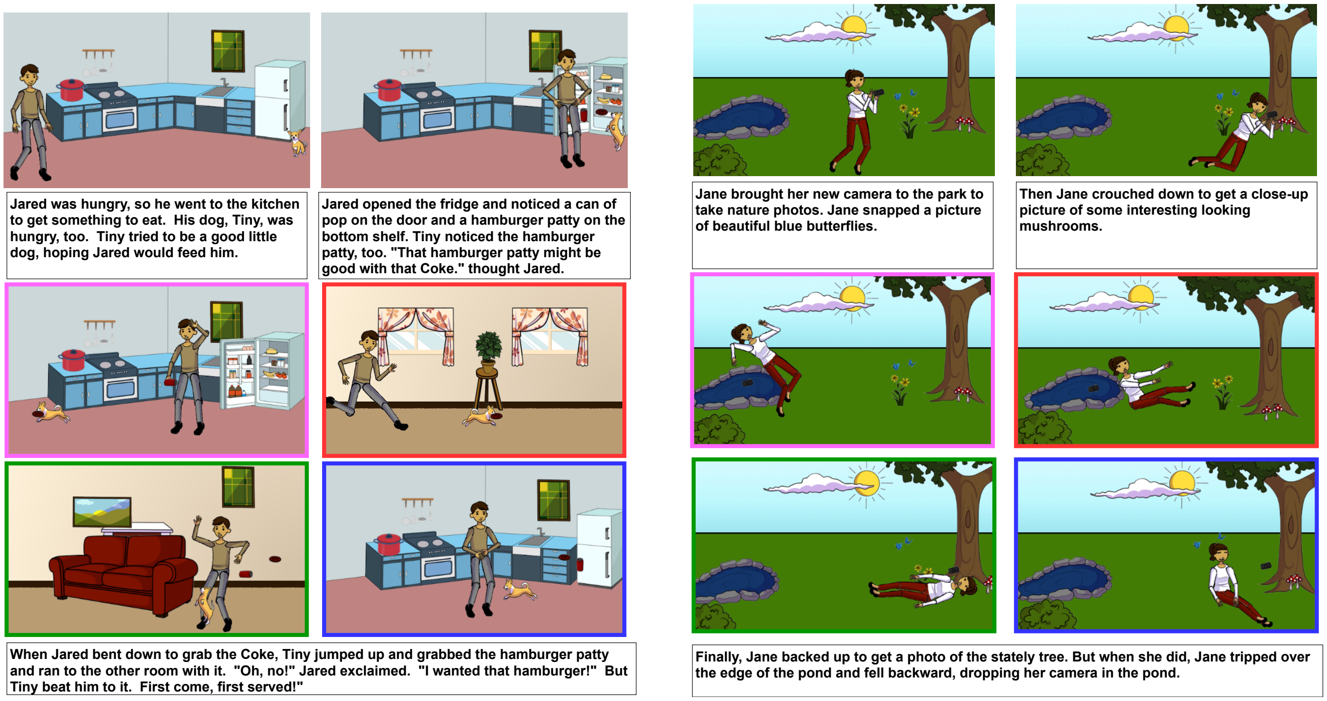

The boy took his dog to the park The dog went running ahead. The boy didn't know why, but the dog ran straight to a sand pile near the trees. What could the dog be looking for?
The dog stopped and sat at the pile and began to bark. The boy knelt down for a closer look. He could see something sticking out of the sand. He found a shovel and started digging.
There was a chipmunk buried in the sand! The boy was able to free him by digging him out of the sand. The chipmunk ran away very fast so he wouldn't get buried again! The boy was happy he saved the chipmunk.
Lizzy and Carol spent some time in the park yesterday. The two of them had decided to have a monkey bar competition.
They had wanted to see who could make the most number of laps in five minutes. After just two minutes, Carol got tired and fell on the ground.
Lizzy jumped off the bars and came to her friend's aid. She was concerned about her fall. Luckily she was not hurt. She was happy that her friend was safe.

Figure 11. Examples of Assistant Illustrator for last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model successfully generated relevant and coherent visual panels.

Analysis:

Left: The generated visual panel is close to ground truth as well as human baselines. It retains the sand, but relieves the chipmunk and has similar pose and orientation to human baseline for Ryan and the dog. Except for the turtle that seems to be on top of Ryan, the overall scene is relevant to the text, coherent to previous visual panels and depicts a meaningful scene.

Right: The proposed model captures the change in expressions and relative locations correctly similar to ground truth or human baseline making for a reasonable illustration of the story. Moreover the model has learned to not vary the position of objects that are still such as the monkey–bars, bush, and tree.



Jared was hungry, so he went to the kitchen to get something to eat. His dog, Tiny, was hungry, too. Tiny tried to be a good little dog, hoping Jared would feed him.

Jared opened the fridge and noticed a can of pop on the door and a hamburger patty on the bottom shelf. Tiny noticed the hamburger patty, too. "That hamburger patty might be good with that Coke." thought Jared.

When Jared bent down to grab the Coke, Tiny jumped up and grabbed the hamburger patty and ran to the other room with it. "Oh, no!" Jared exclaimed. "I wanted that hamburger!" But Tiny beat him to it. First come, first served!"

Jane brought her new camera to the park to take nature photos. Jane snapped a picture of beautiful blue butterflies.

Then Jane crouched down to get a close-up picture of some interesting looking mushrooms.

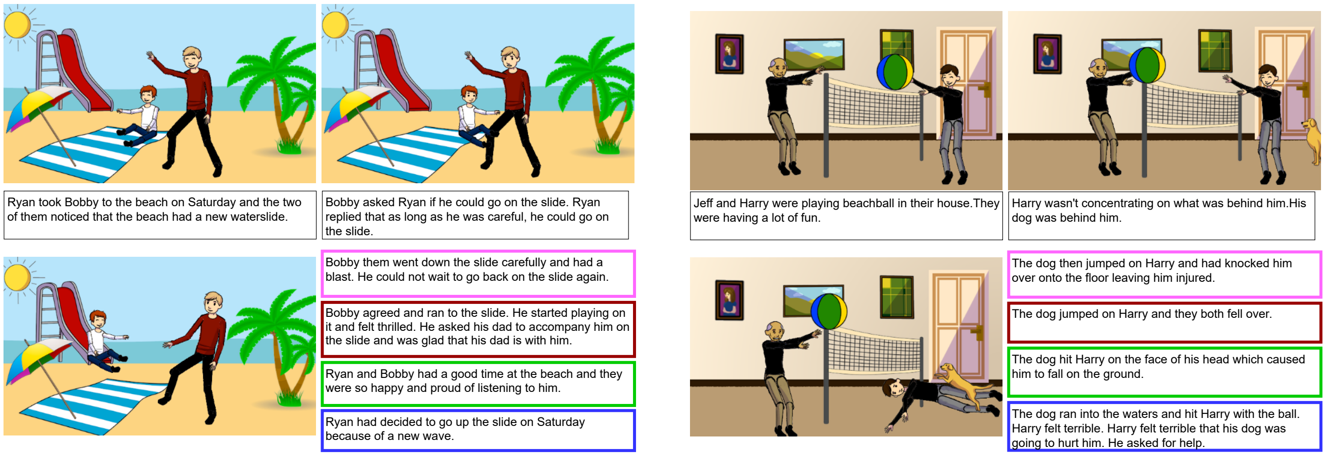
Finally, Jane backed up to get a photo of the stately tree. But when she did, Jane tripped over the edge of the pond and fell backward, dropping her camera in the pond.

Figure 12. Examples of Assistant Illustrator for last panel generation result by Ground truth, Human Baseline, Proposed model and Uni-modal where the proposed model failed to generate relevant and coherent visual panels.

Analysis:

Left: The ground truth illustration for the last panel visualizes the Jared and dog before the dog goes to the other room while the human baseline visualizes them in another room. The proposed model also takes them to another room but lack of any details on what the room is in the story makes it difficult for the model to place them in reasonable locations. However, the model still got all the relevant objects such as couch and fireplace for the living room and dog, Jared, hamburger and the soda for the story.

Right: The model illustrates falling down but did not capture where exactly Jane falls down which should have been near the pool. Predicting each of the attributes is a separate task in itself (e.g. spatial reasoning in abstract scenes formulated as a separate task in [4]), making the overall task complex.



Ryan took Bobby to the beach on Saturday and the two of them noticed that the beach had a new waterslide.

Bobby asked Ryan if he could go on the slide. Ryan replied that as long as he was careful, he could go on the slide.

Bobby them went down the slide carefully and had a blast. He could not wait to go back on the slide again.

Bobby agreed and ran to the slide. He started playing on it and felt thrilled. He asked his dad to accompany him on the slide and was glad that his dad is with him.

Ryan and Bobby had a good time at the beach and they were so happy and proud of listening to him.

Ryan had decided to go up the slide on Saturday because of a new wave.

Jeff and Harry were playing beachball in their house.They were having a lot of fun.

Harry wasn't concentrating on what was behind him.His dog was behind him.

The dog then jumped on Harry and had knocked him over onto the floor leaving him injured.

The dog jumped on Harry and they both fell over.

The dog hit Harry on the face of his head which caused him to fall on the ground.

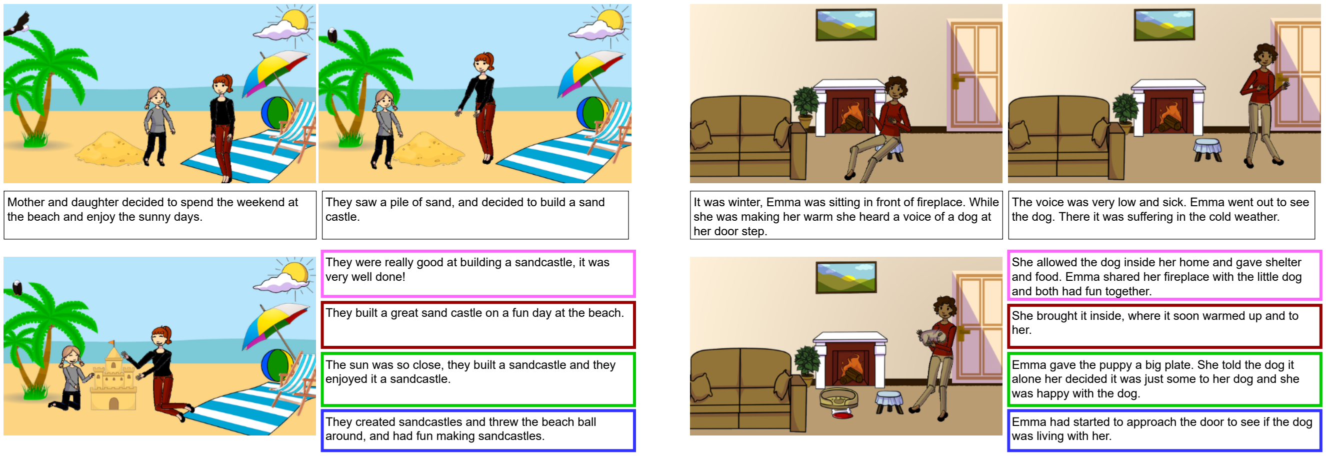
The dog ran into the waters and hit Harry with the ball. Harry felt terrible. Harry felt terrible that his dog was going to hurt him. He asked for help.

Figure 13. Examples of Assistant Writer for last last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model successfully generated reasonable text to complete the story.

Analysis:

Left: The proposed model scored equally for relevance and meaningfulness against human baseline while the human baseline won against coherence. The proposed model's generation won in all metrics against the unimodal GPT2 model. The generated text is a reasonable ending for the current story.

Right: In this example the generated text by the proposed model captures the content with high relevance and coherence to the rest of the story achieving higher scores in the human evaluation.



Mother and daughter decided to spend the weekend at the beach and enjoy the sunny days.

They saw a pile of sand, and decided to build a sand castle.

They were really good at building a sandcastle, it was very well done!

They built a great sand castle on a fun day at the beach.

The sun was so close, they built a sandcastle and they enjoyed it a sandcastle.

They created sandcastles and threw the beach ball around, and had fun making sandcastles.

It was winter, Emma was sitting in front of fireplace. While she was making her warm she heard a voice of a dog at her door step.

The voice was very low and sick. Emma went out to see the dog. There it was suffering in the cold weather.

She allowed the dog inside her home and gave shelter and food. Emma shared her fireplace with the little dog and both had fun together.

She brought it inside, where it soon warmed up and to her.

Emma gave the puppy a big plate. She told the dog it alone her decided it was just some to her dog and she was happy with the dog.

Emma had started to approach the door to see if the dog was living with her.

Figure 14. Examples of Assistant Writer result for last panel generation by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model generated irrelevant or incoherent text to end the story.

Analysis:

Left: The proposed model's generation shows how it loses on coherence while trying to be relevant to the corresponding visual panel. GPT–2 generates much more coherent text and keeping with the context of the other text panels makes it preferable.

Right: This is another example of incoherent text generated by the proposed model. while the objects are perceived as we would like, the text is not comprehensible. Initializing the text parts of the model with pre–trained language models as pointed in future work would help overcome this limitation.