6 Appendix

6.1 Formula derivation

This subsection provides an introduction to the derivation process of Formulas 1, 3 , 4, and 5 in the main text.

First is Formula 1. According to Bayes rule, we can obtain:

$$P(y_t|s,x,y_{1:t-1}) = \frac{P(s,x,y_{1:t-1},y_t)}{P(s,x,y_{1:t-1})} \tag{9}$$

$$P(s|x,y_{1:t-1},y_t) = \frac{P(s,x,y_{1:t-1},y_t)}{P(x,y_{1:t-1},y_t)} \tag{10}$$

$$P(y_t|x,y_{1:t-1}) = \frac{P(x,y_{1:t-1},y_t)}{P(x,y_{1:t-1})} \tag{11}$$

$$P(s|x,y_{1:t-1}) = \frac{P(s,x,y_{1:t-1})}{P(x,y_{1:t-1})}. \tag{12}$$

Then, the probability distribution $P(y_t|s,x,y_{1:t-1})$ can be transformed into:

$$P(y_t|s,x,y_{1:t-1}) = \frac{P(s|x,y_{1:t-1},y_t)P(y_t|x,y_{1:t-1})}{P(s|x,y_{1:t-1})} \tag{13}$$

which is Formula 1.

Second is Formula 3. According to the law of total probability, if s represents the desired style and $\bar{s}$ represents the undesired style, the probability distribution $P(x,y_{1:t-1},y_t)$ can be transformed into:

$$P(x,y_{1:t-1},y_t) = \sum_{s'\in\{s,\bar{s}\}} P(s',x,y_{1:t-1},y_t). \tag{14}$$

Then, by substituting formula 14 into formula 10, we can obtain:

$$P(s|x,y_{1:t-1},y_t) = \frac{P(s,x,y_{1:t-1},y_t)}{\sum_{s'\in\{s,\bar{s}\}} P(s',x,y_{1:t-1},y_t)} \tag{15}$$

which is Formula 3.

Third is Formula 4. According to the multiplication theorem of probability, the probability distribution $P(s',x,y_{1:t-1},y_t)$ can be transformed into:

$$P(s',x,y_{1:t-1},y_t) = P(s',x)P(y_{1:t-1},y_t|s',x). \tag{16}$$

Then, as mentioned in the main text, for any given s', $P(s',x) = P(s')P(x)$, and then $P(s',x,y_{1:t-1},y_t)$ can be transformed into:

$$P(s',x,y_{1:t-1},y_t) = P(s')P(x)P(y_{1:t-1},y_t|s',x) \tag{17}$$

which is Formula 4.

Then, by substituting formula 17 into formula 15, we can obtain:

$$P(s|x,y_{1:t-1},y_t) = \frac{P(s)P(y_{1:t-1},y_t|s,x)}{\sum_{s'\in\{s,\bar{s}\}} P(s')P(y_{1:t-1},y_t|s',x)} \tag{18}$$

which is Formula 5.

6.2 Generation examples by our framework

Some examples of stylized captions for images generated by our framework.

Factual: a dog is laying on the floor with its legs crossed.
Pos: a cute dog is laying on a wooden floor.
Neg: a scared dog sits on the floor in front of its owner.
Roman: a dog playing happily on a red carpet.
Humor: a dog looking for some fun to do.

Factual: a dog is riding a surfboard in the water.
Pos: a great dog is enjoying the waves of a wave.
Neg: a bad dog jumping in the air with a surfboard.
Roman: a dog playing happily in the surf.
Humor: a dog running to get to the beach to surf.

Factual: a dog running through a lush green field.
Pos: a happy dog is running through the grass.
Neg: a crazy dog is chasing after a small animal.
Roman: a dog running happily through the field.
Humor: a dog running fast to get to his destination.

Factual: a dog is sitting in a boat on the water.
Pos: a calm dog is sitting on a boat in the water.
Neg: a dead dog is sitting on a boat in the water.
Roman: a dog relaxes in a canoe as it travels the water.
Humor: a dog looking for his owner in a canoe.

Factual: a dog is standing in the water with its mouth open.
Pos: a friendly dog is playing in the water.
Neg: a dirty dog is laying on the floor.
Roman: a dog playing outside in the rain after having a bath.
Humor: a dog looking for his master to come home.

Factual: a snowboarder is in mid air doing a trick.
Pos: a strong guy rides his snowboard down a hill.
Neg: a dead man in the air with a snowboard on his feet.
Roman: a snow boarder soaring high in air to show off for his fans.
Humor: a skate boarder jumps high to do a trick.

Factual: two men in white shirts and black ties playing a guitar.
Pos: two handsome guys sitting in a chair playing a guitar.
Neg: two dead people sitting down playing a guitar on stage.
Roman: two musicians perform for a crowd at a concert.
Humor: two musicians, singing, playing guitar and singing.

Fig. 6: Some examples of stylized captions for images generated by our framework.

6.3 Comparison of standard and generative discriminator

This subsection provides a detailed explanation of the differences between the standard discriminator and generative discriminator.

The function of a standard discriminator with some style s is to take the entire sentence $y_{1:T}$ as input and outputs a corresponding score $P(s|y_{1:T})$ to indicate whether the sentence align with the desired style, as shown in Figure 7.

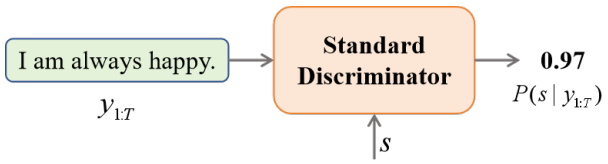


Fig. 7: The function of a standard discriminator.

If the standard discriminator is to be used as a post-processing module to guide the factual model to generate sentences, when generating the t-th word y_t, it is necessary to input all possible combinations of the generated words $y_{1:t-1}$ and all candidate words for y_t (namely all words in the vocabulary) to the standard discriminator. Then, the standard discriminator outputs the probability of every sentence aligning with s and further combines them into the probability distribution $P(s|y_{1:t})$ to indicate whether every possible combination aligns with the desired style, which is computationally intensive, as shown in Figure 8.

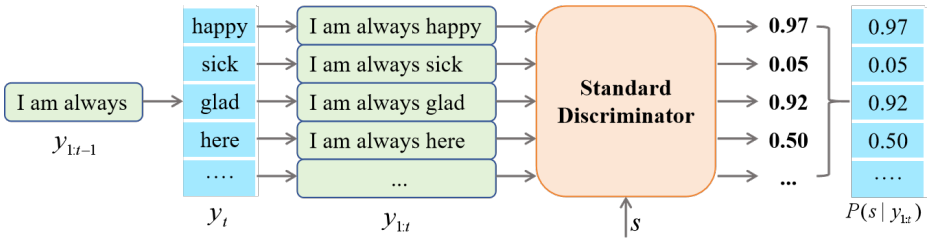


Fig. 8: How to use the standard discriminator as a post-processing module.

Compared to standard discriminator, the generative discriminator with some style s only needs the generated words $y_{1:t-1}$ as input and then directly predict the probability distribution $P(s|y_{1:t})$ to indicate whether every possible words in the vocabulary aligns with the desired style, which clearly save computation, as shown in Figure 9. And the subsection 3.2 in the main text provides a detailed explanation of how we implement such a generative discriminator functionality. such a generative discriminator functionality.

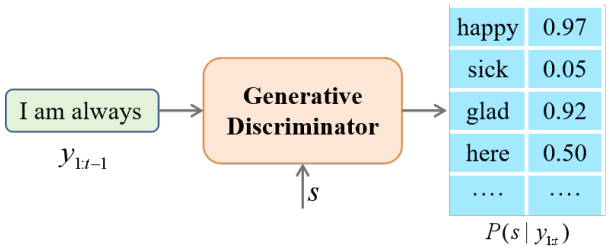


Fig. 9: The function of a generative discriminator.

In addition, the standard discriminator can only judge whether a complete sentence align with some desired style. However, if it is to be used to guide the generation of factual model, especially for sentences where stylistic elements are located towards the end, it may fail to give the correct guidance. It is necessary to segment the sentence into segments and then use them to train the standard discriminator, enabling it to predict whether an incomplete sentence may align with the desired style in the future. In contrast, the generative discriminator inherently possesses the ability to judge whether a incomplete sentence may align with the desired style in the future.

6.4 Comparison examples in ablation study

This subsection introduces some examples of captions for images generated by the factual model, style model, factual+style framework and PPCAP framework on Senticap dataset.

FM: a group of people standing on top of a lush green field.
SM: a woman reaches for a ski sitting in the beautiful snow.
FS: a woman in a wetsuit surfs through blue clear water.
PPCap: a beautiful woman in a bikini riding a wave on top of a surfboard.

FM: a man is standing next to a bus.
SM: a large green busy train is traveling past some wooden crates.
FS: a man is walking down a beautiful street glancing at the bus stop.
PPCap: a nice man is standing in front of a bus.

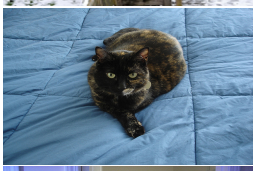
FM: a man sitting on a bench overlooking the ocean.
SM: a man sits next to a tranquil pond in the middle of a clear field.
FS: a man sits on a green bench next to an amazing tree.
PPCap: a happy man sitting on a bench overlooking the ocean.

FM: a red and white bus parked next to a building.
SM: a large green busy train is traveling past some wooden crates.
FS: a large bus drives down a charming street in the city.
PPCap: a great image of a bus and a car on the street.

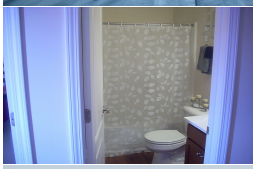
FM: a woman in a bikini riding a wave on her surfboard.
SM: a group of men on sunny beach with surfboards and one sitting on
FS: a group of people are flying kites on a lovely beach.
PPCap: a great group of people are flying kites in the park.

Fig. 10: Some examples of positive captions for images in alblation study.

FM: a bench covered in snow next to a tree.
SM: a lonely chair sits across from a couch
FS: a group of dead people sitting next to each other under blue
PPCap: a lonely bench sits in the snow with a lot of trees.



FM: a cat laying on a bed with its eyes closed.
SM: a shy cat lays on top of a couch near a jacket.
FS: a little cat crouches next to a brick rough wall.
PPCap: a shy cat sits on a bed with its eyes closed.



FM: a bathroom with a toilet, shower and sink.
SM: a dirty bathroom with a tub, toilet, and sink.
FS: a bathroom with a large bathtub and sink under a large broken mirror
PPCap: a dirty bathroom with a toilet and a shower curtain.

FM: two men on snowboards in the snow.
SM: a man skiing and two others playing with a dangerous dog.
FS: two people on skis standing on a cold mountain
PPCap: two stupid people on snowboards in the snow.

FM: a train on the tracks with a building in the background.
SM: a damaged statue of a giraffe is next to a barn
FS: a large brick damaged building stretches into the blue sky and
PPCap: a damaged building with a train on the tracks.

Fig. 11: Some examples of positive captions for images in alblation study.