Abstract—Semantic communication technologies enable wireless edge devices to communicate effectively by transmitting semantic meanings of data. Edge components, such as vehicles in next-generation intelligent transport systems, use well-trained semantic models to encode and decode semantic information extracted from raw and sensor data. However, the limitation in computing resources makes it difficult to support the training process of accurate semantic models on edge devices. As such, edge devices can buy the pretrained semantic models from semantic model providers, which is called “semantic model trading”. Upon collecting semantic information with the semantic models, the edge devices can then sell the extracted semantic information, e.g., information about urban road conditions or traffic signs, to the interested buyers for profit, which is called “semantic information trading”. To facilitate both types of the trades, effective incentive mechanisms should be designed. Thus, in this article, we propose a hierarchical trading system to support both semantic model trading and semantic information trading jointly. The proposed incentive mechanism helps to maximize the revenue of semantic model providers in the semantic model trading, and effectively incentivizes model providers to participate in the development of semantic communication systems. For semantic information trading, our designed auction approach can support the trading between multiple semantic information sellers and buyers, while ensuring individual rationality, incentive compatibility, and budget balance, and moreover, allowing them to achieve higher utilities than the baseline method.

Index Terms—Auction, incentive mechanism, semantic communication.

I. INTRODUCTION

With the advancement of sixth-generation (6G) mobile communication technology, data transmission rate in the conventional communication systems is increasing but approaching the Shannon limit. Meanwhile, the remaining available spectrum resources are becoming increasingly scarce. To solve this dilemma, semantic communication technologies are proposed [1], which aims to transmit the extracted semantic information relevant to the communications goal. Because the data amount that needs to be transmitted can be reduced significantly while ensuring the effectiveness of communications [2], semantic communications can be widely used in intelligent wireless networks, to enable smart transportation [3], smart logistic [4], smart cities [5], smart homes [6], and smart healthcare [7].

Existing semantic communication systems [2], [8] are pretrained with labeled datasets with certain channel models. However, a main drawback is that the accuracy and performance of a pretrained semantic model decrease when the background knowledge or communication environment changes, i.e., the mismatch between the knowledge base/channel model used in the training and the actual knowledge base/channel model. To reduce the gap in performance, fine-tuning of the model parameters can be done based on the real channel models [9] and new background knowledge [2]. However, edge and Internet of Things (IoT) devices with limited computation power might not have enough resources for fine-tuning. Moreover, the results of...
To model realistic semantic communication systems, inspired by the model trading framework in collaborative edge learning [10], we can adopt a trading system in which model providers trade the trained model with other devices. Specifically, the semantic model provider has more resources to train quality semantic models with the relevant knowledge base and channel models, and the edge devices can obtain the semantic model (semantic encoder/decoder) from the model providers. Using the semantic model, edge devices can extract semantic information from the collected raw data. This enables semantic information exchange between edge devices. Furthermore, as the semantic information is helpful for the decision making of smart agents [11], the trading of semantic information should also be studied. Using the semantic models, the edge devices collect and trade the semantic information with interested information buyers. For example, one vehicle can buy semantic information [12], [13] from nearby vehicles/smart sensors about the conditions of the surrounding environment.

To promote the above two types of trade in the semantic communication system, i.e., semantic model trading and semantic information trading, we should design novel and effective incentive mechanisms:

1) **Semantic model trading**: To encourage the participation of semantic model providers, incentive mechanisms are designed so that they are rewarded for supplying quality semantic models. In general, edge devices are willing to pay more for semantic models that can achieve better semantic performance. We are the first to propose a deep learning (DL) based auction mechanism to determine an allocation of the semantic model to the edge devices and the price to be paid by the edge devices to the model providers. We show analytically that the DL-based auction attains the properties of truthfulness while maximizing the revenue of the model providers.

2) **Semantic information trading**: To facilitate the semantic information trading between multiple semantic information buyers and edge devices, e.g., vehicles that are interested in collecting semantic information about the conditions of the surrounding environment [12], [13], we introduce a double auction mechanism to model the competition between the buyers and edge devices. In the auction, we propose semantic-based valuation functions, i.e., the valuation of the information is a function of semantic performance of the edge devices. In particular, the semantic information buyers are willing to pay more for the semantic information with higher accuracy, and hence the edge devices have more incentive to obtain better models from the semantic model trading. Moreover, the proposed double auction mechanism shows the desired properties of individual rationality, incentive compatibility, and budget balance, which are all significant properties to achieve sustainable and rational trading.

While many recent works have focused on improving the performance of semantic communication systems [1], [2], few have addressed the designs of incentive mechanisms for semantic communication systems. By achieving the aforementioned two kinds of trade, we propose a novel hierarchical trading system to enhance the economically-sustainable development of semantic communication systems. The main contributions of our article are:

- We propose an incentive design framework for the semantic model trading and semantic information trading to support the deployment of semantic communication systems. Our designed mechanisms support the development of semantic communication systems by motivating the participation of model providers to build and share high-quality semantic engines, buyers to acquire relevant and useful semantic information, and semantic information sellers to facilitate other stakeholders for semantic information exchange.

- We model the competition in the semantic model trading and semantic information trading with auction mechanisms. Different from conventional auctions, our auction can maximize the revenue of semantic model providers while achieving the properties of individual rationality and incentive compatibility. The simulation results are provided based on a case study on semantic text transmission where we derive the valuation functions based on the sentence similarity score and bilingual evaluation understudy (BLEU) score [14].

- We propose an effective feature reduction method for data transmission under a limited data transfer budget. In contrast to existing works of feature reduction techniques for semantic communication systems [15], [16], our method does not increase communication cost and reduce the performance gap between partial feature and full feature. Compared with our previous work [17], the significant extensions in this article include:

  - In contrast to previous work [17] in which the incentive mechanism is customized for wireless powered devices, we propose a general framework that can be applied to semantic communication systems with different purposes. A case study is provided to show a concrete application of the generalized framework on semantic text transmission. A new method for effective feature reduction is proposed to reduce the gap of performance when only partial data is transmitted because of limited communication resources.

  - While the previous work [17] focuses on semantic information transfer, we consider both and joint semantic model trading and semantic information trading in this article. In the experiments to investigate the relationship between model trading and semantic trading, we show that the accuracy of the model obtained in model trading could affect the utility of the device in information trading.

  - To model realistic semantic communication systems, multiple semantic information buyers and sellers are considered instead of a single buyer setting in the previous work [17]. A new double auction mechanism is proposed in this article to account for the multi-buyers multi-sellers scenario in the semantic information trading.

Our article is organized as follows. In Section II, we discuss the related works of semantic communication systems and incentive mechanism design. In Section III, we detail the system model and problem formulation. In Section IV, we present a case study of semantic model trading and semantic information exchange.
trading for semantic text transmission. In Section V, we present
the numerical results, and Section VI concludes the article.

II. RELATED WORK

A. Vehicular Networks

With the development of vehicular infrastructure in recent
years, vehicles can be seen as important network players with
computing, caching and communication capabilities [18], [19].
However, as the number of vehicles increases, the vehicular
network relies heavily on reliable real-time communication and
interactions for complex operations [20], [21], such as route
planning and collision avoidance. Thus, timely and accurate
information updates are vital to the development of the vehicular
networks. This implies that the conventional communication
paradigm which seeks the lowest latency is no longer a sustain-
able development direction. To make fast and accurate decisions
in vehicular networks, it is important to leverage the semantic
meaning of information [22]. The authors in [23] design a
resource allocation algorithm for semantic video transmission in
vehicular networks. By using the proposed algorithm [23],
the semantic understanding accuracy of the video transmission
is optimized by a multi-agent deep Q-network. The simulation
results show that the proposed method can achieve as high as
70% improvement for the density of correctly detected objects,
compared with the conventional QoS and QoE based resource
allocation methods.

However, it is not realistic to train a usable semantic model
for each vehicle, due to the limited computing resources and
the dynamic positioning of vehicles [24]. Therefore, we will
consider a semantic model trading system in this article. More-
over, considering the importance of semantic information in the
vehicular networks, vehicles can then sell the semantic informa-
tion to potential buyers. The trading of semantic information is
gaining attraction especially for the sustainable development of
large-scale multi-agent systems.

B. Deep Learning Enabled Semantic Communication Systems

Conventional communication systems focus on transmitting
bits or symbols with minimum error from the transmitter to the
receiver, and the performance is evaluated at the bits or symbols
level. In contrast to the traditional communication systems,
semantic communication system aims to communicate at the
semantic level, where performance is evaluated by the recovery
of the meanings of the data instead of bits accuracy. Semantic
communication systems for text [2], speech signals [8], and
multimodal data [25] first encode the data by a semantic encoder
and send the encoded semantic information to the receivers.
The receivers then decode the received signals with semantic
decoders to recover the original data. Typically, the semantic
encoders and decoders are implemented by end-to-end DL net-
works and trained with labeled data.

To improve the encoding efficiency, several works focus on
reducing the size of the data during transmission. The authors
in [15] mask the bits according to the original sentence length
to save the transmission resources. For the image classification
task, the authors in [16] use the gradient of the neural network
to select important features. However, the proposed method
requires extra storage cost to store the gradients of weights of
the network. Most of the existing data reduction techniques are
implemented together with the training process. A drawback
is that, after the model is trained and parameters are fixed,
the gradients, is stored to select important features during
data reduction, our approach does not require extra storage
for data reduction. From the simulation results, our method
effectively reduces the performance gap when only partial data
is transmitted.

Our method uses controlled dropout regularization [26] dur-
ing the training process. The dropout layer drops units of the
encoded text features before transmission mimicking the data
reduction process. Instead of the need to mask the bit accord-
ing to the sentence length [15], our method applies the same
mechanism to all sentences. The dropout layer also does not
require extra memory to store the gradients in contrast to [16].
The simulation result shows that our method helps the semantic
communication model to maintain the performance at every
level of data reduction.

C. Incentive Mechanism Design

In real-world settings, data transmissions are limited by the
communication resources such as bandwidth and energy. In-
centive mechanisms are designed to encourage certain parties
to contribute to a communication network. For example, in
a multi-node wireless powered communication network, selfish
wireless nodes are not willing to charge other nodes by
consuming their resources. To encourage the participation of
these nodes, [27] proposed incentive schemes to deal with the
selfishness of wireless nodes with Age of Information (AoI)
based utility functions. In collaborative edge learning, incentive
mechanisms are used to incentivise the data owners to provide
the updated model parameters for global model aggregation [28].

Auction mechanisms are used to motivate and attract partic-
ipsants for data collection and computation tasks in vehicular
network. In [29], auction mechanism is used to encourage more
vehicles to participate in the vehicular crowdsensing (VCS)
system to improve the event detection of road networks. In the
proposed auction mechanism, reputations of the participants are
considered when determining the payment to winning partici-
pants. A double auction mechanism is proposed in [30] to facil-
itate the coded distributed computing (CDC) task by matching
the edge servers and the vehicles. The proposed mechanism also
helps to decide the price paid by the vehicles for the resources
of the edge servers.

Given that most of the communication networks are using
conventional communication systems, semantic-aware incentive
mechanisms are needed to be designed to motivate the participa-
tion of all parties in the development of semantic communication
systems. In our previous work [17], semantic-aware incentive
mechanism is proposed to solve the energy allocation problem
in wireless powered communication systems. The proposed

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 1. The system model which includes semantic model trading (lower part) and semantic information trading (upper part). In the semantic model trading, edge devices trade with semantic model providers to obtain semantic encoders/decoders for semantic communications. In the semantic information trading, edge devices equipped with semantic encoders/decoders trade semantic information with buyers.

In contrast to [29], [31], where valuation/pricing is formulated by data quantity, we consider the semantic aspect in our valuation function to improve the semantic performance of the system. Different from existing double auction mechanisms [30], [32] where prices are determined by fixed rules, our method utilizes the learning ability of deep learning networks to optimize the prices such that the revenue of the sellers is maximized. Maximizing the revenue of the sellers helps to attract more model providers to provide high-quality semantic models which improve the performance of semantic communication systems. The simulation results show that our auction mechanisms outperform conventional auction mechanisms (without the deep learning network) in terms of seller’s utility while satisfying the desired properties of individual rationality, incentive compatibility, and budget balance.

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a semantic communication network (Fig. 1) that consists of a set \( M = \{1, 2, \ldots, m, \ldots, M\} \) of \( M \) edge devices. To perform semantic encoding and decoding, the edge devices have to obtain the trained semantic models from the semantic model providers. Model trading is a common practice in collaborative edge learning, and in particular, federated learning [10], where the model providers (sellers) receive incentives for providing trained models to the participants (buyers). In the case of semantic communications, the models being traded are the semantic encoders and decoders used for semantic information encoding and decoding, respectively. Devices with limited computation and communication resources can obtain high-quality semantic models from model trading. Moreover, it is shown that mismatches in communication channels and background knowledge of the communication environment degrade the performance of a pretrained semantic communication model [2]. Therefore, trading with model providers that perform machine learning training based on the relevant background knowledge and communication environment helps to improve the semantic performance of the devices. For example, devices can trade with the model providers that collect training data from the same certain geographical area as the buyers [33].

To encourage the participation of model providers, incentive mechanisms should be designed to ensure that model providers are appropriately rewarded from the semantic model trading process. Similar to incentive mechanisms designed for the model trading in FL, the devices have to compete to obtain the semantic models from the semantic model providers. Intuitively, the devices are willing to pay more if the model obtained can achieve high semantic performance.

Besides, there exists a set \( N = \{1, 2, \ldots, n, \ldots, N\} \) of \( N \) semantic information buyers that are interested to obtain semantic information from the devices. For example, this may be semantic information trading between UAVs in real time [11], and collection of semantic information for image classification tasks for autonomous vehicles [16]. In this case, incentive mechanism design is also needed to facilitate the trading of such semantic information.

In the following, we propose two auction mechanisms for semantic model trading and semantic information trading. In the semantic model trading, we adopt a DL-based auction mechanism to derive the semantic-aware valuation of the semantic models. The semantic model trading could be a channel to supply the semantic model for the devices to extract semantic information. Then, the semantic information from devices with higher accuracy is more valuable to semantic information buyers. For semantic information trading, we study
the double auction mechanism for information trading between multiple buyers and multiple semantic information sellers. We further investigate how the semantic model obtained from the semantic model trading affects the results in the semantic information trading. The overall system is shown in Fig. 1. The lower part of Fig. 1 shows semantic model trading and the upper part illustrates the semantic information trading.

B. Auction for Semantic Model Trading

The valuation of the devices for the model provided by the service provider is given by:

\[ v_m = A_p - A_m, \]

where \( A_p \) is the accuracy of the model from the model provider, and \( A_m \) is the accuracy of the current model of device \( m \) (\( A_m = 0 \) if the device does not own any model). The accuracy metric can be the text similarity score for semantic text transmission [2], signal-to-distortion ratio (SDR) for semantic speech signal transmission [8], and answer accuracy in visual question answering (VQA) [25]. In every round of the single-item auction, the model provider, i.e., the auctioneer collects bids \( (b_1, b_2, \ldots, b_M) \) from all smart devices, i.e., bidders, and then decides the winner, \( m^* \), and corresponding payment price, \( \theta_m \). The utility of the device is given by \( u_m = v_m - \theta_m \), if the device is the winner and \( u_m = 0 \) otherwise.

Traditional single-item auctions such as the first-price auction and Second-Price Auction (SPA) can be used to determine the winner and price. For an auction to be optimal [34], it should attain the properties of incentive compatibility and individual rationality. Individual rationality guarantees that the utility of the devices is non-negative by participating in the auction, i.e., \( u_m \geq 0 \). Incentive compatibility ensures that each device submits bids according to their true valuations, respectively, i.e., \( b_m = v_m \), regardless of the actions of other devices, and the utility of each device is maximized by submitting the truthful bid. In the first-price auction, the highest bidder wins and pays the exact bid submitted, maximizing the revenue gain of the model provider but does not guarantee incentive compatibility. In SPA, the highest bidder wins but pays the price of the second highest bidder. SPA ensures incentive compatibility but does not maximize the revenue of the model provider.

We adopt a DL-based optimal auction mechanism [35] that can maximize the revenue of the seller while achieving the properties of incentive compatibility and individual rationality. The auctioneer (i.e., the model provider) does not have a priori knowledge about the bidders and optimal decisions in determining the winner. Nevertheless, the model provider can learn from experience and adjust the auction decision using DL-based optimal auction. The DL-based auction consists of three major functions: monotone increasing function, \( \Phi_m \), allocation rule, \( z_m \), and conditional payment rule, \( \theta_m \). Firstly, the input bids, \( b = (b_1, \ldots, b_m, \ldots, b_M) \), are transformed by \( Q \) groups of \( S \) linear functions, followed by the \( \min \) and \( \max \) operations, i.e., the transformed bid,

\[ \tilde{b}_m = \Phi_m(b_m) = \min_{q \in Q} \max_{s \in S} (w_{qs}^m b_m + \beta_{qs}^m), \]

where \( w_{qs}^m \in \mathbb{R}_+, q = 1, \ldots, Q, s = 1, \ldots, S \) and \( \beta_{qs}^m \in \mathbb{R} \). The allocation probability is

\[ z_m(\tilde{b}) = \frac{e^{\tilde{b}_m}}{\sum_{j=1}^{M+1} e^{\tilde{b}_j}}, \]

Theorem 1: (35) For any set of strictly monotonically increasing function \( \{\Phi_1, \ldots, \Phi_M\} \), an auction defined by allocation rule \( z_m = z_0 \circ \Phi_m \) and the payment rule \( \theta_m = \Phi_m^{-1} \circ \theta_0 \circ \Phi_m \) has the properties of incentive compatibility and individual rationality, where \( z_0 \) and \( \theta_0 \) are the allocation and payment rule of a second price auction with zero reserve, respectively, and \( \circ \) indicates function composition, i.e., \( f \circ g(x) = f(g(x)) \).

To ensure that the auction learned by the network achieves incentive compatibility and individual rationality, we constrain the allocation and payment rules of the network by following Theorem 1. After the monotone transformation, the transformed bids are passed to separate networks that approximate the allocation and payment rule. The allocation rule which follows the second price auction with zero reserve (SPA-0) allocation rule is approximated by a softmax function to maximize the allocation probability of the highest bid, i.e.,

\[ \theta_m^0(\tilde{b}) = \text{ReLU} \left( \max_{j \neq m} \tilde{b}_j \right), \]

where \( \text{ReLU}(x) = \max(x, 0) \) is used to ensure that the payment is non-negative. To obtain the payment price, the inverse transformation function is applied on the SPA-0 price of the

---

**Algorithm 1: DL-Based Auction (DLA) Algorithm.**

**Input:** Bids of devices \( b = (b_1, \ldots, b_m, \ldots, b_M) \)

**Output:** Winner and Payment Price

1: **Initialization:**
   \[ w = [w_{qs}^m] \in \mathbb{R}_+^{M \times QS}, \beta = [\beta_{qs}^m] \in \mathbb{R}^{M \times QS} \]

2: **while** Loss function \( \hat{R}(w, \beta) \) is not minimized **do**
3:   Compute transformed bids \( \tilde{b}_m = \Phi_m(b_m) = \min_{q \in Q} \max_{s \in S} (w_{qs}^m b_m + \beta_{qs}^m) \)
4:   Compute the allocation probabilities \( z_m(\tilde{b}) = \text{softmax}(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_{M+1}; \kappa) \)
5:   Compute the SPA-0 payments \( \theta_m^0(\tilde{b}) = \text{ReLU} \left( \max_{j \neq m} \tilde{b}_j \right) \)
6:   Compute the conditional payment \( \theta_m = \Phi_m^{-1}(\theta_m^0(\tilde{b})) \)
7:   Compute the loss \( \hat{R}(w, \beta) \)
8:   Update parameters \( w \) and \( \beta \) using SGD optimizer

9: **end while**
10: **return** Winner \( m^* \) and payment price \( \theta_m \).
transformed bids, i.e.,
\[ \theta_m = \Phi_m^{-1}(\theta_m'(b_m)) , \]
where the inverse transformation function can be expressed by:
\[ \Phi_m^{-1}(y) = \max_{q \in Q} \min_{s \in S}(w_{qs}^m - y - \beta_q^m) . \]
To maximize the revenue, the network optimizes a loss function that is the negative value of the seller revenue. The loss function is given by
\[ \hat{R}(w, \beta) = - \sum_{m=1}^{M} z_m(b_m) \theta_m . \]

After the auction, model providers need to ensure that the performance is optimized for the winning buyers, i.e., model providers should help to finetune the performance of the semantic model using the real channel condition to reduce performance gap due to channel mismatch [8], [25]. The accuracy of the model should be verified after the finetuning, e.g., sentence similarity should be evaluated for text semantic transmission [2]. Feedback ratings could be given to the model providers and the reputation of the model provider can be considered in the pricing scheme [29].

Future updates are required when the communication scenarios are different [2]. When there is a change in the background knowledge, incremental training could be provided with the new knowledge base. The transfer learning method can be adopted to speed up the training time as shown in [2]. Instead of re-training all parameters, transfer learning freezes certain parts of the model and finetunes the targeted encoder/decoder of the model. In the case of different knowledge backgrounds, only part of the semantic encoder and decoder needs to be finetuned by transfer learning.

C. Auction for Semantic Information Trading

We consider \( N \) semantic information buyers and \( M \) devices, where the buyers are interested in buying semantic information from the devices. Considering that the devices obtain the semantic models from the semantic model trading, the semantic information buyers are willing to pay more for the semantic information from the devices with high accuracy, \( A_m \). The semantic encoder and decoder of the sellers and buyers can be obtained from the model providers. To ensure security and privacy, only the model buyers with access rights can retrieve the trained model [36]. Alternatively, if there are enough communication resources, the semantic encoders and decoders can be trained jointly to ensure good performance at anytime. The training of semantic encoder and decoder can be done by optimizing loss functions of respective communication goals [2], [8], [25].

In our system model, we consider that the buyers and sellers are using the same encoder and decoder system as this is recommended for devices operating in the same geographical area [33]. The key advantage of this setting is that the devices are optimized for the real channel condition of the transmission so that there is no degradation of performance due to channel mismatch [2]. Moreover, it is not advisable for buyers and sellers with different semantic models to trade their information because of the poor performance without the mapping of the embedding space [37].

We propose a single-round double auction for the one-to-one mapping of the buyers and the sellers. Similar to [30], [32], a trusted third party, e.g., service provider, is needed as the auctioneer to assist the auction process. In the double auction, there are
- A set of semantic information buyers \( \mathcal{N} = \{1, \ldots, n, \ldots, N\} \)
- A set of semantic information sellers \( \mathcal{M} = \{1, \ldots, m, \ldots, M\} \), the devices that provide semantic information to the buyers
- A trusted third party, the auctioneer, e.g., service provider.

Noted that service provider is not the same as the model provider in semantic model trading (Section III-B).

Based on the semantic performance, each buyer has different preferences for the devices. Let \( b_m = (b_m^1, \ldots, b_m^M) \) denote the bid vector of buyer \( n \), where \( b_m^n \) is the bid of buyer \( n \) for device \( m \), i.e., the price that buyer \( n \) is willing to pay for receiving semantic information from device \( m \).

Let \( a = (a_1, \ldots, a_M) \) denote the ask vector of the devices, where \( a_m \) is the ask of device \( m \), i.e., the price that device \( m \) is willing to receive for trading the semantic information. The value of the semantic information from device \( m \) to buyer \( n \) can be expressed as
\[ v_n^m = A_n^m(a_m) , \]
where \( A_n^m(a_m) \) is the accuracy of the semantic information transmitted by device \( m \) to buyer \( n \), and \( A_m \) (determined by the semantic model trading) is the upper bound of the achievable accuracy of the current semantic model.

Let \( p_n \) be the price that buyer \( n \) pays, the utility of buyer \( n \) is given by
\[ u_n^b = \begin{cases} v_n^m - p_n & \text{if buyer } n \text{ wins the auction}, \\ 0 & \text{otherwise}. \end{cases} \]

Note that to compare the utility of buyer \( n \) when it wins different devices, we also use \( u_n^b \) and \( u_n^m \) to denote the utility of buyer \( n \) when it wins the semantic information of device \( m \) and \( m' \), respectively.

Following [38], the data collection cost is given by
\[ c_d^m = d_\gamma m , \]
where \( d_\gamma m \) and \( \gamma m \) are the data size and unit data cost, respectively. The computational cost can be formulated as
\[ c_c^p m = d_\Gamma m , \]
where \( \Gamma \) is the unit computational cost to extract semantic information from the collected data. This cost can be due to the energy consumption [28] or edge/cloud computation resource rental fee [39]. The communication cost for device \( m \) to transmit the semantic information is
\[ c_m = P_m N_m \frac{\nu}{R} , \]
where \( P_m \) is the communication power, \( N_m \) is the number of bits used to represent the semantic information, \( R \) is the transmission.
Algorithm 2: Candidate Determination and Pricing.

Input: \( N, M, \mathcal{B}, \mathbf{a} \)

Output: \( N_c, M_c, \mathcal{P}_c, \mathcal{Y}_c, \sigma \)

1: for \( m \in M \) do
2: \( n, \theta_n = DLA(\{b_n^{m,n}, \forall n \in N\}) \)
3: if \( \theta_n \geq a_m \) then
4: \( \hat{\sigma}(m) = n \), buyer \( n \) is added into \( N_c \), and seller \( m \) is added into \( M_c \)
5: \( p_n = y_m = \theta_n \)
6: \( p_n \) and \( y_m \) are added into \( \mathcal{P}_c \) and \( \mathcal{Y}_c \), respectively
7: end if
8: end for

rate in bits per second, and \( \nu_m \) is the unit energy cost for communication. The cost of the semantic model is given by

\[
c_m^{md} = \frac{\theta_m}{T_m},
\]  

(13)

where \( \theta_m \) is the price paid for the current semantic model (determined by the model trading auction in Section III-B) and \( T_m \) is the expected number of transmissions with the model.

The total cost for device \( m \) to transmit the semantic information is then defined as follows:

\[
C_m = c_m^{dp} + c_m^{ep} + c_m^{cm} + c_m^{md} = d_m \gamma_m + d_m \Gamma_m + P_m \frac{N_m}{R} \nu_m + \frac{\theta_m}{T_m},
\]  

(14)

Let \( y_m \) be the payment to device \( m \), the utility of the device \( m \) is given by

\[
u_m^s = \begin{cases} 
  y_m - C_m & \text{if device } m \text{ wins the auction,} \\
  0 & \text{otherwise.}
\end{cases}
\]  

(15)

The proposed double auction has two stages, the candidate-determination and pricing stage, and the candidate-elimination stage. The algorithms for the two stages are shown in Algorithms 2 and 3 respectively. Note that the DLA refers to the DL-Based Auction in Algorithm 1. In the candidate-determination and pricing stage, the auctioneer determines the buyer candidates of each device, the prices that the buyer candidates pay, and the payment to be rewarded to the devices.

Let \( n \) and \( \theta_n \) denote the winning buyer and payment price determined by DLA, respectively. For each device \( m \), all bids are sent to DLA to determine the winner and payment price. If the payment price is not lower than the ask \( a_m \), i.e., \( \theta_n \geq a_m \), then the buyer \( n \) is added to the set of buyer candidates \( N_c \) with price \( p_n = \theta_n \), and device \( m \) is added to the set of seller candidates \( M_c \) with payment \( y_m = \theta_n \).

After the first stage, each buyer candidate may win more than one device. In the candidate-elimination stage, for each buyer candidate, the algorithm selects the best device such that the buyer yields the highest utility in (9). If more than one device yields the same highest utility for the buyer, the best device is randomly selected.

In the following, we prove that the double auction mechanism in our model satisfies the properties of individual rationality, incentive compatibility, and budget balanced. Theorem 2 shows that all winning buyers and sellers are rewarded with non-negative utilities. This is to ensure that all participants will not have negative utilities and increase their interest to join the auction. By showing that the auction is incentive compatible in Theorem 3, we ensure that the bids and asks are based on true valuation so that the information is transmitted to the buyers that value it the most.

**Theorem 2:** The proposed double auction mechanism is individually rational. All winning buyers and sellers are rewarded with non-negative utilities i.e. \( p_n \leq b_n^{m} \) and \( y_m \geq a_m \).

**Proof:** From Algorithm 2, since DLA has the property of individual rationality [35], we have \( \theta_n \leq b_n^{m} \). Therefore \( p_n \leq b_n^{m} \) and \( y_m \geq a_m \), individual rationality is satisfied in the candidate determination and pricing stage. Since Algorithm 3 does not change the value of \( p_n \) and \( y_m \), the individual rationality is preserved after the candidate eliminations.

**Theorem 3:** The proposed double auction mechanism is incentive compatible. All buyers and sellers submit their bids and asks truthfully as they cannot improve their utilities by submitting bids and asks that are different from their true valuations.

**Proof:** We prove the incentive compatibility by the following lemmas:

1. The proposed double auction mechanism is truthful for the sellers (as shown in Lemma 1).
2. The proposed double auction mechanism is truthful for the buyers (as shown in Lemma 2).

**Lemma 1:** The proposed double auction mechanism is truthful for the sellers.

**Proof:** To prove that the proposed double auction mechanism is truthful for the sellers, we discuss the three possible outcomes for the sellers in the following subsets:

1. Subset \( M_w \), sellers that win the auction,
2) Subset $\mathcal{M}_c \setminus \mathcal{M}_w$, sellers that are selected as candidates but are eliminated during the candidate elimination stage, and
3) Subset $\mathcal{M} \setminus \mathcal{M}_c$, sellers that are not selected as candidates.

In each of the subsets, we discuss the cases where the sellers bid untruthfully. In each case, we show that the sellers cannot achieve higher utilities with the untruthful bids. Note that tilde $\tilde{\cdot}$ is shown for the notations to indicate the outcomes of the untruthful cases.

1) For seller $m \in \mathcal{M}_w$:
   \begin{enumerate}
   \item Case 1: Seller $m$ does not win the auction with untruthful ask, $\tilde{u}_m^s = 0 \leq u_m^s$.
   \item Case 2: Seller $m$ wins the auction with untruthful ask. In this case, the payment does not change because the input bids to DLA are not changed, i.e., $\tilde{u}_m^s = u_m^s$.
   \end{enumerate}
2) For seller $m \in \mathcal{M}_c \setminus \mathcal{M}_w$, changing ask does not change the price as discussed in the case of seller $m \in \mathcal{M}_w$. Therefore seller $m \in \mathcal{M}_c \setminus \mathcal{M}_w$ does not win the auction regardless of the value of $a_m$, $\tilde{u}_m^s = u_m^s = 0$.
3) For seller $m \in \mathcal{M} \setminus \mathcal{M}_c$:
   \begin{enumerate}
   \item Case 1: Seller $m$ does not win by asking untruthfully, i.e., $m \notin \mathcal{M}_w$, therefore the utility remains unchanged, $\tilde{u}_m^s = u_m^s = 0$.
   \item Case 2: Seller $m$ wins by asking untruthfully, i.e., $m \in \mathcal{M}_w$. Let buyer $n$ be the winner of semantic information from $m$ with price $\tilde{p}_n = \tilde{y}_m$. To win the auction, $m$ has to ask lower than the true valuation such that $\tilde{a}_m < C_m$. As the payment is not affected by $\tilde{a}_m$, we have $\tilde{y}_m = y_m$ and since $m$ does not win by asking truthfully, $y_m < C_m$, therefore $m$ suffers negative utility in this case, i.e., $\tilde{u}_m = y_m - C_m < 0 = u_m^s$.
   \end{enumerate}

Therefore we can conclude that the sellers cannot obtain a higher utility by asking untruthfully.

\begin{lemma}
The proposed double auction mechanism is truthful for the buyers.
\end{lemma}

\begin{proof}
To prove that the proposed double auction mechanism is truthful for the buyers, we discuss the two possible outcomes for the buyers in the following subsets:
1) Subset $\mathcal{N}_w$, buyers that win the auction, and
2) Subset $\mathcal{N} \setminus \mathcal{N}_w$, buyers that lose the auction.

In each of the subsets, we discuss the cases where the buyers ask untruthfully. In each case, we show that the buyers cannot achieve higher utilities with the untruthful asks. Note that tilde $\tilde{\cdot}$ is shown for the notations to indicate the outcomes of the untruthful cases.

1) For buyer $n \in \mathcal{N}_w$, assuming $n$ wins seller $m$ by bidding truthfully. Let us consider the following cases when buyer $n$ bids untruthfully:
   \begin{enumerate}
   \item Case 1: Buyer $n$ loses with untruthful bid, $\tilde{u}_n^b = 0 = u_n^b$.
   \item Case 2: Buyer $n$ wins the same seller $m$ with untruthful bid, given individual rationality property of DLA, we have $\tilde{u}_n^b \leq u_n^b$.
   \item Case 3: Buyer $n$ wins with a different seller $m'$ with untruthful bid. Let us consider the following cases when buyer $n$ bids truthfully:
   \begin{itemize}
   \item Seller $m' \in \mathcal{M}_c$ and $\tilde{\sigma}(m') = n$. Since buyer $n$ wins $m$ in the truthful case, we have $\tilde{u}_{n,m}^b \leq u_{n,m}^b$. Given that DLA has the property of individual rationality, we have $\tilde{u}_{n,m'}^b \leq u_{n,m'}^b$. Thus we know that $\tilde{u}_{n,m'}^b \leq u_{n,m'}^b$.
   \item Seller $m' \in \mathcal{M}_c$ and $\tilde{\sigma}(m') \neq n$. It means that there is another buyer candidate $m'$ with higher or equal bid for $m'$, i.e., $b_{n,m'}^b \geq b_{n,m}^b$. When buyer $n$ wins $m'$ by bidding untruthfully, since DLA satisfies the individual rationality constraint, we have $\tilde{u}_{n,m'}^b \leq 0$. From Theorem 2, we know that $u_{n,m}^b \geq 0$ (all winning buyers and sellers are rewarded with non-negative utility), thus we have $\tilde{u}_{n,m}^b \leq u_{n,m}^b$.
   \end{itemize}
\end{enumerate}
\end{enumerate}

\begin{enumerate}
\item Case 1: Buyer $n$ loses with untruthful bid, $\tilde{u}_n^b = 0 = u_n^b$.
\item Case 2: Buyer $n$ wins seller $m$ by bidding untruthfully. Since DLA has the property of individual rationality, we have $\tilde{u}_{n,m}^b \leq u_{n,m}^b$. From Theorem 2, we know that $u_{n,m}^b \geq 0$, thus we have $\tilde{u}_{n,m}^b \leq u_{n,m}^b$.
\end{enumerate}

2) For buyer $n \in \mathcal{N} \setminus \mathcal{N}_w$ with utility $u_n^b = 0$. We consider the following cases when buyer $n$ bids untruthfully:

\begin{enumerate}
\item Case 1: Buyer $n$ loses with untruthful bid, $\tilde{u}_n^b = 0 = u_n^b$.
\item Case 2: Buyer $n$ wins seller $m$ by bidding untruthfully. Since DLA has the property of individual rationality, we have $\tilde{u}_n^b \leq u_n^b$.
\end{enumerate}

Therefore we can conclude that the buyers cannot obtain a higher utility by bidding untruthfully.

\begin{theorem}
The proposed double auction mechanism is budget balanced. The total price paid by the winning buyers is not less than the total payment to the winning sellers, i.e., \( \sum_{n \in \mathcal{N}_w} p_n \geq \sum_{m \in \mathcal{M}_w} y_m \).
\end{theorem}

\begin{proof}
According to Algorithms 2 and 3, the price that winning buyers pay and the payment received by winning sellers are equal for every winning seller-buyer pairs. Thus, we have

\begin{equation}
\sum_{n \in \mathcal{N}_w} p_n = \sum_{m \in \mathcal{M}_w} y_m = 0.
\end{equation}

\end{proof}

We can conclude that the double auction mechanism is budget balanced.

In Algorithm 2, since there are $M$ sellers in set $\mathcal{M}$, the time complexity of the candidate determination and pricing stage is $O(M)$. In Algorithm 3, we know that $|\mathcal{M}_c| \leq |\mathcal{M}| = M$. In the worst case, the for-loop runs for $\frac{M(M-1)}{2}$ times. Therefore, Algorithm 3 has the time complexity of $O(M^2)$. Overall, the proposed double auction mechanism is a polynomial time algorithm with the time complexity of $O(M^2)$.

\section{IV. CASE STUDY: SEMANTIC TEXT TRANSMISSION}

In Section III, we introduce general auction mechanisms for semantic model and information trading. In this section, we apply the proposed auction mechanisms to the case study: semantic text transmission [2]. In this case study, we evaluate the performance of the proposed auction mechanisms by considering text semantic model trading and text semantic information trading. Specifically we show how the accuracy, $A$ can be formulated in text semantic transmission. Based the accuracy formulation, we derive the valuations of the semantic model trading and semantic information trading for semantic text transmission.
We use DeepSC [2] as the semantic communication system for the case study. DeepSC is one of the few existing works with good performance in semantic text transmission. DeepSC outperforms conventional communications and the baseline method [40] especially in the low signal-to-noise (SNR) region for all tested channels, i.e., additive white Gaussian noise (AWGN), Rayleigh fading channel, and Rician fading channel. Consistency of DeepSC indicates that our results in the case study can be applied to different channels.

Valuation functions (1) and (8) show that the buyers value semantic information based on the accuracy of the semantic transmission. Buyers are willing to pay more for information with higher accuracy. Therefore it is important to improve the model accuracy to attract higher bids. In this section, we propose a feature reduction method to improve the accuracy when only partial data is transmitted due to limited communication resources.

A. Deep Learning Enabled Semantic Communication Systems

We consider that $M$ devices perform text data transmission with DL enabled semantic communication systems, e.g., voice controlled devices (Google Nest Hub,Amazon Echo, and Apple HomePod). In DL enabled semantic communication system, collected sentences, $S = \{s_1, s_2, \ldots, s_N\}$, are encoded by semantic encoder and channel encoder. The encoded signal can be represented by

$$X = enc_c(enc_s(S)),$$

where $X \in \mathbb{R}^{N_x \times L \times D}$, $N_x$ is the number of sentences, $L$ is the sentence length, $D$ is the output dimension of channel encoder, $enc_c(\cdot)$ is the channel encoder, and $enc_s(\cdot)$ is the semantic encoder. Note that all inputs are padded to length $L$ before passing to the encoders. After winner determination of the double auction, winning devices transmit encoded information to the winning buyers. At the buyer, signal received can be expressed as

$$Y = HX + A,$$

where $H$ is the channel gain between the transmitter and receiver and $A \sim \mathcal{N}(0, \sigma_n^2)$ is the additive white Gaussian noise (AWGN). The decoded sentences are given by

$$\hat{S} = dec_c(dec_s(Y)),$$

where $dec_c(\cdot)$ and $dec_s(\cdot)$ are the semantic decoder and channel decoder of the receiver.

We adopt the network architecture of DeepSC [2] where the semantic encoder and decoder are implemented as multiple Transformer [41] encode and decode layers, and channel encoder as dense layers with different units. Our incentive mechanism can be easily extended to other network architectures by following the same evaluation procedure.

The BLEU score and the sentence similarity are two of the critical performance metrics of the text-based semantic communication system. The BLEU score measures an exact matching of words in the original and recovered sentences without considering their semantic information. In contrast to the BLEU score, sentence similarity is calculated by the cosine similarity of the extracted semantic features from original and recovered sentences. In our model, a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [42] model is used for the semantic features extraction. Let $s$ and $\hat{s}$ denote one sentence from $S$ and $\hat{S}$, respectively. The BLEU score can be expressed as

$$\log \text{BLEU} = \min \left(1 - \frac{l_s}{l_{\hat{s}}}, 0 \right) + \sum_{i=1}^{l} u_i \log p_i,$$

where $l_s$ and $l_{\hat{s}}$ are the lengths of the original and recovered sentences respectively, $u_i$ is the weight of $i$-grams, and $p_i$ is the $i$-grams score, which is given by

$$p_i = \frac{\sum_{k=1}^{K_i} \min(C_k(\hat{s}), C_k(s))}{\sum_{k=1}^{K_i} \min(C_k(\hat{s}))},$$

where $K_i$ is the number of elements in $i$-th grams, and $C_k(\cdot)$ is the frequency count function for the $k$-th element in $i$-grams.

The sentence similarity is given by

$$\text{similarity}(\hat{s}, s) = \frac{B(s) \cdot B(\hat{s})^T}{\|B(s)\| \|B(\hat{s})\|},$$

where $B(\cdot)$ is a pre-trained BERT model used to measure the sentence similarity.

In general, to obtain a higher BLEU score and similarity score, we need to increase the output dimension $D$ of the encoder [42]. However, increasing $D$ comes at the cost of larger data size, and the amount of data that devices can send is limited by the communication resources, e.g., energy supply to the devices [17]. Specifically, the similarity score and the BLEU score of device $m$ can be expressed as

$$s_m = f_{sim}(D) = f_{sim}\left(\frac{N_m}{N_s \times L \times b_f}\right),$$

and

$$\text{BLEU}_m = f_{BLEU}(D) = f_{BLEU}\left(\frac{N_m}{N_s \times L \times b_f}\right),$$

respectively, where $f_{sim}(\cdot)$ and $f_{BLEU}(\cdot)$ are simple lookup to obtain the scores of the model, $b_f$ is the number of bits used by a unit feature, and $N_m$ is the total number of bits that the device $m$ can transmit. The values of $f_{sim}(\cdot)$ and $f_{BLEU}(\cdot)$ can be obtained by using different output dimension $D$ to evaluate the similarity score and the BLEU score, respectively. A unit feature is a single entry of $X \in \mathbb{R}^{N_x \times L \times D}$, and $b_f$ is the number of bits used to represent a float type data. In our model, the data size in (10) and (11) is given by the number of words collected, i.e., $d = N_s \times L$. The total number of bits affects the communication cost as shown in (12).

From (23) and (24), it is clear that the scores are affected by the size of data and model performance. Since each device

---

1[Online]. Available: https://www.cnet.com/home/smart-home/how-to-set-up-your-new-google-nest-hub-or-nest-hub-max/

2[Online]. Available: https://www.androidauthority.com/amazon-echo-5th-gen-3095027/

3[Online]. Available: https://www.apple.com/sg/newsroom/2021/10/apple-introduces-homepod-mini-in-new-bold-and-expressive-colors/
has a different model performance and data to be sent, the similarity score and the BLEU score are different among the devices. BERT model adopts self-attention [41] modules which have complexity per layer of $O(L^2 \cdot D)$ where $L$ is the sequence length and $D$ is the representation dimension. The evaluation of sentence similarity can be done once and stored in the device for future use. Model providers could help with the evaluation after the model transmission by performing the inference of BERT model at the providers’ ends. Specifically, the similarity score and BLEU score can be stored in a lookup table and the devices can obtain the scores from the stored result based on the output dimension.

### B. Semantic-Aware Valuation for Auctions

In the semantic model trading, the devices bid according to the performance of the semantic model (1), i.e.,

$$b_m = v_m = A_p - A_m.$$  

(25)

The accuracy of the model from the model provider can be expressed as follows:

$$A_p = \lambda_m s_p + \beta_m \text{BLEU}_p,$$  

(26)

where $s_p$ and $\text{BLEU}_p$ are the similarity score and the BLEU score achievable by the model provided, respectively, $\lambda_m$ is the preference for the similarity score by the device $m$, $\beta_m$ is the preference for the BLEU score by the device $m$, and $\lambda_m + \beta_m = 1$. If $\beta_m > \lambda_m$, it indicates that the device has more interest in the exact recovery of words whereas $\lambda_m > \beta_m$ indicates higher interest in the matching of the semantic meaning. For example, some medical devices [43] would have higher $\beta_m$ because the exact recovery of medical terms is more important, whereas devices that collect data for text classification [44] would have higher $\lambda_m$.

The accuracy of the current model of device $m$ is given by:

$$A_m = \lambda_m s_m + \beta_m \text{BLEU}_m,$$  

(27)

where $s_m$ and $\text{BLEU}_m$ are the similarity score and the BLEU score achievable by the current model. In the semantic information trading, based on the communication environment and resources, each device can achieve different semantic performance when transmitting information to the buyers. Therefore, based on the semantic performance, each buyer has different preferences for the devices.

The value of the semantic information from device $m$ to buyer $n$ is given by:

$$v_n^m = \lambda_n s_m + \beta_n \text{BLEU}_m,$$  

(28)

where $\lambda_n$ is the preference for the similarity score, and $\beta_n$ is the preference for the BLEU score by the buyer $n$. As the auction is truthful for all buyers and sellers, the buyers and sellers submit bids and asks according to their true valuations, i.e., $b_n^m = v_n^m$ and $a_m = C_m$. Again, the cost of collecting the information by device $m$ can be obtained from (14).

In this case study, similarity score and BLEU score are used to model the accuracy in the valuation function. Our incentive mechanism can be extended to other semantic communication systems by adapting the valuation function to different goals of the semantic transmission. For example, SDR can be used for semantic speech signal transmission [8] and answer accuracy can be used for VQA task [25]. Regardless of the valuation function, the properties of individual rationality, incentive compatibility, and budget balance are guaranteed with the same winner and price determination algorithms in the proposed mechanisms.

### C. Feature Reduction Technique

Let $N_m$ denote the number of bits that device $m$ can send to the buyer. Based on the bit budget $N_m$, not all features of the encoded information can be sent. However, the semantic communication model is trained with a fixed number of features with output dimension $D$. A sample of feature representation output by semantic encoder with 16 features is shown in Fig. 2. Sentences decoded from partial features have a lower similarity score and BLEU score than that decoded from all features. Deep neural networks need to fine-tune the model parameters to reduce the gap in performance. Unfortunately, devices that operate on limited resources might not be able to fine-tune the model in real-time because it is both time and energy consuming. Therefore an effective feature reduction method is required for these devices to minimize the gap in performance when they have to communicate with a limited bit budget.

We propose a simple feature reduction method where the performance can be adjusted by a regularization technique [26] during the training of the model. Consider that the model on device is pre-trained with output dimension $D$, under the limited bit budget, the encoded signal, $X \in \mathbb{R}^{N_m \times L \times D}$ is reduced to $X' \in \mathbb{R}^{N_m \times L \times D'}$, where $0 < D' < D$. At the receiver, the received signal $Y' \in \mathbb{R}^{N_m \times L \times D'}$ is padded with zeros to become $Y \in \mathbb{R}^{N_m \times L \times D}$. The proposed data reduction method is illustrated in Fig. 3. To obtain $f_{\text{sim}}(\cdot)$ and $f_{\text{BLEU}}(\cdot)$, we first train the DeepSC model by using the data with dimension $D$ and next apply the trained model to evaluate the similarity scores for output dimension $d, \forall d \in [1, D]$. Then, we can obtain $f_{\text{sim}}(\cdot)$ and $f_{\text{BLEU}}(\cdot)$ from the evaluation results of test datasets.

To reduce the degradation of performance, we add a controlled dropout [26] layer before the channel decoding layer of the receiver. The conventional dropout [45] technique randomly drops units (Fig. 4(a)) in the training process to solve the overfitting issue of the deep neural networks. In contrast to conventional dropout, controlled dropout drops units intentionally, i.e., dropping a selected index of a dimension, as shown in Fig. 4(b). Specifically, the conventional dropout technique randomly drops units in the training process, e.g., unit $(x, y, z)$ is dropped by randomly selecting $x, y,$ and $z$. In contrast to conventional dropout, controlled dropout first selects an index $d_i$ randomly, and then all units from $0, 0, d_i)$ to $(N_s - 1, L - 1, d_i)$ are dropped. This is to mimic the process of the data reduction. An illustration of the effect of controlled dropout during training is shown in Fig. 3. In our experiments, we drop units from a certain index of the output dimension. As shown in [26], we can obtain a better performance than conventional dropout when the index is randomly selected. Following [26], the index is randomly selected with a dropout rate, $p_{\text{drop}}, 0 < p_{\text{drop}} < 1$. 

\[\text{Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.}\]
Controlled dropout helps the model to generalize to the reduced features during the training.

V. NUMERICAL RESULTS

In this section, we evaluate the performance of the proposed auction mechanisms and feature reduction method. The values of experiment parameters are presented in Table I. The similarity and BLEU scores are sampled according to the simulation settings in [28] for the DL-based auction. The dropout rate is set according to [26]. Following [38], we set the cost-related parameters in the double auction as shown in Table I.

![Fig. 2](image)

Sample output of semantic encoder with 16 features, input sentence: “thirdly it criticises the shortcomings but in a positive manner”.

![Fig. 3](image)

Illustration of controlled dropout during the training.

![Fig. 4](image)

An illustration of (a) conventional dropout and (b) controlled dropout.

![Fig. 5](image)

BLEU and similarity score under different output dimensions.

A. Evaluation of DeepSC With Feature Reduction

We first investigate the improvement of semantic performance under the proposed feature reduction method. With the help of the DeepSC, we set the output dimension of encoder $D$ to 16, and train the model under AWGN channel for 200 epochs. The training and test data is obtained from the proceedings of the European Parliament [46]. We use 7347 English sentences in the dataset for our evaluation, and use the rest of the English sentences for training. The performance scores are considered for the evaluation of the proposed double auction mechanism.

As described in Section IV-C, we add a controlled dropout layer between the physical layer and receiver. The dropout probability is set as 0.1 which means 10% of the features are dropped randomly in a controlled setting. We record the similarity score and the BLEU score for the output dimensions from 1 to 16, which is shown in Fig. 5. Regardless of the application of controlled dropout, we observe that the performance degrades...
as the output dimension decreases. The reason is that fewer features are transmitted. However, when the output dimension changes from 1 to 15, the performance of model trained with controlled dropout constantly outperforms conventional dropout in the baseline model. In other words, as the output dimension decreases, the baseline model has a larger performance gap compared to the model with controlled dropout. Specifically, the reduction of the similarity score per output dimension is 0.05 in the baseline model and 0.04 in the proposed model. For the reduction of the BLEU score per output dimension, it is 0.06 for the baseline model and 0.05 for the proposed model. This result clearly shows that the proposed model can maintain a similarity score of 0.80 even after 25% of feature reduction ($D = 12$) whereas the baseline model can achieve only a similarity score of 0.60 with the same output dimension. As shown in Table II, the recovered sentence has higher similarity when the controlled dropout is applied.

However, we notice that the best performance achieved by the baseline model at $D = 16$ is slightly higher than that of the proposed model. The BLEU score and the similarity score for our proposed model are 0.89 and 0.91, respectively, but both scores are 0.94 for the baseline model. The reason is that the accuracy is slightly dropped due to the generalization of the feature reduction. Overall, the gap in performance at fewer output dimensions is compensated by the controlled dropout during training.

### B. Evaluation of Deep Learning Based Auction Mechanism

Without loss of generality, we consider that the devices do not own any semantic model initially, i.e., $A_m = 0$ and $b_m = v_m = A_p$. To obtain the bid profiles, we consider $A_p \sim U[0, 0.4]$ and $v_m \sim U[0.5, 0.9]$. We collect 1000 training samples with 10 bidders (devices) in each of the samples and perform training for 500 epochs. From Fig. 6, we observe that the DL-based auction can always achieve higher revenue than that of the SPA, regardless of the values of $A_p$. The reason is that the DL-based auction mechanism can adapt to different bid profiles by optimizing the parameters in the DL network. Moreover, we observe that, while SPA is incentive compatible, it does not maximize the revenue of the model providers. In contrast to SPA, the DL-based auction maximizes the revenue of model providers while keeping the desired properties of incentive compatibility and individual rationality, which helps to attract more model providers to offer quality semantic encoder/decoder for semantic communications.

### C. Evaluation of Double Auction Mechanism

To evaluate the performance of the double auction mechanism, we generate 1000 samples and average the simulation results. We set the number of sellers to $M = 20$ and evaluate the performance under different number of buyers. The DL-based auction network can be scaled according to the number of sellers and buyers, while ensuring revenue maximization by optimizing the network parameters. Note that in the following discussion, we refer semantic information buyers as buyers and devices as sellers for simplicity.

To validate that the double auction mechanism is individually rational and budget balanced, we record the values of ask, bid, and price in one of the samples with $M = 20$ and $N = 10$. The values are shown in Fig. 7. We observe that there are totally 7 winning seller-buyer pairs, and the utilities for all of the winning pairs are positive. This means that the winning sellers are paid higher than their cost, and the winning buyers pay no more than their true valuation for the semantic information. Therefore, both buyers and sellers have incentives to participate in the auction. For the losing sellers and buyers, their utilities are zero. This shows that the property of individual rationality is achieved because all of the buyers and sellers are awarded with a non-negative utility. The price paid by winning sellers is equal
to the payment received by the winning buyers. Thus, the budget balanced property is satisfied.

The average utility of the winning buyers and sellers are presented in Fig. 8, which is obtained by averaging the values of 1000 samples. Intuitively, as the number of buyers increases, the sellers have more choices to achieve higher utilities. From Fig. 8, we observe that the auction mechanism helps to increase the average utility of the winning sellers as the number of buyers grows. Thus, our proposed mechanism can attract more sellers to participate in the information exchange with semantic communication systems.

To investigate the impact of DL in the double auction, we compare the average utilities of winning sellers with and without DL mechanism. The results without the DL mechanism (i.e., the baseline) are obtained by using the double auction mechanism proposed in [32]. In the baseline method [32], the double auction mechanism follows a similar two-stage approach. However, in the candidate-determination and pricing stage of [32] the winning buyers and prices are not determined by the DLA algorithm. Instead of DLA algorithm, the baseline method [32] use a set of fixed rules to decide the winners and prices. It is shown in Fig. 9 that the average utility of the winning sellers is higher when DL mechanism is adopted in the double auction. The reason is that the DL mechanism helps to maximize the revenue of the sellers. By varying the number of buyers in the experiments, it shows that the DL-based auction mechanism can adapt to different number of participants. While the maximization of the revenue is guaranteed by the DL-based auction, increasing the number of sellers expands the computation time of the polynomial time algorithm (as discussed in Section III-C). Polynomial time complexity is the common overall runtime for the double auction mechanisms that ensure the truthfulness of the auction [32],[47]. Polynomial runtime is considered efficient and tractable [48], i.e., the outcome can be obtained in a reasonable period of time. Hence, the polynomial runtime algorithm is more suitable for large-scale implementation than an exponential runtime algorithm. Moreover, increasing the number of sellers and buyers does not affect the performance because the DL network can adjust the decision by optimizing the revenue with the loss function (7). The mechanism remains individually rational and incentive compatible regardless of the size of the input.

As shown in Fig. 10, the average utility of the winning sellers is higher when the sellers set the payment price $\theta_m \geq 0.5$. The reason is that the sellers with higher $\theta_m$ obtain the semantic model which has a higher BLEU score and similarity score. Hence the buyers are willing to pay more to obtain more accurate information. This insight is verified in Fig. 11, in which we can see that the similarity score and the BLEU score are higher for sellers with $\theta_m \geq 0.5$. The higher similarity and BLEU scores motivate the buyers to submit higher bids to the sellers, which results in higher utilities as shown in Fig. 10. Furthermore, it is shown in Fig. 12 that there are more sellers with $\theta_m \geq 0.5$ from the winning sellers. In other words, the seller with higher $\theta_m$ has a higher chance to win the auction, regardless of the number of buyers.

To verify the truthfulness of the double auction, the sellers and buyers are randomly chosen to evaluate their utilities when their bid and ask are different from their true valuation. In Fig. 13(a),

![Fig. 8. Average utility of winning buyers and sellers.](image1)

![Fig. 9. Average utility of winning sellers with and without deep learning (baseline) in the double auction mechanism.](image2)

![Fig. 10. Average utility of winning sellers with different ranges of $\theta_m$.](image3)

![Fig. 11. (a) Average similarity score of sellers (b) Average BLEU score of sellers with different cost of the semantic model, $\theta_m$.](image4)
In this article, we have proposed incentive mechanisms for both semantic model trading and semantic information trading. We developed the valuation functions for general semantic communications and performed a case study of the proposed auctions for semantic text transmission. To improve the system performance, we have proposed an effective feature reduction method to support devices with limited transmission resources. The simulation results show that the proposed method helps to increase significantly the utility of devices in the semantic information trading. Moreover, with the double auction mechanism, we have matched the buyers and devices effectively. It is also shown that the revenue of the semantic model provider can be maximized while keeping the properties of incentive compatibility and individual rationality.

For future research directions, we can consider the semantic-aware incentive mechanism design in non-text-based transmissions such as wireless images and video transmission, and other semantic-based intelligent tasks. For future works, considering that the raw data collected from different regions decays over time, we can count the age of information in the value functions of raw data. The difference in the age of information can also be taken into account in the evaluation of transmission accuracy. Moreover, we can consider that the semantic information, which is extracted from different types of raw data, e.g., text, image, and audio, has different values.

**VI. CONCLUSION AND FUTURE DIRECTIONS**

From the experiment results, we observe that the sellers that pay higher prices for the semantic models can achieve better similarity and BLEU scores in the double auction. It is shown that the sellers with better performance are more likely to win the auction and obtain higher utilities. Numerical results also show that the proposed double auction is incentive compatible, individually rational, and budget balanced.
K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in Proc. Annua. Meeting Assoc. Comput. Linguistics, 2002, pp. 311–318.

P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, “Deep source-channel coding for sentence semantic transmission with HARQ,” IEEE Trans. Commun., vol. 70, no. 8, pp. 5225–5240, 2022.

Y. Yang et al., “Semantic communications with artificial intelligence tasks: Reducing bandwidth requirements and improving artificial intelligence task performance,” IEEE Ind. Electron. Mag., 2022.

Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, and S. Sun, “Economics of semantic communication system in wireless powered Internet of Things,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2022, pp. 8637–8641.

M. Li, J. Gao, L. Zhao, and X. Shen, “Adaptive computing scheduling for edge-assisted autonomous driving,” IEEE Trans. Veh. Technol., vol. 70, no. 6, pp. 5318–5331, Jun. 2021.

Z. Q. Liew, H. A. Hzain, W. Zhuang, and X. Shen, “Joint RAN slicing and computation offloading for autonomous vehicular networks: A learning-assisted hierarchical approach,” IEEE Open J. Veh. Technol., vol. 2, pp. 272–288, 2021.

W. Wu et al., “Dynamic RAN slicing for service-oriented vehicular networks via constrained learning,” IEEE J. Sel. Areas Commun., vol. 39, no. 7, pp. 2076–2089, Jul. 2021.

A. Nanda, D. Pathania, J. F. Rodrigues, and S. A. Kozlow, “Internet of autonomous vehicles communications security: Overview, issues, and directions,” IEEE Wireless Commun., vol. 26, no. 4, pp. 60–65, Aug. 2019.

N. Pappas and M. Kontouris, “Goal-oriented communication for real-time tracking in autonomous systems,” in Proc. IEEE Int. Conf. Auton. Syst., 2021, pp. 1–5.

M. Zhu, C. Feng, J. Chen, C. Guo, and X. Gao, “Video semantics based resource allocation algorithm for spectrum multiplexing scenarios in vehicular networks,” in Proc. IEE/CIC Int. Conf. Commun. China Workshops, 2021, pp. 31–36.

S. K. Tayabay et al., “5G vehicular network resource management for improving radio access through machine learning,” IEEE Access, vol. 8, pp. 6792–6800, 2020.

H. Xie, Z. Qin, and G. Y. Li, “Task-oriented multi-user semantic communications for VQA,” IEEE Wireless Commun. Lett., vol. 11, no. 3, pp. 553–557, 2021.

B. Ko, H.-G. Kim, K.-J. Oh, and H.-J. Choi, “Controlled dropout: A different approach to using dropout on deep neural network,” in Proc. IEEE Int. Conf. Big Data Smart Comput., 2017, pp. 358–362.

H. Zheng, K. Xiong, P. Fan, Z. Zhong, and K. B. Letaief, “Age of information-based wireless powered communication networks with self-charging nodes,” IEEE J. Sel. Areas Commun., vol. 39, no. 5, pp. 1393–1411, May 2021.

W. Y. B. Lim et al., “Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning,” IEEE Trans. Parallel Distrib. Syst., vol. 33, no. 3, pp. 536–550, Mar. 2022.

H. Yu, Y. Yang, H. Zhang, R. Liu, and Y. Ren, “Reputation-based reverse combination auction incentive method to encourage vehicles to participate in the VCS system,” IEEE Trans. Netw. Sci. Eng., vol. 8, no. 3, pp. 2469–2481, Jul.–Sep. 2021.

J. S. Ng, W. Y. B. Lim, Z. Xiong, D. Niyato, C. Leung, and C. Miao, “A double auction mechanism for resource allocation in coded vehicular edge computing,” IEEE Trans. Veh. Technol., vol. 71, no. 2, pp. 1832–1845, Feb. 2022.

J. Wang, C. Jiang, Z. Bie, T. Q. Quek, and Y. Ren, “Mobile data transactions in device-to-device communication networks: Pricing and auction,” IEEE Wireless Commun. Lett., vol. 5, no. 3, pp. 300–303, Jun. 2016.

A.-L. Jin, W. Song, P. Wang, D. Niyato, and P. Ju, “Auction mechanisms toward efficient resource sharing for cloudlets in mobile cloud computing,” IEEE Trans. Serv. Comput., vol. 9, no. 6, pp. 895–909, Dec. 2016.

W. Yang et al., “Semantic communication meets edge intelligence,” IEEE Wireless Commun., vol. 29, no. 5, pp. 28–35, 2022.

R. B. Myerson, “Optimal auction design,” Math. Oper. Res., vol. 6, no. 1, pp. 51–64, Feb. 1981.

P. Dütting, Z. Feng, H. Narasimhan, D. Parkes, and S. S. Ravindran, “Optimal auctions through deep learning,” in Proc. 39th Annu. ACM Symp. Theory Comput., 2007, pp. 607–609. [Online]. Available: https://aclanthology.org/2020.emnlp-main.491

Y. Jiao, P. Wang, D. Niyato, B. Lin, and D. I. Kim, “Toward an automated auction framework for wireless federated learning services market,” IEEE Trans. Mobile Comput., vol. 20, no. 10, pp. 3034–3048, Oct. 2021.

T. Q. Dinh, B. Liang, T. Q. Quek, and H. Shin, “Online resource procurement and allocation in a hybrid edge-cloud computing system,” IEEE Trans. Wireless Commun., vol. 19, no. 3, pp. 2137–2149, Mar. 2020.

N. Farsad, M. Rao, and A. Goldsmith, “Deep learning for joint source-channel coding of text,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Calgary, AB, Canada, 2018, pp. 2326–2330.

A. Savwani et al., “Attention is all you need,” in Proc. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017, pp. 5998–6008.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

M. Dhyani and R. Kumar, “An intelligent Chatbot using deep learning with Bidirectional RNN and attention model,” Mater. Today: Proc., vol. 34, pp. 817–824, 2021.

F. P. Shah and V. Patel, “A review on feature selection and feature extraction for text classification,” in Proc. Int. Conf. Wireless Commun., Signal Process. Netw., 2016, pp. 2264–2268.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

P. Koehn et al., “Europarl: A parallel corpus for statistical machine translation,” in Proc. MT Summit, 2005, vol. 5, pp. 79–86.

W. Sun, J. Liu, Y. Yue, and H. Zhang, “Double auction-based resource allocation for mobile edge computing in industrial Internet of Things,” IEEE Trans. Ind. Informat., vol. 14, no. 10, pp. 4692–4701, Oct. 2018.

A. Cobham, “The intrinsic computational difficulty of functions,” in Proc. Log., Methodol. Philosophy Sci., Proc. 1964 Int. Congr. (Stud. Log. Found. Math.), Y. Bar-Hillel, Ed., 1965, pp. 24–30.

Zi Qin Liew received the degree (Hons.) in electronic and electrical engineering in 2018 from the Nanyang Technological University, Singapore, where he is currently working toward the Ph.D. degree with Alibaba Group and the Alibaba-NTU Joint Research Institute. His research interests include wireless communications and incentive mechanisms.

Hongyang Du (Graduate Student Member, IEEE) received the B.Sc. degree from Beijing Jiaotong University, Beijing, China, in 2021. He is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, Energy Research Institute @ NTU, Nanyang Technological University, Singapore, under the Interdisciplinary Graduate Program. His research interests include semantic communications, reconfigurable intelligent surface, and communication theory. He was recognized as an Exemplary Reviewer of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2021. He was the recipient of IEEE Daniel E. Noble Fellowship Award in 2022.

Wei Yang Bryan Lim received the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2022 under the Alibaba Ph.D. Talent Programme. He is currently a Wallenberg-NTU Presidential Postdoctoral Fellow. He was the recipient of the Most Promising Industrial Postgraduate Student Award from Nanyang Technological University. His works have won Best Paper Awards including in the IEEE Wireless Communications and Networking Conference (WCNC) and IEEE SPCC Technical Committee Best Paper Award. He regularly is the Technical Programme Committee Member in flagship conferences, is part of the guest editor team of special issues in IEEE Wireless Communication Magazine and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATION, and is currently the Review Board Member of IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.}
Zehui Xiong (Member, IEEE) received the Ph.D. degree from Nanyang Technological University (NTU), Singapore. He was a Visiting Scholar with Princeton University, Princeton, NJ, USA, and University of Waterloo, Waterloo, ON, Canada. He is currently an Assistant Professor with the Singapore University of Technology and Design, Singapore, and also an Honorary Adjunct Senior Research Scientist with Alibaba-NTU Singapore Joint Research Institute, Singapore. He has authored or coauthored more than 200 research papers in leading journals and flagship conferences and many of them are ESI Highly Cited Papers. His research interests include wireless communications, Internet of Things, blockchain, edge intelligence, and Metaverse. He was the recipient of the ten Best Paper Awards in international conferences and is listed in the World’s Top 2% Scientists identified by Stanford University. He is currently the editor or guest editor for many leading journals including IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING. He was also the recipient of IEEE Early Career Researcher Award for Excellence in Scalable Computing, IEEE Technical Committee on Blockchain and Distributed Ledger Technologies Early Career Award, IEEE Internet Technical Committee Early Achievement Award, IEEE TCI Rising Star Award, IEEE TCCLD Rising Star Award, IEEE Best Land Transportation Paper Award, IEEE CSIM Technical Committee Best Journal Paper Award, IEEE SPCC Technical Committee Best Paper Award, IEEE VTS Singapore Best Paper Award, Chinese Government Award for Outstanding Students Abroad, and NTU SCSE Best Ph.D. Thesis Runner-Up Award. He is currently the Associate Director of Future Communications R&D Programme.

Dusit Niyato (Fellow, IEEE) received B.Eng. degree from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand, in 1999, and Ph.D. degree in electrical and computer engineering from the University of Southern California, Los Angeles, CA, USA, in 1999. He was a Tenured Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. Since 2007, he has been an Sungkyunkwan University-Fellowship and then a Distinguished Professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is a Fellow of the Korean Academy of Science and Technology and a Member of the National Academy of Engineering of Korea. He has been the first recipient of the NRF of Korea Engineering Research Center in Wireless Communications for RF Energy Harvesting since 2014. He has been listed as a 2020/2022 Highly Cited Researcher by Clarivate Analytics. Since 2001, he has been the Editor, Editor at Large, and Area Editor of Wireless Communications for the IEEE TRANSACTIONS ON COMMUNICATIONS. From 2002 to 2011, he was the Editor and a Founding Area Editor of Cross-Layer Design and Optimization for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. From 2008 to 2011, he was the Co-Editor-in-Chief for the IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS. He was the Founding Editor-in-Chief for the IEEE WIRELESS COMMUNICATIONS LETTERS, from 2012 to 2015. He was selected the 2019 recipient of the IEEE Communications Society Joseph Lo Cicero Award for Exemplary Service to Publications. He was the General Chair for IEEE ICC 2022 in Seoul.

Chunyan Miao (Fellow, IEEE) is currently the Director of the Joint Nanyang Technological University-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore. She is the Chair of the School of Computer Science and Engineering, Nanyang Technological University. She is the Editor-in-Chief of the International Journal of Information Technology published by the Singapore Computer Society.

Dong In Kim (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He was a Tenured Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. Since 2007, he has been an Sungkyunkwan University-Fellowship and then a Distinguished Professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is a Fellow of the Korean Academy of Science and Technology and a Member of the National Academy of Engineering of Korea. He has been the first recipient of the NRF of Korea Engineering Research Center in Wireless Communications for RF Energy Harvesting since 2014. He has been listed as a 2020/2022 Highly Cited Researcher by Clarivate Analytics. Since 2001, he has been the Editor, Editor at Large, and Area Editor of Wireless Communications I for the IEEE TRANSACTIONS ON COMMUNICATIONS. From 2002 to 2011, he was the Editor and a Founding Area Editor of Cross-Layer Design and Optimization for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. From 2008 to 2011, he was the Co-Editor-in-Chief for the IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS. He was the Founding Editor-in-Chief for the IEEE WIRELESS COMMUNICATIONS LETTERS, from 2012 to 2015. He was selected the 2019 recipient of the IEEE Communications Society Joseph Lo Cicero Award for Exemplary Service to Publications. He was the General Chair for IEEE ICC 2022 in Seoul.